IBM SPSS Neural Networks 21

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 98.

Cette version s'applique à IBM® SPSS® Statistics 21 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Réseaux neuronaux fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Réseaux neuronaux doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics offre des informations complètes, cohérentes et précises permettant aux preneurs de décision d'améliorer leurs performances professionnelles. Un portefeuille complet de solutions de business intelligence, d'analyses prédictives, de performance financière et de gestion de la stratégie, et d'applications analytiques permet une connaissance claire et immédiate et offre des possibilités d'actions sur les performances actuelles et la capacité de prédire les résultats futurs. En combinant des solutions du secteur, des pratiques prouvées et des services professionnels, les entreprises de toute taille peuvent générer la plus grande productivité, automatiser les décisions en toute confiance et apporter de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les entreprises à prédire des événements futurs et à agir de manière proactive en fonction de ces prédictions pour apporter de meilleurs résultats. Des clients dans les domaines commerciaux, gouvernementaux et académiques se servent de la technologie IBM SPSS comme d'un avantage concurrentiel pour attirer ou retenir des clients, tout en réduisant les risques liés à l'incertitude et à la fraude. En intégrant le logiciel IBM SPSS à leurs opérations quotidiennes, les entreprises peuvent effectuer des prévisions, et sont capables de diriger et d'automatiser leurs décisions afin d'atteindre leurs objectifs commerciaux et d'obtenir des avantages concurrentiels mesurables. Pour plus d'informations ou pour contacter un représentant, visitez le site *http://www.ibm.com/spss*.

Support technique

Un support technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, visitez le site IBM Corp. à l'adresse *http://www.ibm.com/support*. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Support technique pour les étudiants

Si vous êtes un étudiant qui utilise la version pour étudiant, personnel de l'éducation ou diplômé d'un produit logiciel IBM SPSS, veuillez consulter les pages Solutions pour l'éducation (*http://www.ibm.com/spss/rd/students/*) consacrées aux étudiants. Si vous êtes un étudiant utilisant une copie du logiciel IBM SPSS fournie par votre université, veuillez contacter le coordinateur des produits IBM SPSS de votre université.

Service clients

Si vous avez des questions concernant votre livraison ou votre compte, contactez votre bureau local. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

IBM Corp. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, accédez au site *http://www.ibm.com/software/analytics/spss/training*.

Contenu

Annexes

Partie I:
Guide de l'utilisateur

Chapitre
1

Introduction aux réseaux neuronaux

Les réseaux neuronaux constituent l'outil de prédilection dans de nombreuses applications prédictives d'exploration de données en raison de leurs puissance, souplesse et convivialité. Les réseaux neuronaux prédictifs sont particulièrement utiles dans les applications où le processus sous-jacent est complexe, comme dans les cas suivants :

- Prévision de la demande des consommateurs pour rationaliser les coûts de production et de livraison.
- Prévision de la probabilité de réponse à un publipostage pour déterminer les ménages d'une liste de mailing auxquels envoyer une offre.
- Evaluation d'un demandeur pour déterminer le risque de prolonger son crédit.
- Détection des opérations frauduleuses dans une base de données de déclarations de sinistre.

Les réseaux neuronaux utilisés dans les applications prédictives, tels que les réseaux de type **perceptron multistrate (MLP, Multilayer Perceptron)** et **fonction à base radiale (RBF, Radial Basis Function)**, sont supervisés, en ce sens que les résultats prévus par le modèle peuvent être comparés aux valeurs connues des variables cible. L'option Réseaux neuronaux vous permet d'ajuster les réseaux MLP et RBF et d'enregistrer les modèles obtenus à des fins d'évaluation.

Qu'est-ce qu'un réseau neuronal ?

Le terme **réseau neuronal** s'applique à une famille de modèles vaguement apparentée, caractérisée par un grand espace de paramètres et une structure flexible, inspirée des études sur le fonctionnement du cerveau. Au fur et à mesure que la famille s'est agrandie, la plupart des nouveaux modèles ont été conçus pour des applications non biologiques, bien que la majeure partie de la terminologie associée reflète leur origine.

Les définitions spécifiques des réseaux neuronaux sont aussi variées que les domaines dans lesquels elles sont employées. Alors qu'aucune définition ne couvre exactement l'ensemble de la famille de modèles, pour l'instant, examinons la description suivante(Haykin, 1998) :

Un réseau neuronal est un processeur massivement distribué en parallèle qui a une propension naturelle à stocker des connaissances empiriques et à les rendre disponibles en vue d'une utilisation. Il ressemble au cerveau sur deux aspects :

- La connaissance est acquise par le réseau à travers un processus d'apprentissage.
- Les connexions entre les neurones, connues sous le nom de pondérations synaptiques, servent à stocker les connaissances.

Pour obtenir des détails sur la raison pour laquelle cette définition est peut-être trop restrictive, reportez-vous à (Ripley, 1996)

Pour différencier les réseaux neuronaux des méthodes statistiques traditionnelles à l'aide de cette définition, ce qui n'est *pas* explicite est tout aussi important que le texte de la définition lui-même. Par exemple, le modèle de régression linéaire traditionnel peut acquérir des connaissances via la

méthode des moindres carrés et stocker cette connaissance dans les coefficients de régression. En ce sens, il s'agit d'un réseau neuronal. En réalité, vous pouvez avancer que la régression linéaire est un cas particulier de certains réseaux neuronaux. Toutefois, la régression linéaire a une structure de modèle stricte et un ensemble d'hypothèses qui sont imposés avant d'acquérir des connaissances de ces données.

En revanche, la définition ci-dessus crée des contraintes minimes pour la structure du modèle et les hypothèses. Par conséquent, un réseau neuronal peut se rapprocher d'un grand nombre de modèles statistiques sans que vous deviez imaginer à l'avance certaines relations entre les variables dépendantes et indépendantes. En fait, la forme de la relation est déterminé pendant le processus d'apprentissage. Si une relation linéaire entre les variables dépendantes et indépendantes est adaptée, les résultats du réseau neuronal devraient se rapprocher étroitement de ceux du modèle de régression linéaire. Si une relation non linéaire est plus adaptée, le réseau neuronal se rapproche automatiquement de la structure du modèle « correcte ».

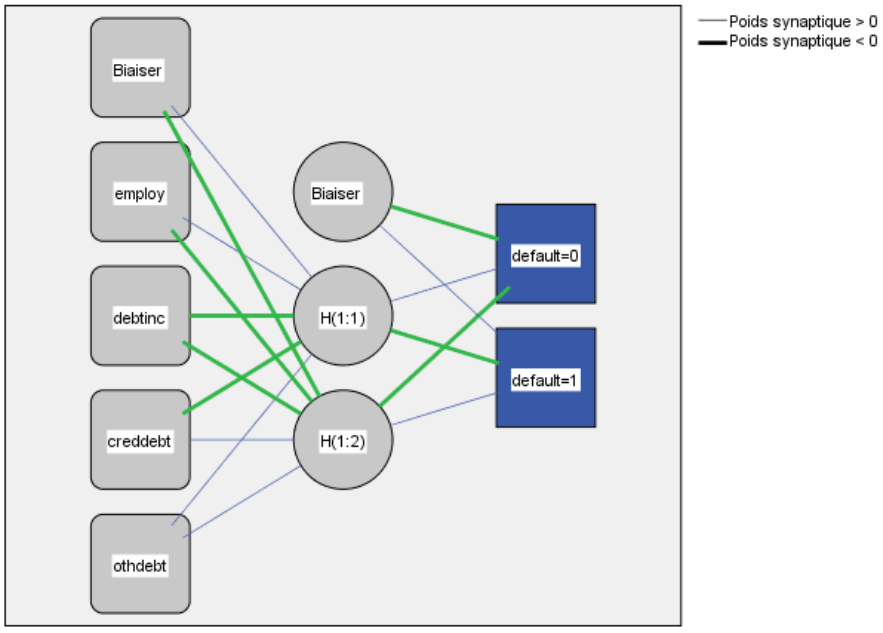
L'inconvénient de cette flexibilité est que les pondérations synaptiques d'un réseau neuronal ne sont pas facilement interprétables. Par conséquent, si vous essayez d'expliquer un processus sous-jacent qui crée des relations entre les variables dépendantes et indépendantes, il serait préférable d'utiliser un modèle statistique plus traditionnel. Toutefois, s'il n'est pas important de pouvoir interpréter le modèle, vous pouvez souvent obtenir plus rapidement des résultats du modèle satisfaisants en utilisant un réseau neuronal.

Structure d'un réseau neuronal

Bien que les réseaux neuronaux imposent des contraintes minimes à la structure du modèle et aux hypothèses, il est utile de comprendre leur **architecture** générale. Le réseau MLP ou RBF est une fonction de variables indépendantes (également appelées entrées) qui minimisent l'erreur de prévision des variables cible (également appelées sorties).

Le produit comporte l'ensemble de données *bankloan.sav*, qui peut s'avérer utile pour l'identification des personnes susceptibles de manquer à leurs engagements parmi un groupe de demandeurs de prêt. Un réseau MLP ou RBF appliqué à ce problème est une fonction des mesures qui minimise l'erreur dans la prévision du manquement. La figure suivante permet de saisir la forme de cette fonction.

Figure 1-1
Architecture d'anticipation avec une strate masquée



Cette structure est appelée architecture d'**anticipation** car le flux des connexions du réseau progresse de la strate d'entrée vers la strate de résultat sans former de boucles de réaction. Dans cette figure :

- La **strate d'entrée** contient les variables indépendantes.
- La **strate masquée** contient les unités ou noeuds non observables. La valeur de chaque unité masquée constitue une fonction des variables indépendantes ; la forme exacte de la fonction dépend, pour une partie, du type de réseau et, pour une autre partie, des spécifications contrôlables par l'utilisateur.
- La **strate de résultat** contient les réponses. Dans la mesure où l'historique du manquement est une variable qualitative comprenant deux modalités, celle-ci est recodée en deux variables indicatrices. Chaque unité de résultat constitue une fonction des unités masquées. La forme exacte de la fonction dépend, pour une partie, du type de réseau et, pour une autre partie, des spécifications contrôlables par l'utilisateur.

Le réseau MLP autorise une seconde strate masquée ; dans ce cas, chaque unité de la seconde strate masquée est une fonction des unités dans la première strate masquée et chaque réponse est une fonction des unités dans la seconde strate masquée.

Chapitre
2

Perceptron multistrate

La procédure Perceptron multistrate produit un modèle de prévision pour une ou plusieurs variables (cible) dépendantes en fonction de valeurs de variables prédites.

Exemples : Voici deux scénarios utilisant la procédure MLP :

Un responsable des prêts dans une banque souhaite pouvoir identifier les caractéristiques qui indiquent les personnes susceptibles de manquer à leurs engagements et d'utiliser ces caractéristiques pour identifier les bons et les mauvais risques de crédit. A l'aide d'un échantillon d'anciens clients, elle peut former un modèle Perceptron multistrate, valider l'analyse grâce à un échantillon traité d'anciens clients, puis utiliser le réseau pour classer les clients éventuels entre bons et mauvais risques de crédit.

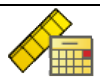
Un hôpital souhaite effectuer un suivi des coûts et des durées de séjour des patients admis pour soigner un infarctus du myocarde (crise cardiaque). Des estimations précises de ces mesures permettent à l'administration de gérer correctement le nombre de lits disponibles lors du traitement des patients. Grâce à l'utilisation des archives sur les traitements d'un échantillon de patients ayant été soignés pour un infarctus du myocarde, l'administration peut former un réseau pour prévoir le coût et la durée du séjour à l'hôpital.

Variables dépendantes Les variables dépendantes peuvent être :

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

  La procédure considère que le niveau de mesure approprié a été assigné à toutes les variables dépendantes, bien que vous puissiez changer provisoirement le niveau de mesure d'une variable en cliquant avec le bouton droit de la souris sur la variable dans la liste des variables source, puis en sélectionnant un niveau de mesure dans le menu contextuel.

Dans la liste des variables, une icône indique le niveau de mesure et le type de données :

	Numérique	**Chaîne**	**Date**	**Heure**
Echelle (continue).		n/a		
Ordinal				
Nominal				

Variables prédites. Les variables prédites peuvent être spécifiées en tant que facteurs (qualitatifs) ou covariables (d'échelle).

Codage des variables indicatrices. La procédure recode provisoirement les variables prédites qualitatives et les variables dépendantes via le codage un-de-*c* pour la durée de la procédure. S'il existe des modalités *c* d'une variable, la variable est stockée comme vecteurs c, la première modalité étant identifiée par (1,0,...,0), la suivante par (0,1,0,...,0), ... et la dernière par (0,0,...,0,1).

Ce système de codage augmente le nombre de pondérations synaptiques et peut résulter en une formation plus lente, mais les méthodes de codage plus compactes aboutissent généralement à des réseaux neuronaux mal ajustés. Si la formation de votre réseau s'effectue très lentement, essayez de réduire le nombre de modalités dans vos variables prédites qualitatives en combinant des modalités similaires ou en supprimant les observations comportant des modalités extrêmement rares.

Les codage un-de-*c* repose entièrement sur les données de formation, même si un échantillon de test ou traité est défini (reportez-vous à Partitions sur p. 9). Ainsi, si les échantillons de test ou traités contiennent des observations avec des modalités de variables prédites ne figurant pas dans les données de formation, ces observations ne sont pas utilisées par la procédure ou dans l'évaluation. Si les échantillons de test ou traités contiennent des observations avec des modalités de variables dépendantes ne figurant pas dans les données de formation, ces observations ne sont pas utilisées par la procédure mais peuvent être notées.

Rééchelonnement. Les variables d'échelle dépendantes et les covariables sont rééchelonnées par défaut pour améliorer la formation du réseau. Le rééchelonnement repose entièrement sur les données de formation, même si un échantillon de test ou traité est défini (reportez-vous à Partitions sur p. 9). En d'autres termes, en fonction du type de rééchelonnement, la moyenne, l'écart-type, la valeur minimale ou la valeur maximale d'une covariable ou d'une variable dépendante ne sont calculées qu'à l'aide des données de formation. Si vous spécifiez une variable pour définir des partitions, il est important que ces covariables ou variables dépendantes présentent des distributions similaires à travers les échantillons de formation, de test et traités.

Pondérations d'effectif. Cette procédure ignore les pondérations d'effectif.

Réplication de résultats. Si vous souhaitez répliquer exactement vos résultats, outre les mêmes paramètres de procédure, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires, le même ordre de données et le même ordre de variables. Vous trouverez ci-après plus de détails sur cet aspect.

- **Génération de nombres aléatoires.** La procédure utilise la génération de nombres aléatoires pendant l'attribution aléatoire de partitions, le sous-échantillonnage aléatoire pour l'initialisation des pondérations synaptiques, le sous-échantillonnage aléatoire pour la sélection automatique de l'architecture et l'algorithme recuit simulé utilisé dans l'initialisation de la pondération et la sélection automatique de l'architecture. Pour reproduire les mêmes résultats aléatoires à l'avenir, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires avant chaque exécution de la procédure Perceptron multistrate. Reportez-vous à Préparation des données pour l'analyse sur p. 37 pour des instructions détaillées.
- **Tri par observation.** Les méthodes de formation en ligne et par mini-commande (reportez-vous à Formations sur p. 13) dépendent explicitement de l'ordre des observations. Toutefois, même la formation par commande dépend de l'ordre des observations car l'initialisation des pondérations synaptiques implique le sous-échantillonnage de l'ensemble de données.

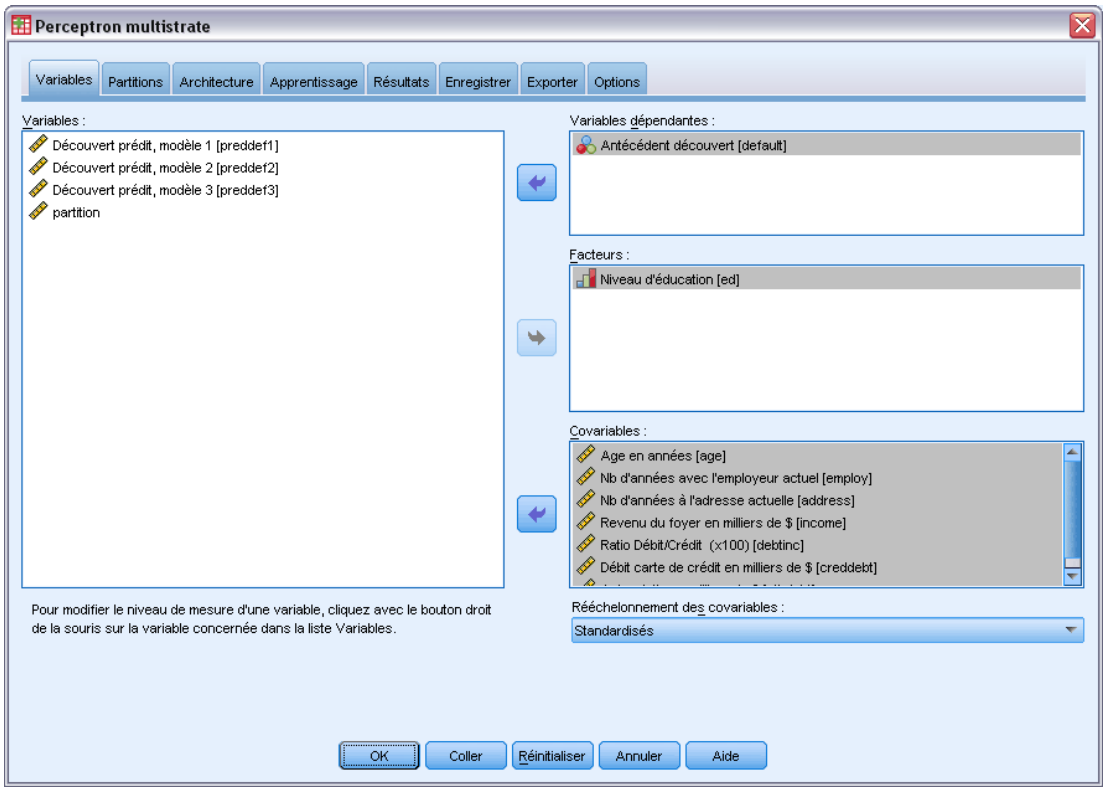
  Pour réduire les effets de tri, classez les observations de manière aléatoire. Pour vérifier la stabilité d'une solution donnée, vous pouvez obtenir différentes solutions dans lesquelles les observations sont triées de différentes manières aléatoires. Si les fichiers sont très volumineux, vous pouvez effectuer plusieurs fois l'opération sur un échantillon des observations triées de différentes manières aléatoires.
- **Ordre des variables.** Les résultats peuvent être influencés par l'ordre des variables dans les listes des facteurs et des covariables en raison du schéma différent de valeurs initiales affectées lorsque l'on change l'ordre des variables. Comme avec les effets d'ordre des observations, vous pouvez essayer différents ordres de variables (il suffit d'utiliser la fonction glisser-déplacer dans les listes de facteurs et de covariables) pour évaluer la stabilité d'une solution donnée.

Création d'un réseau Perceptron multistrate

A partir des menus, sélectionnez :

Analyse > Réseaux neuronaux > Perceptron multistrate...

Figure 2-1
Perceptron multistrate : l'onglet Variables



- Sélectionnez au moins une variable dépendante.
- Sélectionnez au moins un facteur ou une covariable.

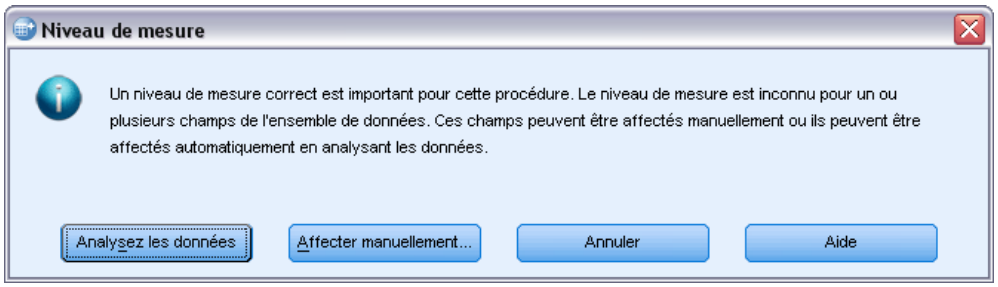
Dans l'onglet Variables, vous pouvez aussi changer la méthode de rééchelonnement des covariables. Les choix sont les suivants :

- **Standardisés.** Soustrayez la moyenne et divisez le résultat par l'écart-type, (x−moyenne)/s.
- **Normalisé.** Soustrayez le minimum et divisez le résultat par l'intervalle, (x−min)/(max−min). Les valeurs normalisées sont comprises entre 0 et 1.
- **Normalisé ajusté.** Version ajustée de la soustraction du minimum et de la division du résultat par l'intervalle, [2*(x−min)/(max−min)]−1. Les valeurs normalisées ajustées sont comprises entre −1 et 1.
- **Aucune.** Aucun rééchelonnement des covariables.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 2-2
Alerte du niveau de mesure

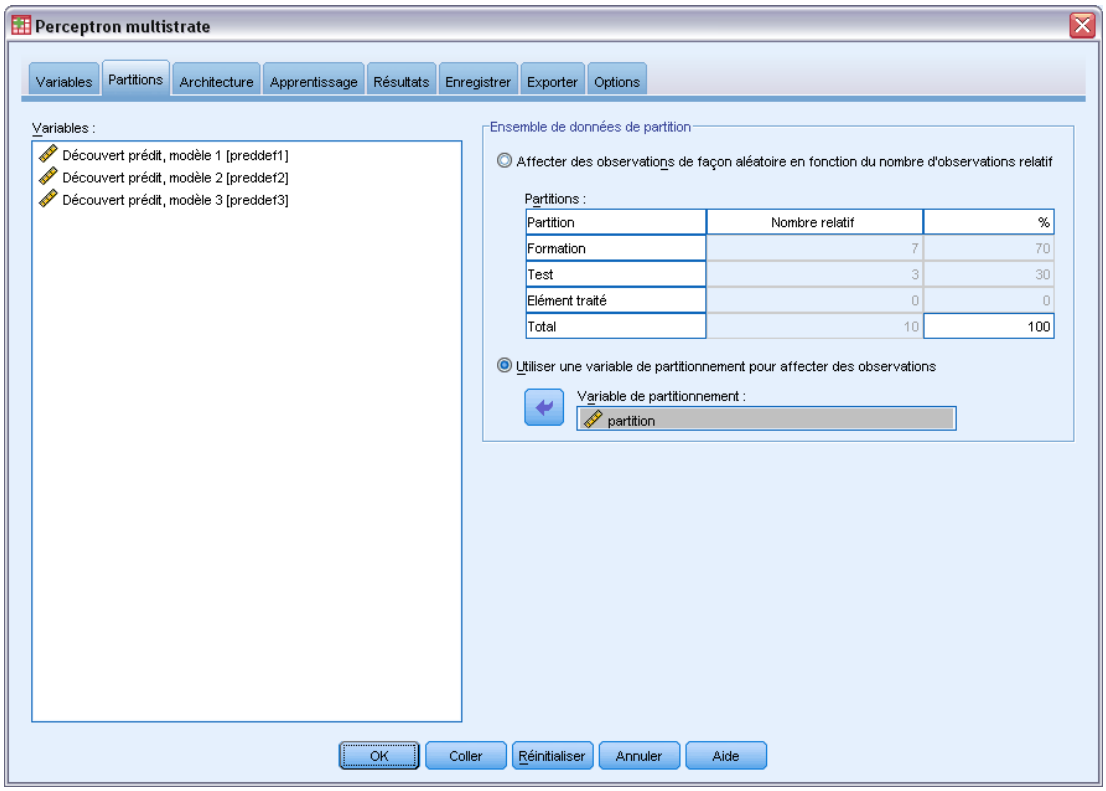


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Partitions

Figure 2-3
Perceptron multistrate : Onglet Partitions



Ensemble de données de partition. Ce groupe indique la méthode de partitionnement de l'ensemble de données actif en échantillons d'apprentissage, de test et traité. L'**échantillon d'apprentissage** comprend les enregistrements de données utilisés pour former le réseau neuronal. Un certain pourcentage d'observations contenues dans l'ensemble de données doit être affecté à l'échantillon d'apprentissage pour l'obtention d'un modèle. L'**échantillon de test** est un ensemble indépendant d'enregistrements de données utilisé pour identifier des erreurs au cours de la formation afin d'empêcher une formation excessive. Nous vous conseillons fortement de créer un échantillon d'apprentissage. Une formation de réseau sera en général plus efficace si l'échantillon de test est plus petit que l'échantillon de fonction. L'**échantillon traité** est un autre ensemble indépendant d'enregistrements de données utilisé pour évaluer le réseau neuronal final ; l'erreur pour l'échantillon traité donne une estimation « honnête » de la capacité de prévision du modèle parce que les observations traitées n'ont pas été utilisées pour construire le modèle.

- **Affecter des observations de façon aléatoire en fonction du nombre d'observations relatif.** Indiquez le nombre d'observations relatif (ratio) affecté de façon aléatoire à chaque échantillon (d'apprentissage, de test et traité). La colonne % rapporte le pourcentage d'observations qui seront affectées à chaque échantillon en fonction des nombres relatifs que vous avez spécifiés.

  Par exemple, si vous spécifiez 7, 3 et 0 comme nombres relatifs pour les échantillons d'apprentissage, de test et traité, ces valeurs correspondent à 70 %, 30 % et 0 %. Si vous spécifiez 2, 1, 1 comme nombres relatifs, ces valeurs correspondent à 50 %, 25 % et 25 %. 1,

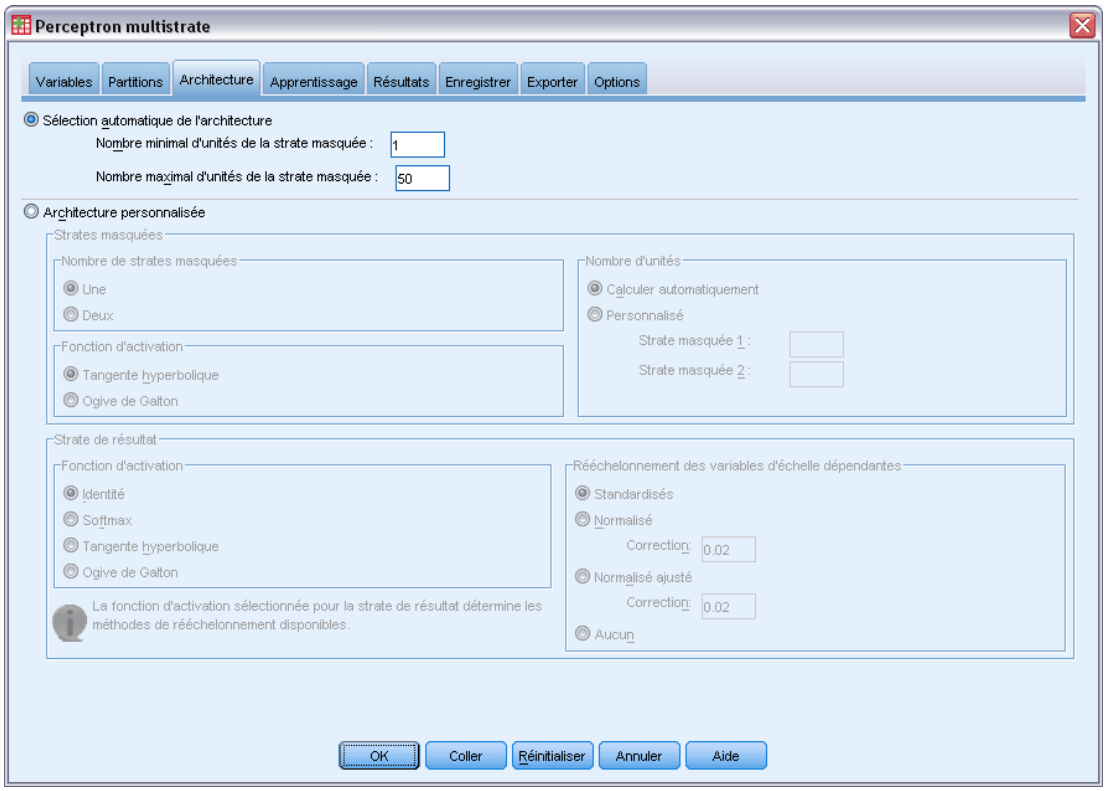
1 et 1 correspond à la division de l'ensemble de données en tiers égaux entre l'apprentissage, le test et l'élément traité.

- **Utiliser une variable de partitionnement pour affecter des observations.** Indiquez une variable numérique qui affecte chaque observation de l'ensemble de données actif à l'échantillon d'apprentissage, de test et traité. Les observations contenant une valeur positive sur la variable sont affectées à l'échantillon d'apprentissage, celles contenant une valeur égale à 0 sont affectées à l'échantillon de test, et celles contenant une valeur négative sont affectées à l'échantillon traité. Les observations contenant des valeurs manquantes sont exclues de l'analyse. Les valeurs manquantes spécifiées par l'utilisateur pour la variable de partitionnement sont toujours considérées comme étant valides.

  Remarque : L'utilisation d'une variable de partitionnement ne garantira pas des résultats identiques dans les exécutions successives de la procédure. Reportez-vous à la section "Réplication de résultats" de la rubrique principale Perceptron multistrate.

Architecture

Figure 2-4
Perceptron multistrate : Onglet Architecture.



L'onglet Architecture permet de spécifier la structure du réseau. La procédure peut sélectionner la « meilleure » architecture automatiquement ou vous pouvez spécifier une architecture personnalisée.

La sélection automatique de l'architecture construit un réseau avec une strate masquée. Spécifiez le nombre minimal et maximal d'unités autorisé dans la strate masquée pour que la sélection automatique de l'architecture calcule le « meilleur » nombre d'unités figurant dans la strate masquée. La sélection automatique de l'architecture utilise les fonctions d'activation par défaut pour les strates masquées et de résultat.

La sélection personnalisée de l'architecture vous permet de contrôler très précisément les strates masquées et de résultat et peut être très utile lorsque vous savez à l'avance quelle architecture vous souhaitez ou lorsque vous devez modeler les résultats de la sélection automatique d'architecture.

Strates masquées

La strate masquée contient des noeuds (unités) de réseau non observables. Chaque unité masquées est une fonction de la somme pondérée des entrées. La fonction des la fonction d'activation et les valeurs des pondérations sont déterminées par l'algorithme d'estimation. Si le réseau contient une seconde strate masquée, chaque unité masquée de la seconde strate est une fonction de la somme pondérée des unités de la première strate. La même fonction d'activation est utilisée dans les deux strates.

Nombre de strates masquées. Le perceptron multicouche peut avoir une ou deux couches cachées.

Fonction d'activation. La fonction d'activation "lie" les sommes pondérées des unités dans une couche aux valeurs des unités dans la couche réussie.

- **Tangente hyperbolique.** Cette fonction observe la forme suivante : $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Elle extrait les arguments de valeurs réelles et les transforme en plage (–1, 1). Lorsque la sélection automatique de l'architecture est utilisée, il s'agit de la fonction d'activation de toutes les unités dans les strates masquées.
- **Ogive de Galton.** Cette fonction observe la forme suivante : $\gamma(c) = 1/(1+e^{-c})$. Elle extrait les arguments de valeurs réelles et les transforme en plage (0, 1).

Nombre d'unités. Le nombre d'unités dans chaque strate peut être défini explicitement ou déterminé automatiquement par l'algorithme d'estimation.

Strate de résultat

La strate de résultat contient les variables (dépendantes) cible.

Fonction d'activation. La fonction d'activation "lie" les sommes pondérées des unités dans une couche aux valeurs des unités dans la couche réussie.

- **Identité :** Cette fonction observe la forme suivante : $\gamma(c) = c$. Elle extrait les arguments de valeurs réelles et les renvoie inchangés. Lorsque la sélection automatique de l'architecture est utilisée, il s'agit de la fonction d'activation des unités de la strate de résultat s'il existe des variables d'échelle dépendantes.
- **Softmax.** Cette fonction observe la forme suivante : $\gamma(c_k) = \exp(c_k)/\Sigma_j\exp(c_j)$. Cette fonction extrait un vecteur des arguments de valeur réels et le transforme en un vecteur dont les éléments sont compris dans la plage (0, 1) et ont pour somme 1. La fonction Softmax n'est disponible que si toutes les variables dépendantes sont des variables qualitatives. Lorsque la

sélection automatique de l'architecture est utilisée, il s'agit de la fonction d'activation des unités de la strate de résultat si toutes les variables dépendantes sont des variables qualitatives.

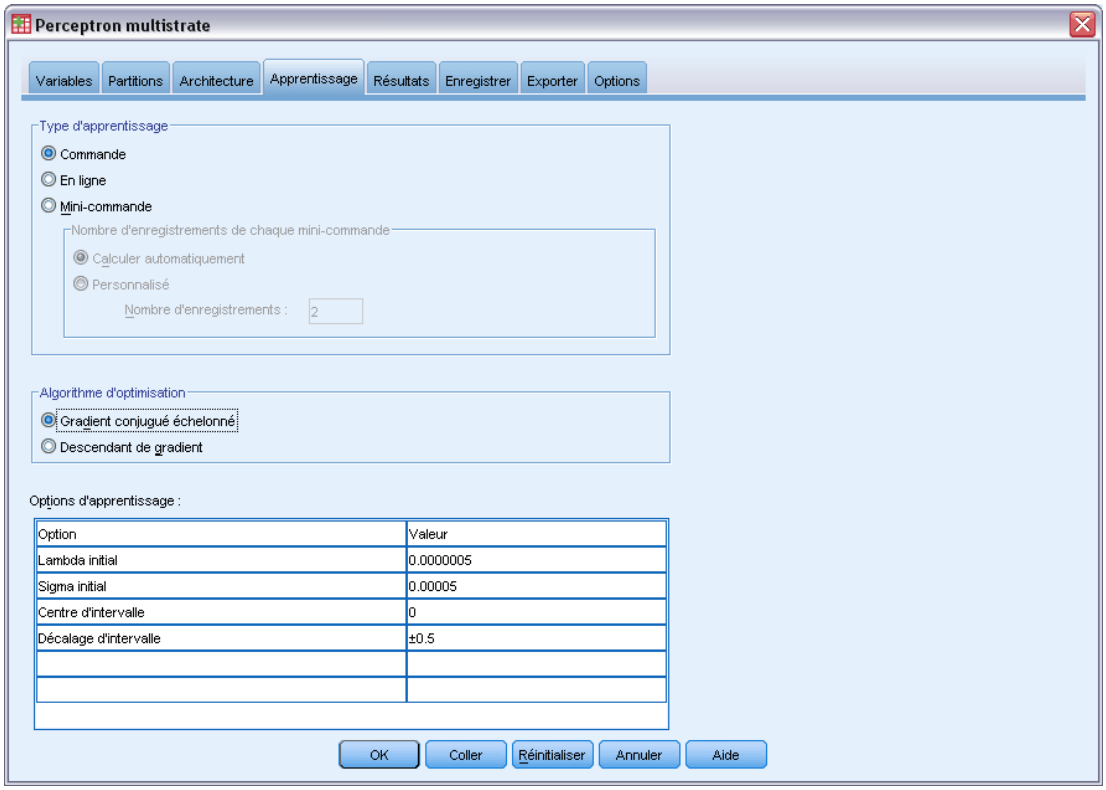
- **Tangente hyperbolique.** Cette fonction observe la forme suivante : $\gamma(c) = \tanh(c) = (e^{c}-e^{-c})/(e^{c}+e^{-c})$. Elle extrait les arguments de valeurs réelles et les transforme en plage (–1, 1).
- **Ogive de Galton.** Cette fonction observe la forme suivante : $\gamma(c) = 1/(1+e^{-c})$. Elle extrait les arguments de valeurs réelles et les transforme en plage (0, 1).

Rééchelonnement des variables d'échelle dépendantes. Ces contrôles ne sont disponibles que si une variable d'échelle dépendante au moins a été sélectionnée.

- **Standardisés.** Soustrayez la moyenne et divisez le résultat par l'écart-type, (x−moyenne)/s.
- **Normalisé.** Soustrayez le minimum et divisez le résultat par l'intervalle, (x−min)/(max−min). Les valeurs normalisées sont comprises entre 0 et 1. Il s'agit de la méthode de rééchelonnement requise pour les variables d'échelle dépendantes si la strate de résultat utilise la fonction d'activation d'ogive de Galton. L'option de correction spécifie un nombre ε qui est appliqué en tant que correction à la formule de rééchelonnement ; avec cette correction, toutes les valeurs de variable dépendante rééchelonnée se situeront dans la plage de la fonction d'activation. En particulier, les valeurs 0 et 1, qui sont présentes dans la formule non corrigée quand x prend sa valeur minimum et maximum, définissent les limites de la plage de la fonction d'ogive de Galton, mais ne sont pas comprises dans cette plage. La formule corrigée est [x−(min−ε)]/[(max+ε)−(min−ε)]. Spécifiez une valeur supérieure ou égale à 0.
- **Normalisé ajusté.** La version ajustée de la soustraction du minimum et de la division du résultat par l'intervalle, , [2*(x−min)/(max−min)]−1. Les valeurs normalisées ajustées sont comprises entre −1 et 1. Il s'agit de la méthode de rééchelonnement requise pour les variables d'échelle dépendantes si la strate de résultat utilise la fonction d'activation de tangente hyperbolique. L'option de correction spécifie un nombre ε qui est appliqué en tant que correction à la formule de rééchelonnement ; avec cette correction, toutes les valeurs de variable dépendante rééchelonnée se situeront dans la plage de la fonction d'activation. En particulier, les valeurs −1 et 1, qui sont présentes dans la formule non corrigée quand x prend sa valeur minimum et maximum, définissent les limites de la plage de la fonction de tangente hyperbolique, mais ne sont pas comprises dans cette plage. La formule corrigée est {2*[(x−(min−ε))/((max+ε)−(min−ε))]}−1. Spécifiez un nombre supérieur ou égal à 0.
- **Aucune.** Pas de rééchelonnement des variables d'échelle dépendantes.

Formations

Figure 2-5
Perceptron multistrate : Onglet formation



L'onglet Formation permet de spécifier la manière dont le réseau doit être formé. Le type de formation et l'algorithme d'optimisation déterminent les options de formation disponibles.

Type de formation. Le type de formation détermine la manière dont le réseau traite les enregistrements. Choisissez l'une des options de formation suivantes :

- **Commande.** Met à jour les pondérations synaptiques uniquement après avoir lu tous les enregistrements de données de formation ; en d'autres termes, la formation par commande utilise les informations issues de tous les enregistrements de l'ensemble de données de formation. La formation par commande est souvent préférée car elle minimise directement le nombre total d'erreurs ; cependant, elle peut nécessiter de mettre à jour les pondérations de nombreuses fois jusqu'à ce que l'une des règles d'arrêt soit observée, ce qui peut nécessiter de nombreuses lectures des données. Elle est très utile pour les petits ensembles de données.
- **En ligne.** Met à jour les pondérations synaptiques après chaque enregistrement de données de formation ; en d'autres termes, la formation en ligne utilise les informations issues d'un enregistrement à la fois. La formation en ligne extrait en permanence un enregistrement et met à jour les pondérations jusqu'à ce que l'une des règles d'arrêt soit observée. Si tous les enregistrements sont utilisés une fois et qu'aucune des règles d'arrêt n'est observée, le processus continue en recyclant les enregistrements de données. La formation en ligne est supérieure à la formation par commande pour les ensembles de données "volumineux" associés à des variables prédites ; en d'autres termes, s'il existe de nombreux enregistrements

et de nombreuses entrées et que leurs valeurs ne sont pas indépendantes les unes des autres, la formation en ligne peut obtenir plus rapidement une réponse raisonnable que la formation par commande.

- **Mini-commande.** Divise les enregistrements de données de formation en groupes de taille approximativement égale, puis met à jour les pondérations synaptiques après lecture d'un groupe ; en d'autres termes, la formation par mini-commande utilise les informations issues d'un groupe d'enregistrements. Le processus recycle ensuite le groupe de données si nécessaire. La formation par mini-commande offre un compromis entre les formations par commande et en ligne et peut être préférable pour les ensembles de données de taille moyenne. La procédure peut automatiquement déterminer le nombre d'enregistrements de formation par mini-commande ou vous pouvez spécifier un entier supérieur à 1 ou inférieur ou égal au nombre maximal d'observations à stocker en mémoire. Vous pouvez définir le nombre maximum d'observations à stocker en mémoire dans l'onglet Options.

Algorithme d'optimisation. Il s'agit de la méthode utilisée pour estimer les pondérations synaptiques.

- **Gradient conjugué échelonné.** Les hypothèses qui justifient l'utilisation des méthodes de gradient conjugué ne s'appliquant qu'aux types de formation par commande, cette méthode n'est pas disponible pour la formation en ligne ou par mini-commande.
- **Descendant de gradient.** Cette méthode doit être utilisée avec les formations en ligne et par mini-commande ; elle peut également être utilisée avec la formation par commande.

Options de formation. Les options de formation vous permettent d'affiner l'algorithme de formation. Vous n'aurez normalement pas besoin de modifier ces paramètres, sauf si le réseau rencontre des problèmes d'estimation.

Les options de formation de l'algorithme gradient conjugué redimensionné sont les suivantes :

- **Lambda initial.** Valeur initiale du paramètre lambda de l'algorithme gradient conjugué redimensionné. Spécifiez un nombre supérieur à 0 et inférieur à 0,000001.
- **Sigma initial.** Valeur initiale du paramètre sigma de l'algorithme gradient conjugué redimensionné. Spécifiez un nombre supérieur à 0 et inférieur à 0,0001.
- **Centre d'intervalle et décalage d'intervalle.** Le centre d'intervalle (a_0) et le décalage d'intervalle (a) define the interval $[a_0-a, a_0+a]$, dans lequel les vecteurs de pondération sont générés aléatoirement lorsque l'algorithme recuit stimulé est utilisé. L'algorithme recuit simulé est utilisé pour décomposer un minimum local, dans le but d'identifier le minimum global, pendant l'application de l'algorithme d'optimisation. Cette approche est utilisée dans l'initialisation de pondération et la sélection automatique de l'architecture. Spécifiez un nombre pour le centre d'intervalle et un nombre supérieur à 0 pour le décalage d'intervalle.

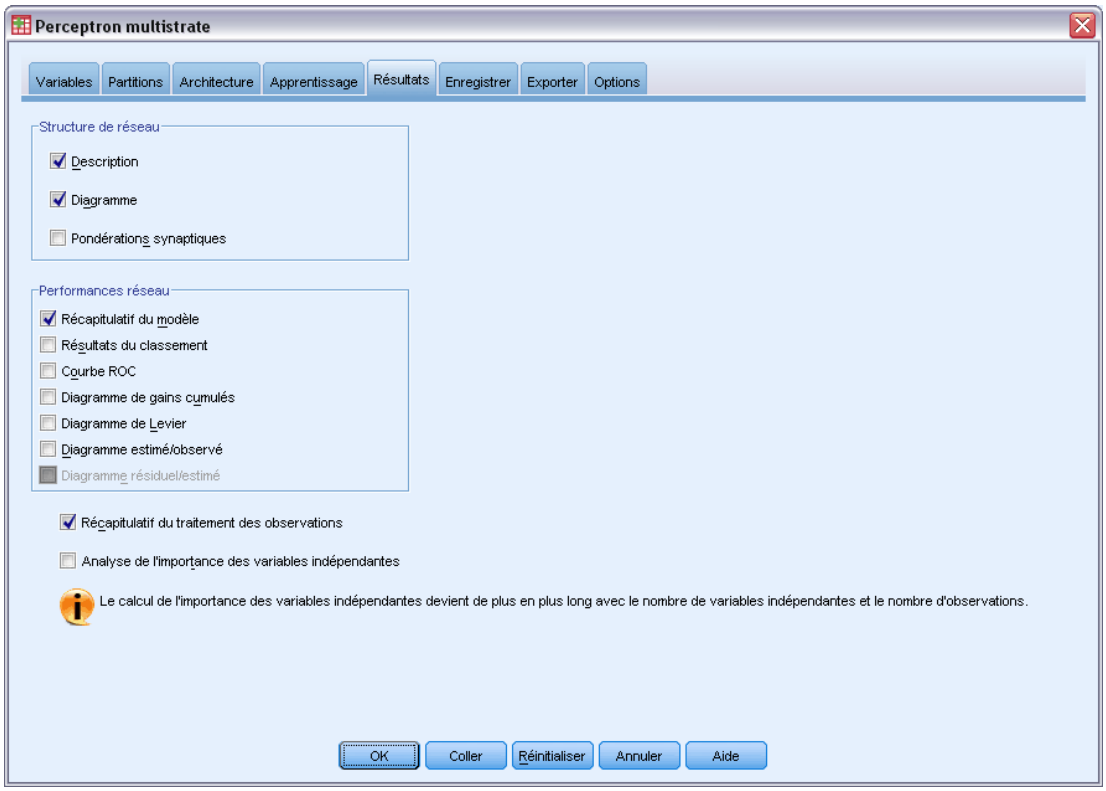
Les options de formation de l'algorithme descendant de gradient sont les suivantes :

- **Taux d'apprentissage initial.** Valeur initiale du taux d'apprentissage de l'algorithme descendant de gradient. Un taux d'apprentissage plus élevé signifie que le réseau se formera plus rapidement, au risque de devenir instable. Spécifiez un nombre supérieur à 0.

- **Limite inférieure du taux d'apprentissage.** Limite inférieure du taux d'apprentissage de l'algorithme descendant de gradient. Ce paramètre ne s'applique qu'aux formations en ligne et par mini-commande. Spécifiez un nombre supérieur à 0 et inférieur au taux d'apprentissage initial.
- **Vitesse.** Paramètre de vitesse initial de l'algorithme descendant de gradient. Ce paramètre permet d'empêcher les instabilités causées par un taux d'apprentissage trop élevé. Spécifiez un nombre supérieur à 0.
- **Réduction du taux d'apprentissage, par période.** Le nombre de périodes (p), ou lectures des données de l'échantillon de formation, pour réduire le taux d'apprentissage initial à la limite inférieure du taux d'apprentissage lorsque l'algorithme descendant de gradient est utilisé avec la formation en ligne ou par mini-commande. Vous pouvez ainsi contrôler le facteur de diminution du taux d'apprentissage $\beta = (1/pK)*\ln(\eta_0/\eta_{bas})$, η_0 étant le taux d'apprentissage initial, η_{bas} la limite inférieure du taux d'apprentissage et K le nombre total de mini-commandes (ou le nombre d'enregistrements de formation, pour la formation en ligne) dans l'ensemble de données de formation. Entrez un entier supérieur à 0.

Résultats

Figure 2-6
Perceptron multistrate : Onglet Résultats



Structure de réseau. Affiche des informations récapitulatives sur le réseau neuronal.

- **Description :** Affiche des informations sur le réseau neuronal, y compris les variables dépendantes, le nombre d'unités d'entrée et de sortie, le nombre de strates et d'unités masquées, ainsi que les fonctions d'activation.
- **Diagramme.** Affiche le diagramme de réseau sous forme de diagramme non modifiable. A mesure que le nombre de covariables et de niveaux de facteur augmente, le diagramme devient plus difficile à interpréter.
- **Pondérations synaptiques.** Affiche les estimations de coefficients qui indiquent la relation existant entre les unités d'une strate donnée et celles de la strate suivante. Les pondérations synaptiques sont basées sur l'échantillon de formation même si l'ensemble de données actif est partitionné en données de formation, de test et traitées. Le nombre de pondérations synaptiques peut être élevé et ces pondérations ne sont généralement pas utilisées pour interpréter les résultats du réseau.

Performances réseau. Affiche les résultats utilisés pour déterminer si le modèle est correct. *Remarque* : Les diagrammes figurant dans ce groupe sont basés sur les échantillons de formation et de test combinés, ou uniquement sur l'échantillon de formation s'il n'existe aucun échantillon de test.

- **Récapitulatif du modèle.** Affiche un récapitulatif des résultats du réseau neuronal par partition et globalement, y compris les erreurs, les erreurs ou les pourcentages relatifs de prévisions incorrectes, la règle d'arrêt utilisée pour arrêter la formation et le temps de formation.

  L'erreur est l'erreur de somme des carrés lorsque la fonction d'activation d'identité, d'ogive de Galton ou de tangente hyperbolique est appliquée à la strate de résultat. Il s'agit de l'erreur d'entropie croisée lorsque la fonction d'activation de Softmax est appliquée à la strate de résultat.

  Les erreurs ou les pourcentages relatifs de prévisions incorrectes sont affichés en fonction des niveaux de mesure de variable dépendante. Si une variable dépendante comporte un niveau de mesure d'échelle, l'erreur relative globale moyenne (par rapport au modèle moyen) est affichée. Si toutes les variables dépendantes sont des variables qualitatives, le pourcentage moyen de prévisions incorrectes est affiché. Les erreurs ou les pourcentages relatifs de prévisions incorrectes sont également affichés pour les variables dépendantes individuelles.
- **Résultats du classement.** Affiche un tableau de classement pour chaque variable dépendante qualitative par partition et globalement. Chaque tableau indique le nombre d'observations classées correctement et incorrectement pour chaque modalité de variable dépendante. Le pourcentage d'observations totales ayant été correctement classées est également indiqué.
- **Courbe ROC** Affiche une courbe ROC (Receiver Operating Characteristic) pour chaque variable dépendante qualitative. Affiche également un tableau indiquant la zone au-dessous de chaque courbe. Pour une variable dépendante donnée, le diagramme ROC affiche une courbe pour chaque modalité. Si la variable dépendante comporte deux modalités, chaque courbe traite la modalité en question comme étant l'état positif par rapport à l'autre modalité. Si la variable dépendante comporte plus de deux modalités, chaque courbe traite la modalité en question comme étant l'état positif par rapport à la somme de toutes les autres modalités.
- **Diagramme de gains cumulés.** Affiche un diagramme de gains cumulés pour chaque variable dépendante qualitative. L'affichage d'une courbe pour chaque modalité de variable dépendante est identique à celui des courbes ROC.

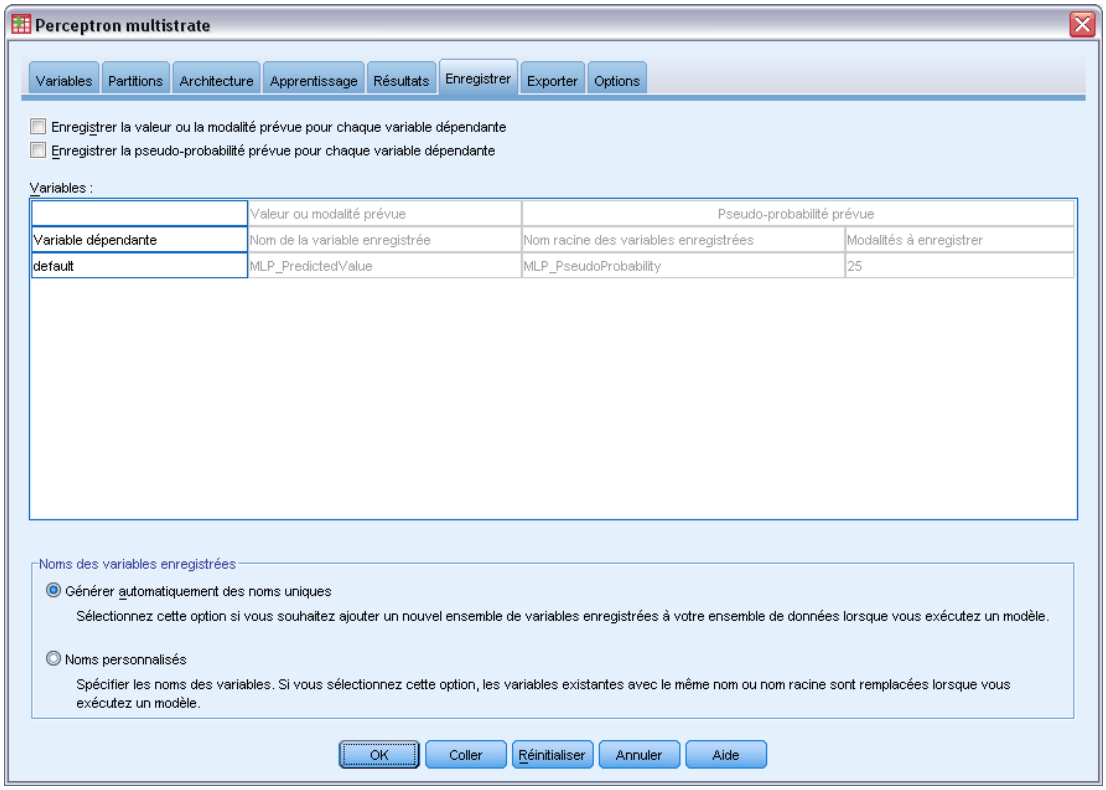
- **Diagramme de Levier.** Affiche un diagramme de levier pour chaque variable dépendante qualitative. L'affichage d'une courbe pour chaque modalité de variable dépendante est identique à celui des courbes ROC.
- **Diagramme estimé/observé.** Affiche un diagramme estimé/observé pour chaque variable dépendante. Pour les variables dépendantes qualitatives, des boîtes à moustaches juxtaposées des pseudo-probabilités prévues sont affichées pour chaque modalité de réponse, avec la modalité de réponse observée comme variable de classe. Pour les variables d'échelle dépendantes, un diagramme de dispersion est affiché.
- **Diagramme résiduel/estimé.** Affiche un diagramme résiduel/estimé pour chaque variable d'échelle dépendante. Il ne doit exister aucun schéma visible entre les résidus et les prévisions. Ce diagramme n'est généré que pour les variables d'échelle dépendantes.

Récapitulatif du traitement des observations. Affiche le tableau récapitulatif de traitement des observations, qui récapitule le nombre d'observations incluses et exclues dans l'analyse, au total et par échantillon de formation, de test et traité.

Analyse de l'importance des variables prédites. Effectue une analyse de sensibilité, qui calcule l'importance de chaque variable prédite dans la détermination du réseau neuronal. L'analyse est basée sur les échantillons de formation et de test combinés, ou uniquement sur l'échantillon de formation s'il n'existe aucun échantillon de test. Ceci produit un tableau et un diagramme qui indiquent l'importance et l'importance normalisée de chaque variable prédite. L'analyse de sensibilité nécessite beaucoup de calculs et de temps si les variables prédites ou les observations sont nombreuses.

Enregistrer

Figure 2-7
Perceptron multistrate : Onglet Enregistrer



L'onglet Enregistrer permet d'enregistrer les prévisions en tant que variables dans l'ensemble de données.

- **Enregistrer la valeur ou la modalité prévue pour chaque variable dépendante.** Cette option enregistre la valeur prévue pour les variables d'échelle dépendantes et la modalité prévue pour les variables dépendantes qualitatives.
- **Enregistrer la pseudo-probabilité prévue ou la catégorie pour chaque variable dépendante.** Cette option enregistre les pseudo-probabilités prévues pour les variables dépendantes qualitatives. Une variable distincte est enregistrée pour chacune des *n* premières modalités, *n* étant spécifié dans la colonne Modalités à enregistrer.

Noms des variables enregistrées. Grâce à la génération automatique de nom, vous conservez l'ensemble de votre travail. Les noms personnalisés vous permettent de supprimer/remplacer les résultats d'exécutions précédentes sans supprimer d'abord les variables enregistrées dans l'éditeur de données.

Probabilités et pseudo-probabilités

Les variables dépendantes qualitatives présentant une erreur d'activation softmax et d'entropie croisée comporteront une valeur prévue pour chaque modalité, chaque valeur prévue étant la probabilité que l'observation appartienne à la modalité.

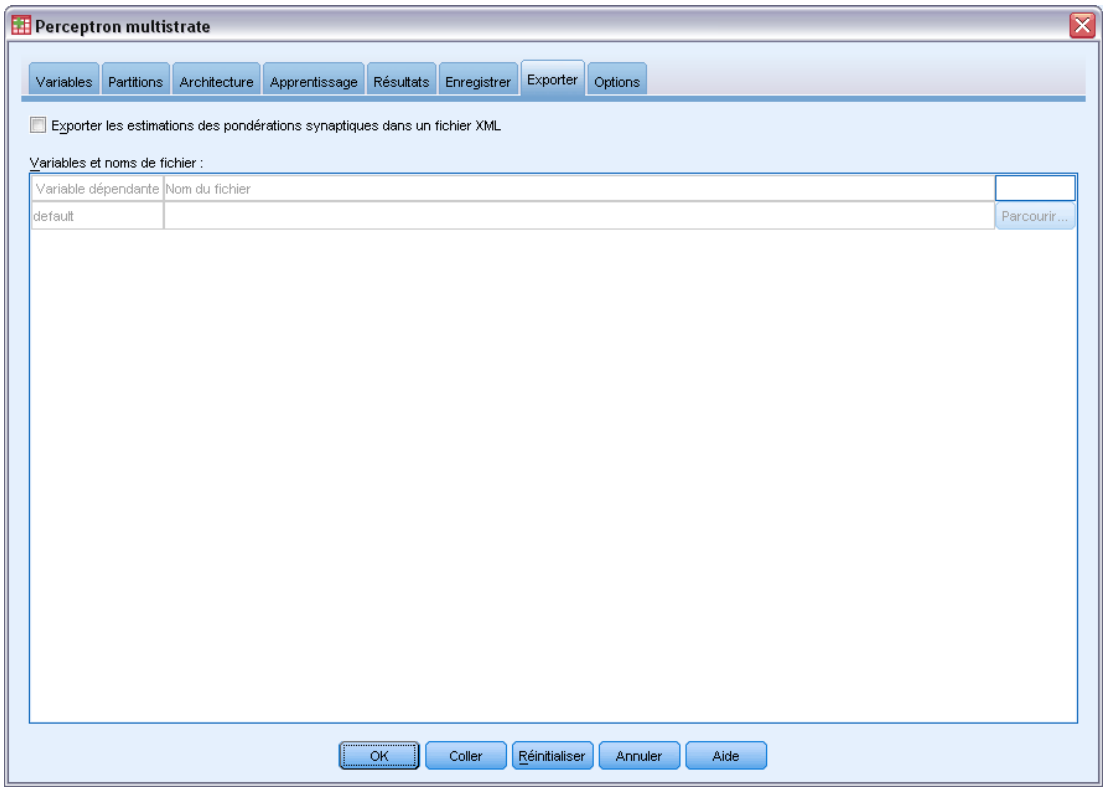
Les variables dépendantes qualitatives présentant une erreur de somme des carrés comporteront une valeur prévue pour chaque modalité, mais les valeurs prévues ne peuvent pas être interprétées comme probabilités. La procédure enregistre ces pseudo-probabilités prévues même si certaines d'entre elles sont inférieures à zéro ou supérieures à 1, ou si la somme d'une variable dépendante donnée n'est pas égale à 1.

Le diagramme de ROC, des gains cumulés et de Levier (reportez-vous Résultats sur p. 15) sont créés en fonction des pseudo-probabilités. Si des pseudo-probabilités sont inférieures à 0 ou supérieures à 1 ou que la somme d'une variable donnée n'est pas égale à 1, elles sont d'abord rééchelonnées pour se situer entre 0 et 1, et avoir pour somme 1. Les pseudo-probabilités sont rééchelonnées en étant divisées par leur somme. Par exemple, si une observation comporte des pseudo-probabilités de 0,50, 0,60 et 0,40 pour une variable dépendante à trois modalités, chaque pseudo-probabilité est alors divisée par la somme 1,50 afin d'obtenir 0,33, 0,40 et 0.27.

Si des pseudo-probabilités sont négatives, la valeur absolue de la plus faible est ajoutée à toutes les pseudo-probabilités avant le rééchelonnement ci-dessus. Par exemple, si les pseudo-probabilités sont -0,30, 0,50, et 1,30, ajoutez d'abord 0,30 à chaque valeur pour obtenir 0,00, 0,80 et 1,60. Divisez ensuite chaque nouvelle valeur par la somme 2,40 pour obtenir 0,00, 0,33 et 0,67.

Exporter

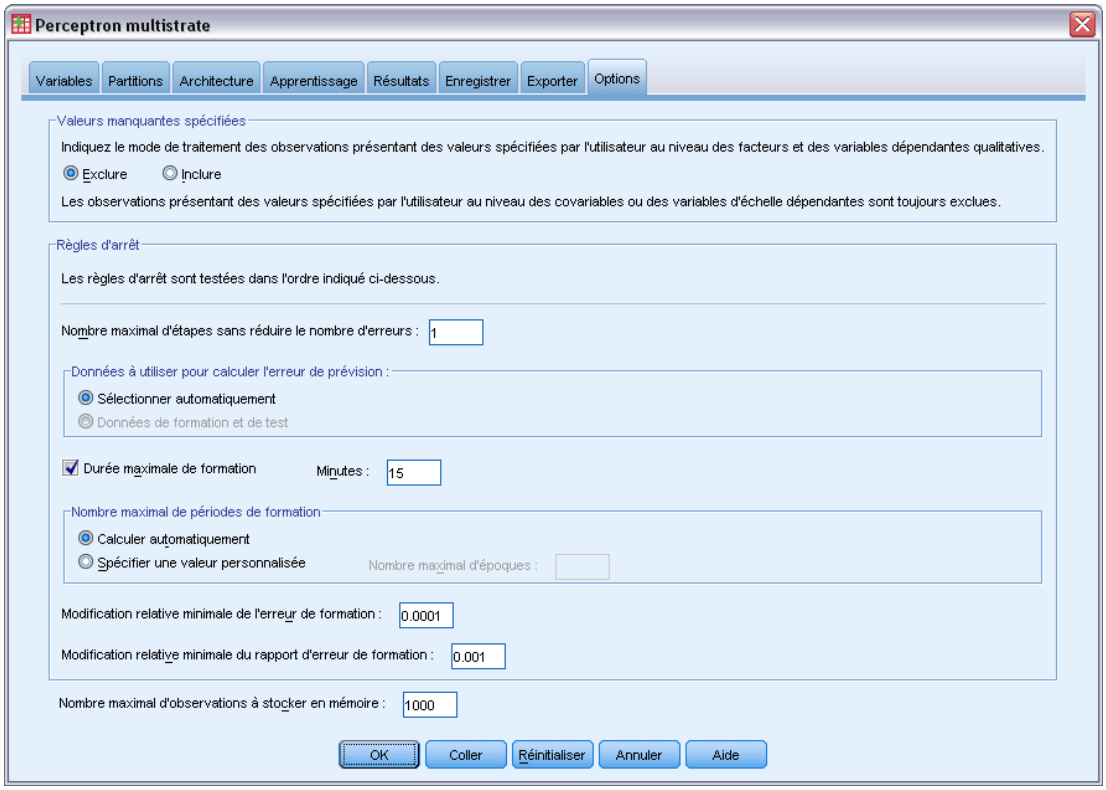
Figure 2-8
Perceptron multistrate : Onglet Exporter



L'onglet Exporter permet d'enregistrer les estimations des pondérations synaptiques de chaque variable dépendante dans un fichier XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation. Cette option n'est pas disponible si des fichiers scindés ont été définis.

Options

Figure 2-9
Perceptron multistrate : Onglet Options



Valeurs manquantes spécifiées. Les facteurs doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les facteurs et les variables dépendantes qualitatives.

Règles d'arrêt. Ces règles déterminent le moment où la formation du réseau neuronal doit être arrêtée. La formation se poursuit avec au moins une lecture des données. La formation peut être arrêtée en fonction des critères suivants, qui sont sélectionnés dans l'ordre indiqué. Dans les définitions de règle d'arrêt qui suivent, une étape correspond à une lecture des données pour les méthodes en ligne et par mini-commande, ainsi qu'à une itération pour la méthode par commande.

- **Nombre maximal d'étapes sans réduire le nombre d'erreurs.** Nombre d'étapes à autoriser avant de vérifier une baisse du nombre d'erreurs. Si le nombre d'erreurs ne diminue pas après le nombre spécifié d'étapes, la formation s'arrête. Spécifiez un entier supérieur à 0. Vous pouvez également spécifier l'échantillon de données utilisé pour calculer les erreurs. L'option Sélectionner automatiquement utilise l'échantillon de test s'il existe et, dans le cas contraire, l'échantillon de formation. La formation par commande garantit la réduction du nombre d'erreurs d'échantillon de formation après chaque lecture des données. Par conséquent, cette option ne s'applique qu'à la formation par commande s'il existe un échantillon de test. L'option Données de formation et de test vérifie les erreurs pour chacun de ces échantillons ; cette option ne s'applique que s'il existe un échantillon de test.

Remarque : Après chaque lecture complète des données, les formations en ligne et par mini-commande nécessitent une lecture supplémentaire des données pour calculer l'erreur de formation. Cette lecture supplémentaire des données pouvant ralentir considérablement la formation, il est généralement recommandé de fournir un échantillon de test et de sélectionner Sélectionner automatiquement dans tous les cas.

- **Durée maximale de formation.** Choisissez de spécifier ou non un nombre maximal de minutes pour l'exécution de l'algorithme. Spécifiez un nombre supérieur à 0.
- **Nombre maximal de périodes de formation.** Nombre maximal de périodes (lectures des données) autorisé. Si le nombre maximal de périodes est dépassé, la formation s'arrête. Entrez un entier supérieur à 0.
- **Modification relative minimale de l'erreur de formation.** La formation s'arrête si le changement relatif dans les erreurs de formation par rapport à l'étape précédente est inférieur à la valeur de critère. Spécifiez un nombre supérieur à 0. Pour les formations en ligne et par mini-commande, ce critère est ignoré si les données de test sont les seules à être utilisées pour calculer les erreurs.
- **Modification relative minimale du rapport d'erreur de formation.** La formation s'arrête si le rapport entre erreurs de formation et erreurs du modèle nul est inférieur à la valeur de critère. Le modèle nul prévoit la valeur moyenne de toutes les variables dépendantes. Spécifiez un nombre supérieur à 0. Pour les formations en ligne et par mini-commande, ce critère est ignoré si les données de test sont les seules à être utilisées pour calculer les erreurs.

Nombre maximal d'observations à stocker en mémoire. Cette option contrôle les paramètres suivants dans les algorithmes de perceptron multistrate. Entrez un entier supérieur à 1.

- Lors de la sélection automatique de l'architecture, la taille de l'échantillon permet de déterminer si la taille de l'architecture réseau est min(1000,*memsize*), *memsize* représentant le nombre maximal d'observations à stocker en mémoire.
- Lors de la formation par mini-commande avec calcul automatique du nombre de mini-commandes, le nombre de mini-commandes est min(max(*M*/10,2),*memsize*), *M* représentant le nombre d'observations de l'échantillon de formation.

Chapitre 3

Fonction à base radiale

La procédure de fonction à base radiale (RBF) produit un modèle de prévision pour une ou plusieurs variables dépendantes (cibles) en fonction des valeurs des variables prédites.

Exemple : Un fournisseur de services de télécommunication a segmenté sa base de clients par type d'utilisation des services en catégorisant les clients en quatre groupes. Un réseau RBF utilisant des données démographiques pour prévoir les groupes d'affectations permet à l'entreprise de personnaliser les offres pour chaque client éventuel.

Variables dépendantes Les variables dépendantes peuvent être :

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

  La procédure considère que le niveau de mesure approprié a été assigné à toutes les variables dépendantes, bien que vous puissiez changer provisoirement le niveau de mesure d'une variable en cliquant avec le bouton droit de la souris sur la variable dans la liste des variables source, puis en sélectionnant un niveau de mesure dans le menu contextuel.

Dans la liste des variables, une icône indique le niveau de mesure et le type de données :

	Numérique	Chaîne	Date	Heure
Echelle (continue).		n/a		
Ordinal				
Nominal				

Variables prédites. Les variables prédites peuvent être spécifiées en tant que facteurs (qualitatifs) ou covariables (d'échelle).

Codage des variables indicatrices. La procédure recode provisoirement les variables prédites qualitatives et les variables dépendantes via le codage un-de-*c* pour la durée de la procédure. S'il existe des modalités *c* d'une variable, la variable est stockée comme vecteurs c, la première modalité étant identifiée par (1,0,...,0), la suivante par (0,1,0,...,0), ... et la dernière par (0,0,...,0,1).

Ce système de codage augmente le nombre de pondérations synaptiques et peut résulter en une formation plus lente, mais les méthodes de codage plus compactes aboutissent généralement à des réseaux neuronaux mal ajustés. Si la formation de votre réseau s'effectue très lentement, essayez de réduire le nombre de modalités dans vos variables prédites qualitatives en combinant des modalités similaires ou en supprimant les observations comportant des modalités extrêmement rares.

Les codage un-de-*c* repose entièrement sur les données de formation, même si un échantillon de test ou traité est défini (reportez-vous à Partitions sur p. 27). Ainsi, si les échantillons de test ou traités contiennent des observations avec des modalités de variables prédites ne figurant pas dans les données de formation, ces observations ne sont pas utilisées par la procédure ou dans l'évaluation. Si les échantillons de test ou traités contiennent des observations avec des modalités de variables dépendantes ne figurant pas dans les données de formation, ces observations ne sont pas utilisées par la procédure mais peuvent être notées.

Rééchelonnement. Les variables d'échelle dépendantes et les covariables sont rééchelonnées par défaut pour améliorer la formation du réseau. Le rééchelonnement repose entièrement sur les données de formation, même si un échantillon de test ou traité est défini (reportez-vous à Partitions sur p. 27). En d'autres termes, en fonction du type de rééchelonnement, la moyenne, l'écart-type, la valeur minimale ou la valeur maximale d'une covariable ou d'une variable dépendante ne sont calculées qu'à l'aide des données de formation. Si vous spécifiez une variable pour définir des partitions, il est important que ces covariables ou variables dépendantes présentent des distributions similaires à travers les échantillons de formation, de test et traités.

Pondérations d'effectif. Cette procédure ignore les pondérations d'effectif.

Réplication de résultats. Si vous souhaitez répliquer vos résultats exactement, outre les mêmes paramètres de procédure, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires et le même ordre de données. Vous trouverez ci-après plus de détails sur cet aspect.

- **Génération de nombres aléatoires.** La procédure utilise la génération de nombres aléatoires pendant l'attribution aléatoire des partitions. Pour reproduire les mêmes résultats aléatoires à l'avenir, utilisez la même valeur d'initialisation pour le générateur de nombres aléatoires avant chaque exécution de la procédure de fonction à base radiale. Reportez-vous à Préparation des données pour l'analyse sur p. 74 pour des instructions détaillées.
- **Tri par observation.** Les résultats dépendent également de l'ordre des données, car l'algorithme de classification en deux étapes intervient dans la détermination des fonctions à base radiale.

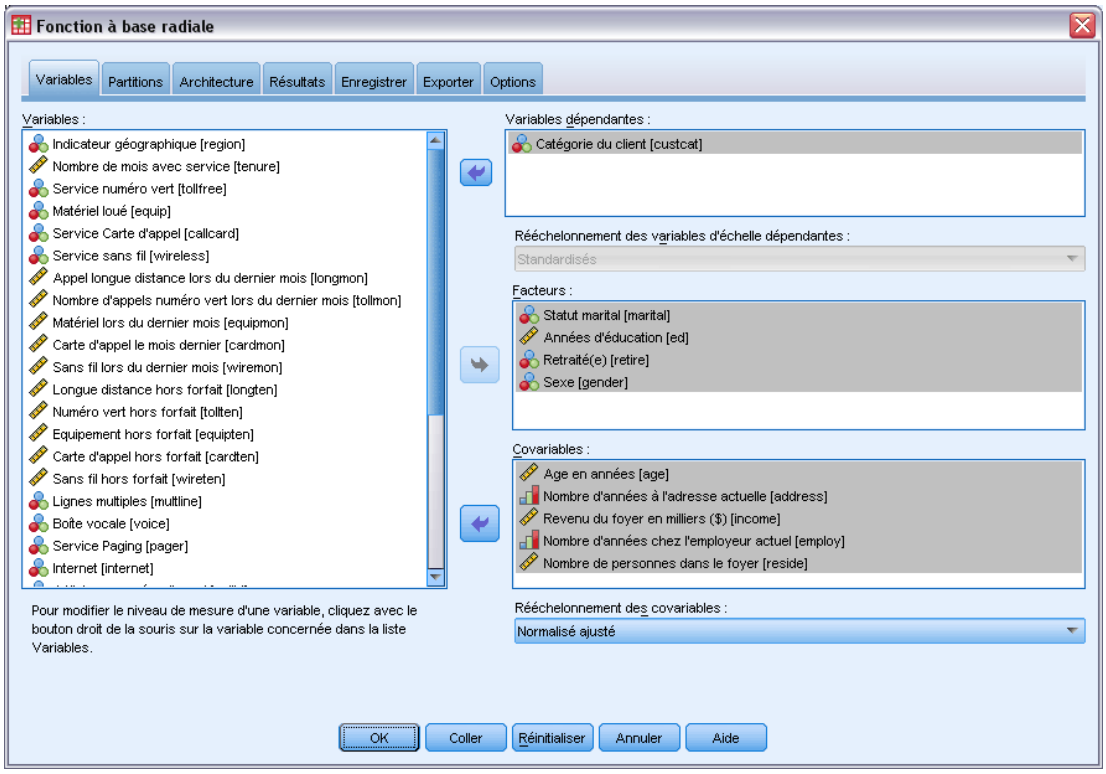
  Pour réduire les effets de tri, classez les observations de manière aléatoire. Pour vérifier la stabilité d'une solution donnée, vous pouvez obtenir différentes solutions dans lesquelles les observations sont triées de différentes manières aléatoires. Si les fichiers sont très volumineux, vous pouvez effectuer plusieurs fois l'opération sur un échantillon des observations triées de différentes manières aléatoires.

Création d'un réseau de fonction à base radiale

A partir des menus, sélectionnez :

Analyse > Réseaux neuronaux > Fonction à base radiale...

Figure 3-1
Fonction à base radiale : l'onglet Variables



- ▶ Sélectionnez au moins une variable dépendante.
- ▶ Sélectionnez au moins un facteur ou une covariable.

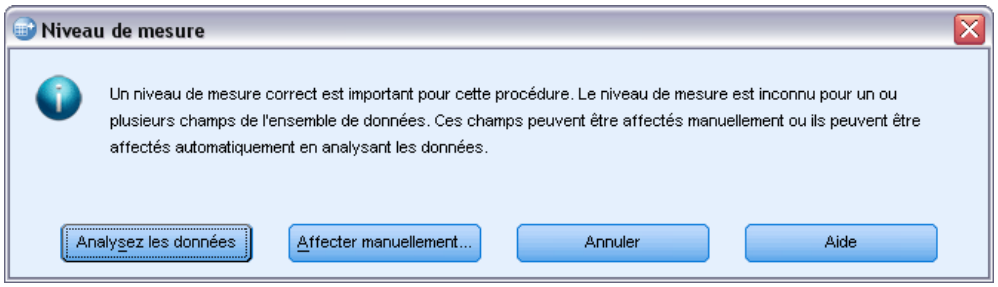
Dans l'onglet Variables, vous pouvez aussi changer la méthode de rééchelonnement des covariables. Les choix sont les suivants :

- **Standardisés.** Soustrayez la moyenne et divisez le résultat par l'écart-type, (x−moyenne)/s.
- **Normalisé.** Soustrayez le minimum et divisez le résultat par l'intervalle, (x−min)/(max−min). Les valeurs normalisées sont comprises entre 0 et 1.
- **Normalisé ajusté.** Version ajustée de la soustraction du minimum et de la division du résultat par l'intervalle, [2*(x−min)/(max−min)]−1. Les valeurs normalisées ajustées sont comprises entre −1 et 1.
- **Aucune.** Aucun rééchelonnement des covariables.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 3-2
Alerte du niveau de mesure

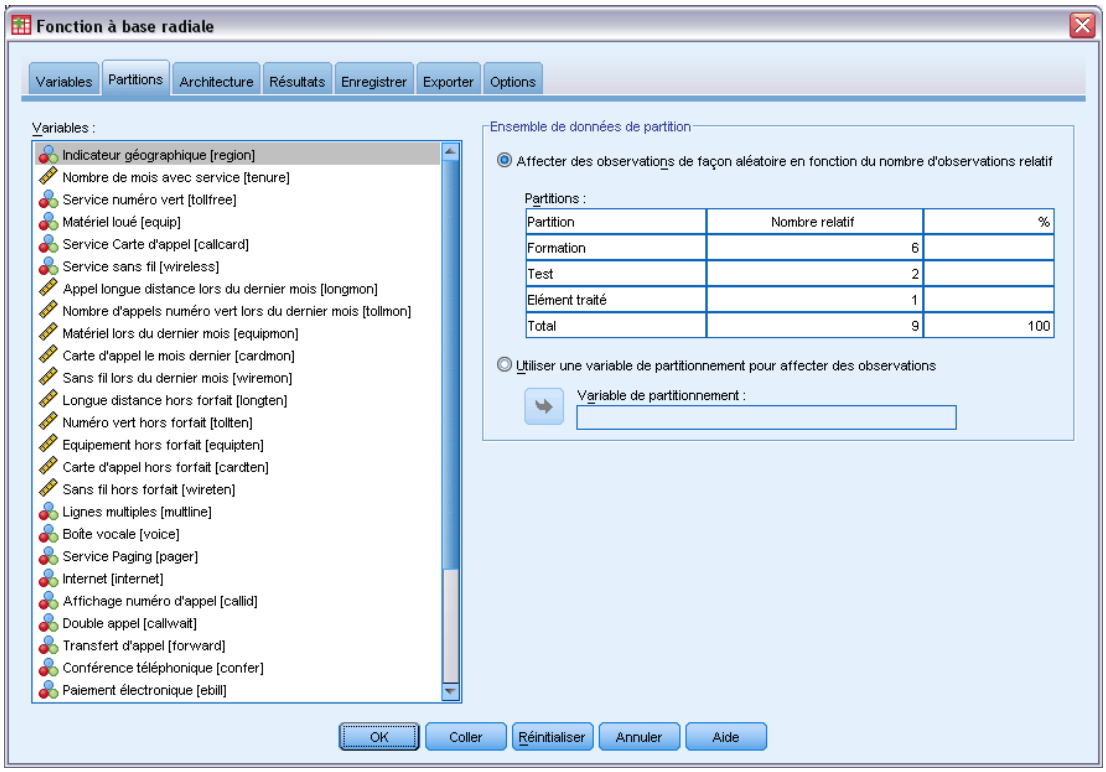


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Partitions

Figure 3-3
Fonction à base radiale : Onglet Partitions



Ensemble de données de partition. Ce groupe indique la méthode de partitionnement de l'ensemble de données actif en échantillons d'apprentissage, de test et traité. L'**échantillon d'apprentissage** comprend les enregistrements de données utilisés pour former le réseau neuronal. Un certain pourcentage d'observations contenues dans l'ensemble de données doit être affecté à l'échantillon d'apprentissage pour l'obtention d'un modèle. L'**échantillon de test** est un ensemble indépendant d'enregistrements de données utilisé pour identifier des erreurs au cours de la formation afin d'empêcher une formation excessive. Nous vous conseillons fortement de créer un échantillon d'apprentissage. Une formation de réseau sera en général plus efficace si l'échantillon de test est plus petit que l'échantillon de fonction. L'**échantillon traité** est un autre ensemble indépendant d'enregistrements de données utilisé pour évaluer le réseau neuronal final ; l'erreur pour l'échantillon traité donne une estimation « honnête » de la capacité de prévision du modèle parce que les observations traitées n'ont pas été utilisées pour construire le modèle.

- **Affecter des observations de façon aléatoire en fonction du nombre d'observations relatif.** Indiquez le nombre d'observations relatif (ratio) affecté de façon aléatoire à chaque échantillon (d'apprentissage, de test et traité). La colonne % rapporte le pourcentage d'observations qui seront affectées à chaque échantillon en fonction des nombres relatifs que vous avez spécifiés.

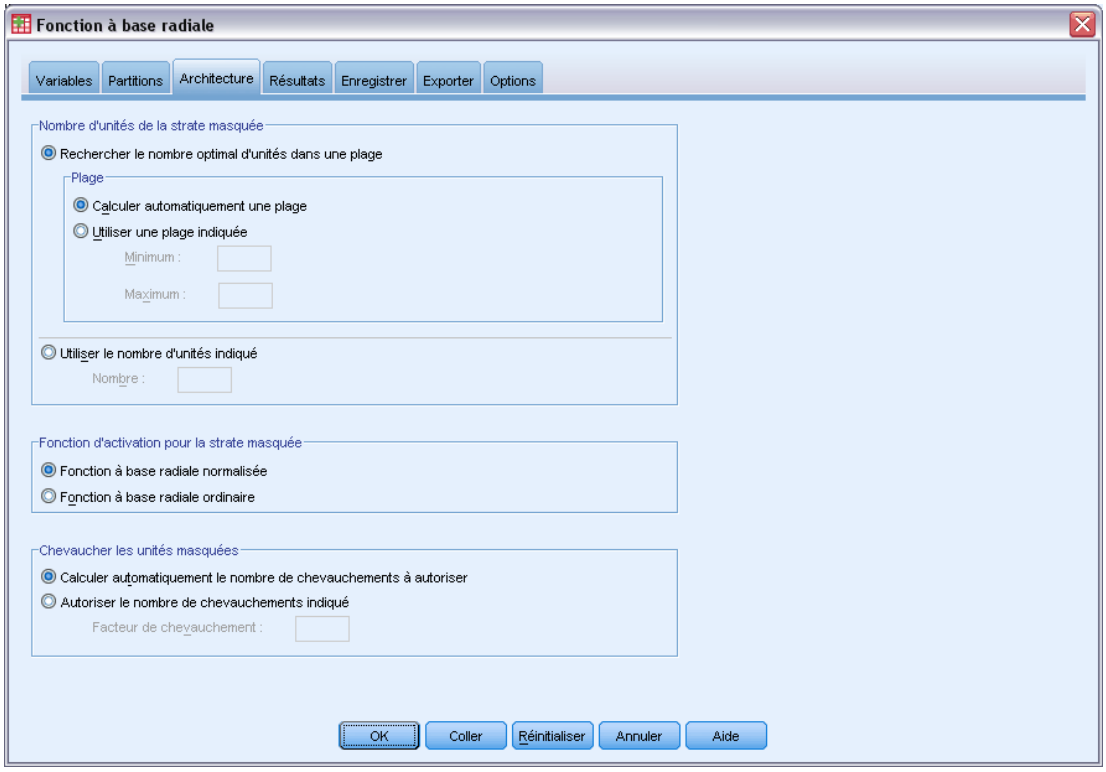
  Par exemple, si vous spécifiez 7, 3 et 0 comme nombres relatifs pour les échantillons d'apprentissage, de test et traité, ces valeurs correspondent à 70 %, 30 % et 0 %. Si vous spécifiez 2, 1, 1 comme nombres relatifs, ces valeurs correspondent à 50 %, 25 % et 25 %. 1,

1 et 1 correspond à la division de l'ensemble de données en tiers égaux entre l'apprentissage, le test et l'élément traité.

- **Utiliser une variable de partitionnement pour affecter des observations.** Indiquez une variable numérique qui affecte chaque observation de l'ensemble de données actif à l'échantillon d'apprentissage, de test et traité. Les observations contenant une valeur positive sur la variable sont affectées à l'échantillon d'apprentissage, celles contenant une valeur égale à 0 sont affectées à l'échantillon de test, et celles contenant une valeur négative sont affectées à l'échantillon traité. Les observations contenant des valeurs manquantes sont exclues de l'analyse. Les valeurs manquantes spécifiées par l'utilisateur pour la variable de partitionnement sont toujours considérées comme étant valides.

Architecture

Figure 3-4
Fonction à base radiale : Onglet Architecture.



L'onglet Architecture permet de spécifier la structure du réseau. La procédure crée un réseau neuronal avec une strate de « fonction à base radiale » masquée ; en général, il n'est pas nécessaire de modifier ces paramètres.

Nombre d'unités de la strate masquée. Vous pouvez choisir le nombre d'unités masquées de trois façons.

1. **Rechercher le nombre optimal d'unités dans une plage calculée automatiquement.** La procédure calcule automatiquement les valeurs minimale et maximale de la plage et recherche le nombre optimal d'unités masquées à l'intérieur de la plage.

   Si un échantillon de test est défini, la procédure utilise le critère de données de test : Le nombre optimal d'unités masquées est celui qui génère la plus petite erreur dans les données de test. Si aucun échantillon de test n'est défini, la procédure utilise le critère d'information bayésien (BIC) : Le nombre optimal d'unités masquées est celui qui génère le plus petit critère d'information bayésien dans les données de formation.

2. **Rechercher le nombre optimal d'unités dans une plage spécifique.** Vous pouvez indiquer la plage de votre choix afin que la procédure y recherche le nombre « optimal » d'unités masquées. Comme dans la méthode précédente, le nombre optimal d'unités masquées dans la plage est déterminé à l'aide du critère de données de test ou du critère d'information bayésien.

3. **Utiliser le nombre d'unités indiqué.** Vous pouvez passer outre à l'utilisation d'une plage et indiquer directement un nombre d'unités spécifique.

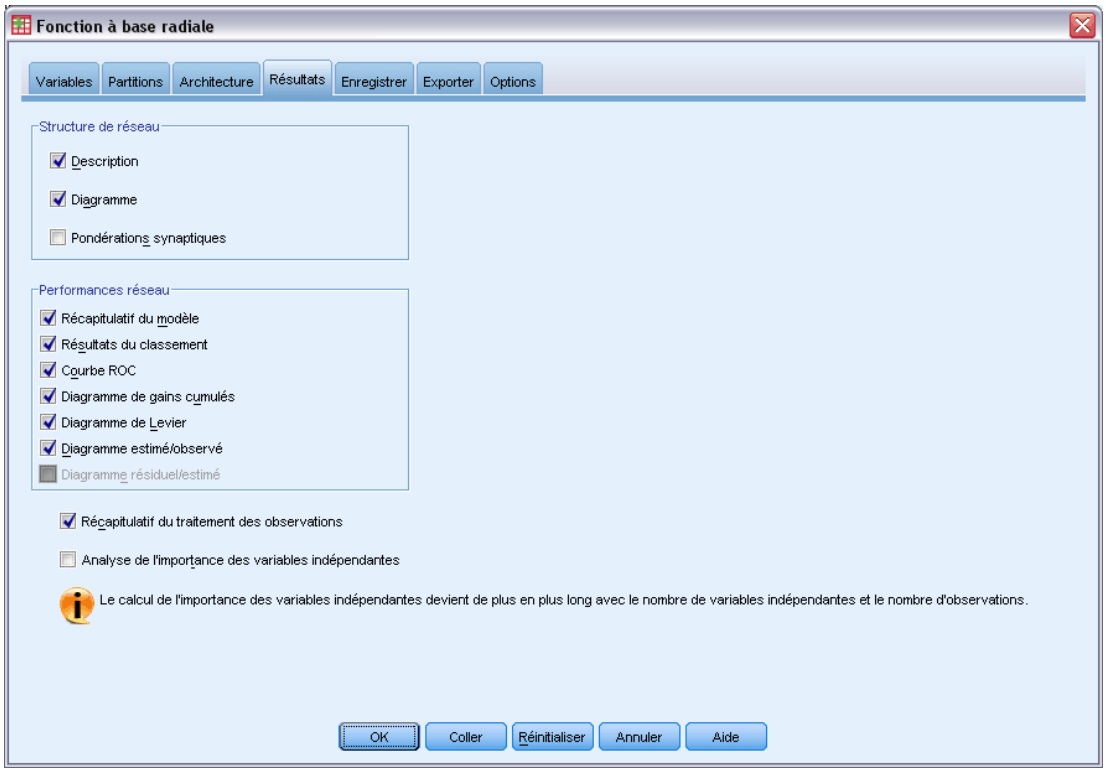
Fonction d'activation pour la strate masquée. La fonction d'activation pour la strate masquée est la fonction à base radiale, qui « lie » les unités d'une strate aux valeurs des unités de la suivante. Pour la strate de résultat, la fonction d'activation est la fonction d'identité ; les unités de résultat sont donc simplement les sommes pondérées des unités masquées.

- **Fonction à base radiale normalisée.** Utilise la fonction d'activation softmax afin que les activations de toutes les unités masquées soient normalisées pour être égales à un.
- **Fonction à base radiale ordinaire.** Utilise la fonction d'activation exponentielle afin que l'activation de l'unité masquée soit une « bosse » gaussienne en guise de fonction des entrées.

Chevaucher les unités masquées. Le facteur de chevauchement est un multiplicateur appliqué à la largeur des fonctions à base radiale. Valeur automatiquement calculée du facteur de chevauchement $1+0,1d$, où d représente le nombre d'unités d'entrée (somme du nombre de modalités dans tous les facteurs et du nombre de covariables).

Résultat

Figure 3-5
Fonction à base radiale : Onglet Résultats



Structure de réseau. Affiche des informations récapitulatives sur le réseau neuronal.

- **Description :** Affiche des informations sur le réseau neuronal, y compris les variables dépendantes, le nombre d'unités d'entrée et de sortie, le nombre de strates et d'unités masquées, ainsi que les fonctions d'activation.
- **Diagramme.** Affiche le diagramme de réseau sous forme de diagramme non modifiable. A mesure que le nombre de covariables et de niveaux de facteur augmente, le diagramme devient plus difficile à interpréter.
- **Pondérations synaptiques.** Affiche les estimations de coefficients qui indiquent la relation existant entre les unités d'une strate donnée et celles de la strate suivante. Les pondérations synaptiques sont basées sur l'échantillon de formation même si l'ensemble de données actif est partitionné en données de formation, de test et traitées. Le nombre de pondérations synaptiques peut être élevé et ces pondérations ne sont généralement pas utilisées pour interpréter les résultats du réseau.

Performances réseau. Affiche les résultats utilisés pour déterminer si le modèle est correct. *Remarque* : Les diagrammes figurant dans ce groupe sont basés sur les échantillons de formation et de test combinés, ou uniquement sur l'échantillon de formation s'il n'existe aucun échantillon de test.

- **Récapitulatif du modèle.** Affiche un récapitulatif des résultats du réseau neuronal par partition et globalement, y compris l'erreur, l'erreur relative ou le pourcentage de prévisions incorrectes, ainsi que la durée de formation.

  L'erreur est l'erreur de la somme des carrés. Apparaissent également les erreurs relatives ou les pourcentages de prévisions incorrectes, suivant les niveaux de mesure des variables dépendantes. Si une variable dépendante comporte un niveau de mesure d'échelle, l'erreur relative globale moyenne (par rapport au modèle moyen) est affichée. Si toutes les variables dépendantes sont des variables qualitatives, le pourcentage moyen de prévisions incorrectes est affiché. Les erreurs ou les pourcentages relatifs de prévisions incorrectes sont également affichés pour les variables dépendantes individuelles.

- **Résultats du classement.** Affiche un tableau de classement pour chaque variable dépendante qualitative. Chaque tableau indique le nombre d'observations classées correctement et incorrectement pour chaque modalité de variable dépendante. Le pourcentage d'observations totales ayant été correctement classées est également indiqué.

- **Courbe ROC** Affiche une courbe ROC (Receiver Operating Characteristic) pour chaque variable dépendante qualitative. Affiche également un tableau indiquant la zone au-dessous de chaque courbe. Pour une variable dépendante donnée, le diagramme ROC affiche une courbe pour chaque modalité. Si la variable dépendante comporte deux modalités, chaque courbe traite la modalité en question comme étant l'état positif par rapport à l'autre modalité. Si la variable dépendante comporte plus de deux modalités, chaque courbe traite la modalité en question comme étant l'état positif par rapport à la somme de toutes les autres modalités.

- **Diagramme de gains cumulés.** Affiche un diagramme de gains cumulés pour chaque variable dépendante qualitative. L'affichage d'une courbe pour chaque modalité de variable dépendante est identique à celui des courbes ROC.

- **Diagramme de Levier.** Affiche un diagramme de levier pour chaque variable dépendante qualitative. L'affichage d'une courbe pour chaque modalité de variable dépendante est identique à celui des courbes ROC.

- **Diagramme estimé/observé.** Affiche un diagramme estimé/observé pour chaque variable dépendante. Pour les variables dépendantes qualitatives, des boîtes à moustaches juxtaposées des pseudo-probabilités prévues sont affichées pour chaque modalité de réponse, avec la modalité de réponse observée comme variable de classe. Pour les variables d'échelle dépendantes, un diagramme de dispersion est affiché.

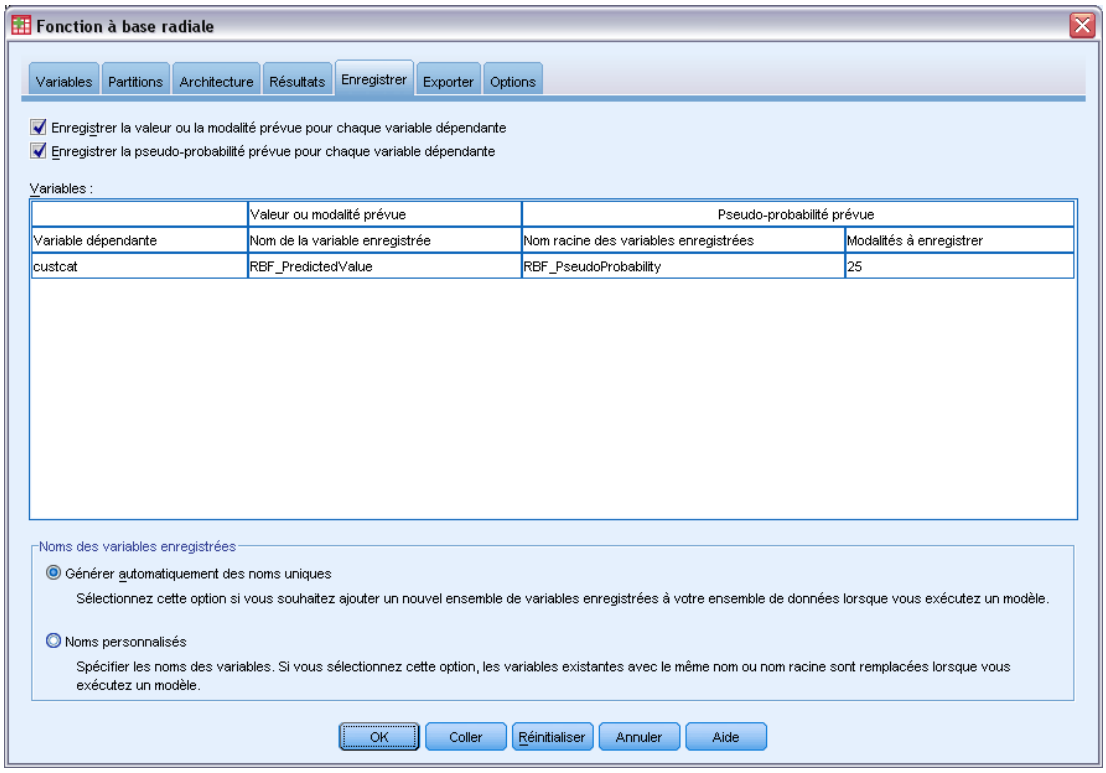
- **Diagramme résiduel/estimé.** Affiche un diagramme résiduel/estimé pour chaque variable d'échelle dépendante. Il ne doit exister aucun schéma visible entre les résidus et les prévisions. Ce diagramme n'est généré que pour les variables d'échelle dépendantes.

Récapitulatif du traitement des observations. Affiche le tableau récapitulatif de traitement des observations, qui récapitule le nombre d'observations incluses et exclues dans l'analyse, au total et par échantillon de formation, de test et traité.

Analyse de l'importance des variables prédites. Effectue une analyse de sensibilité, qui calcule l'importance de chaque variable prédite dans la détermination du réseau neuronal. L'analyse est basée sur les échantillons de formation et de test combinés, ou uniquement sur l'échantillon de formation s'il n'existe aucun échantillon de test. Ceci produit un tableau et un diagramme qui indiquent l'importance et l'importance normalisée de chaque variable prédite. L'analyse de sensibilité nécessite beaucoup de calculs et de temps si les variables prédites ou les observations sont nombreuses.

Enregistrer

Figure 3-6
Fonction à base radiale : Onglet Enregistrer



L'onglet Enregistrer permet d'enregistrer les prévisions en tant que variables dans l'ensemble de données.

- **Enregistrer la valeur ou la modalité prévue pour chaque variable dépendante.** Cette option enregistre la valeur prévue pour les variables d'échelle dépendantes et la modalité prévue pour les variables dépendantes qualitatives.
- **Enregistrer la pseudo-probabilité prévue pour chaque variable dépendante.** Cette option enregistre les pseudo-probabilités prévues pour les variables dépendantes qualitatives. Une variable distincte est enregistrée pour chacune des *n* premières modalités, *n* étant spécifié dans la colonne *Modalités à enregistrer*.

Noms des variables enregistrées. Grâce à la génération automatique de nom, vous conservez l'ensemble de votre travail. Les noms personnalisés vous permettent de supprimer ou de remplacer les résultats d'exécutions précédentes sans supprimer d'abord les variables enregistrées dans l'éditeur de données.

Probabilités et pseudo-probabilités

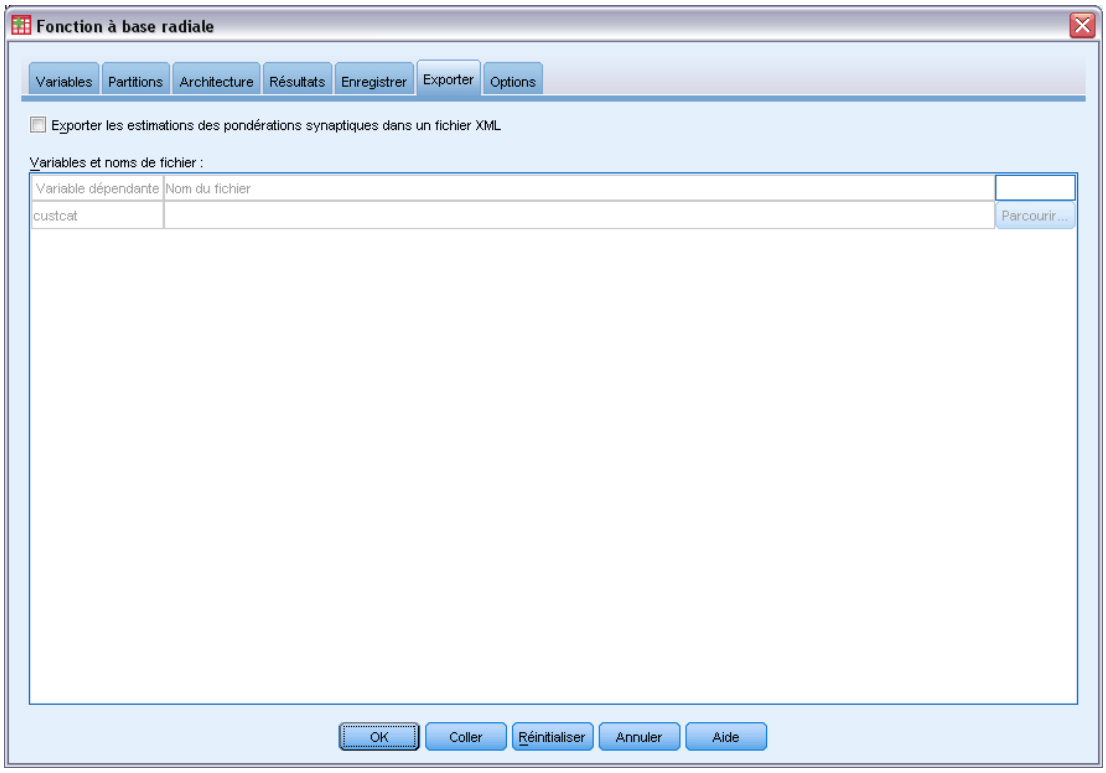
Les pseudo-probabilités prévues ne peuvent pas être interprétées comme des probabilités, car la procédure de fonction à base radiale utilise l'erreur de la somme des carrés et la fonction d'activation d'identité pour la strate de résultat. La procédure enregistre ces pseudo-probabilités prévues même si certaines d'entre elles sont inférieures à 0 ou supérieures à 1, ou si la somme d'une variable dépendante donnée n'est pas égale à 1.

Le diagramme de ROC, des gains cumulés et de Levier (reportez-vous Résultat sur p. 30) sont créés en fonction des pseudo-probabilités. Si des pseudo-probabilités sont inférieures à 0 ou supérieures à 1 ou que la somme d'une variable donnée n'est pas égale à 1, elles sont d'abord rééchelonnées pour se situer entre 0 et 1, et avoir pour somme 1. Les pseudo-probabilités sont rééchelonnées en étant divisées par leur somme. Par exemple, si une observation comporte des pseudo-probabilités de 0,50, 0,60 et 0,40 pour une variable dépendante à trois modalités, chaque pseudo-probabilité est alors divisée par la somme 1,50 afin d'obtenir 0,33, 0,40 et 0.27.

Si des pseudo-probabilités sont négatives, la valeur absolue de la plus faible est ajoutée à toutes les pseudo-probabilités avant le rééchelonnement ci-dessus. Par exemple, si les pseudo-probabilités sont -0.30, 0,50, et 1,30, ajoutez d'abord 0,30 à chaque valeur pour obtenir 0,00, 0,80 et 1,60. Divisez ensuite chaque nouvelle valeur par la somme 2,40 pour obtenir 0,00, 0,33 et 0,67.

Exporter

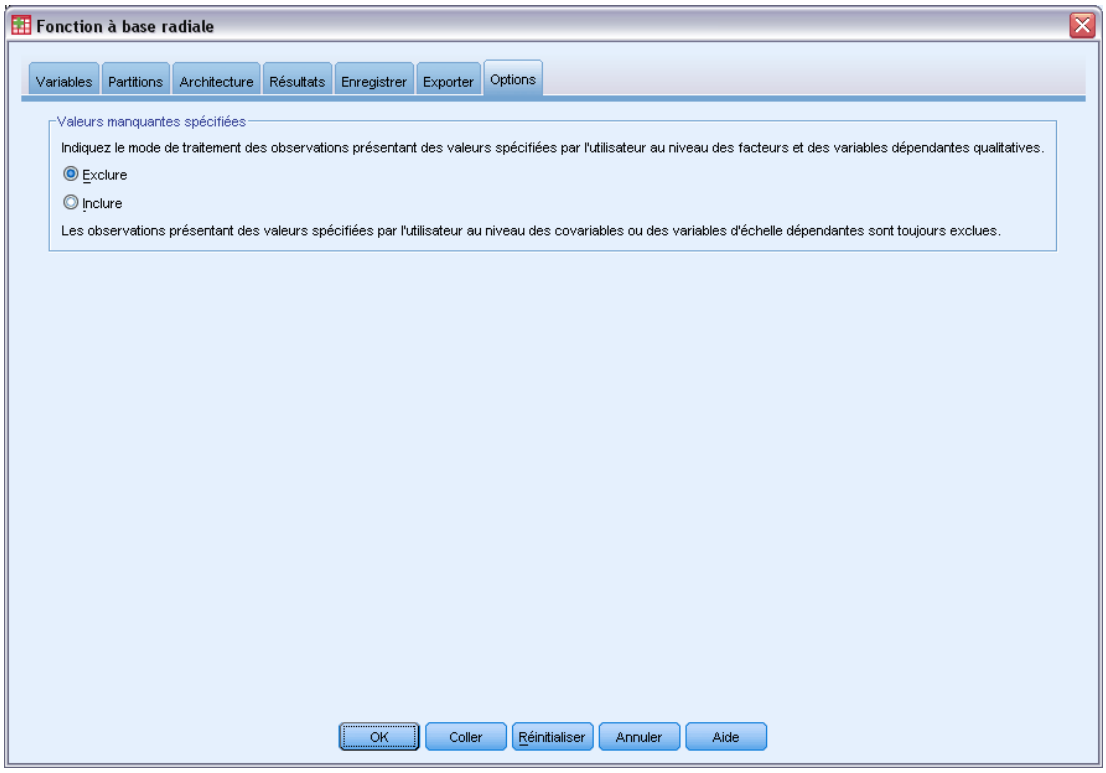
Figure 3-7
Fonction à base radiale : Onglet Exporter



L'onglet Exporter permet d'enregistrer les estimations des pondérations synaptiques de chaque variable dépendante dans un fichier XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation. Cette option n'est pas disponible si des fichiers scindés ont été définis.

Options

Figure 3-8
Fonction à base radiale : Onglet Options



Valeurs manquantes spécifiées. Les facteurs doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les facteurs et les variables dépendantes qualitatives.

Partie II: Exemples

Chapitre
4

Perceptron multi-couches

La procédure Perceptron multistrate produit un modèle de prévision pour une ou plusieurs variables (cible) dépendantes en fonction de valeurs de variables explicatives.

Utilisation du perceptron multistrate pour évaluer le risque de crédit

Un responsable des prêts dans une banque souhaite pouvoir identifier les caractéristiques qui indiquent les personnes susceptibles de manquer à leurs engagements et d'utiliser ces caractéristiques pour identifier les bons et les mauvais risques de crédit.

Supposez que les informations sur les 850 clients précédents et éventuels soient contenues dans le fichier *bankloan.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 87.Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Utilisez un échantillon aléatoire de ces 700 clients pour créer un perceptron multistrate, en laissant le reste des clients de côté pour valider l'analyse. Utilisez ensuite le modèle pour classer les 150 clients éventuels entre bon et mauvais risques de crédit.

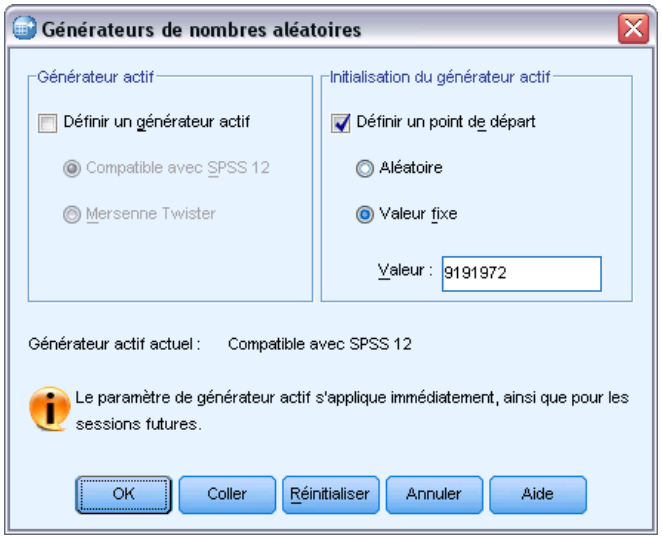
Par ailleurs, le responsable des prêts a auparavant analysé les données en utilisant une régression logistique (dans l'option Régression) et se demande dans quelle mesure le perceptron multistrate peut-il être comparé en tant qu'outil de classement.

Préparation des données pour l'analyse

Définir le générateur aléatoire vous permet de reproduire l'analyse exactement.

- Pour définir le générateur aléatoire, à partir des menus, sélectionnez :
  Transformer > Générateurs de nombres aléatoires…

Figure 4-1
Boîte de dialogue Générateurs de nombres aléatoires

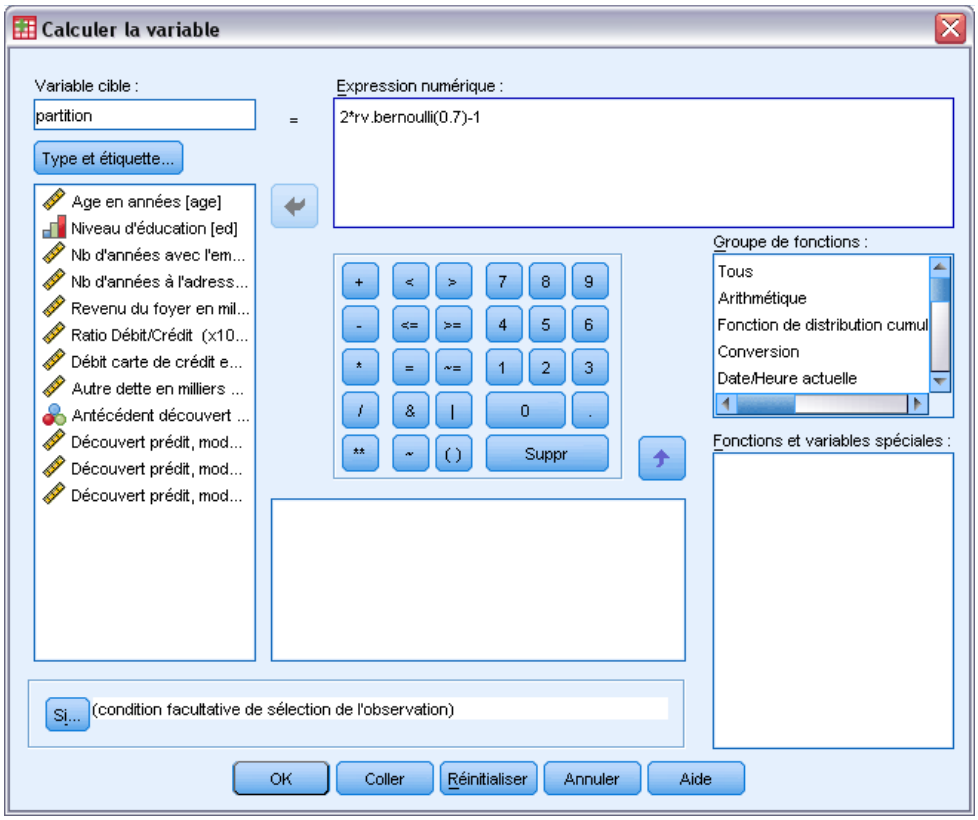


- Sélectionnez Définir un point de départ.
- Sélectionnez Valeur fixe et tapez la valeur 9 191 972.
- Cliquez sur OK.

Dans l'analyse de régression logistique précédente, environ 70 % des clients passés étaient attribués à l'échantillon d'apprentissage et 30 % à un échantillon traité. Vous devez recourir à une variable de partitionnement pour recréer exactement les échantillons utilisés dans ces analyses.

- Pour créer la variable de partition, dans les menus, choisissez :
  Transformer > Calculer la variable...

Figure 4-2
Boîte de dialogue Calculer la variable



- Saisissez partition dans la zone de texte Variable cible.
- Tapez 2*rv.bernoulli(0,7)-1 dans la zone Expression numérique.

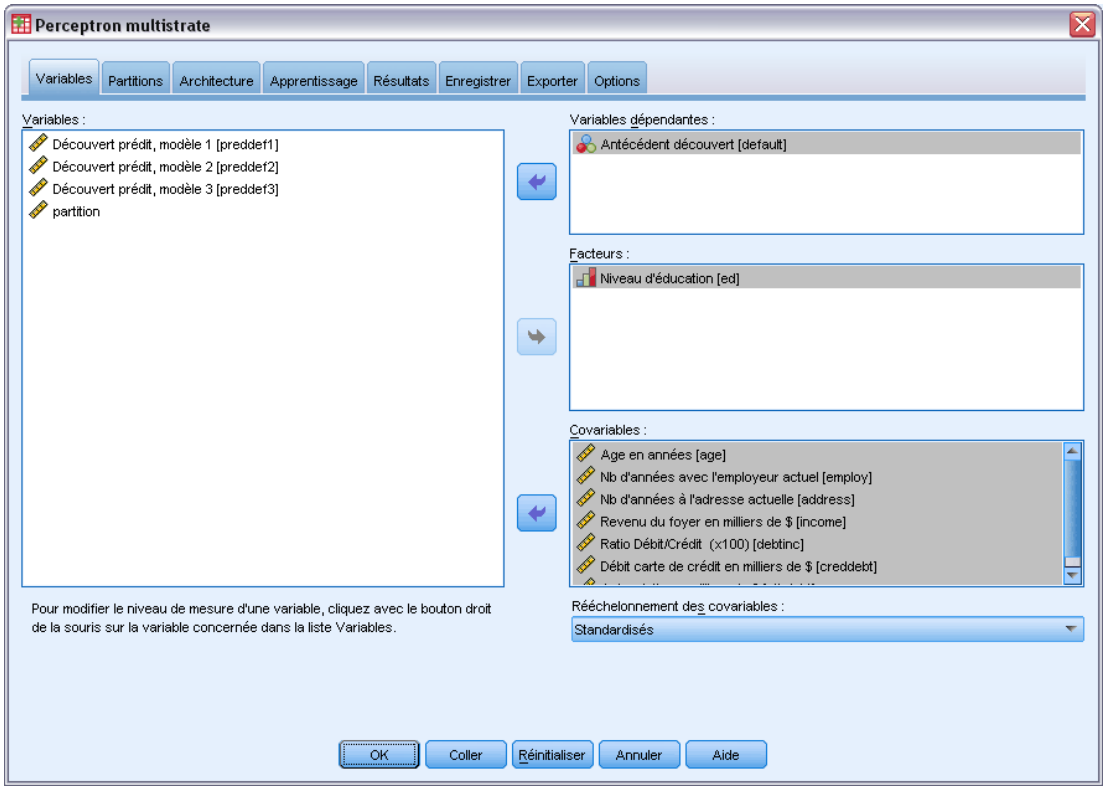
  Vous définissez ainsi les valeurs de *validation* comme variables de **Bernoulli** générées aléatoirement avec un paramètre de probabilité de 0,7, modifiées de manière à prendre la valeur 1 ou −1, au lieu de 1 ou 0. N'oubliez pas que les observations contenant des valeurs positives sur la variable de partitionnement sont affectées à l'échantillon de formation, celles avec des valeurs négatives sont affectées à l'échantillon traité et celles avec une valeur égale à 0 sont affectées à l'échantillon de test. Nous n'allons pas indiquer d'échantillon de test pour l'instant.
- Cliquez sur OK dans la boîte de dialogue Calculer la variable.

  Environ 70 % des clients ayant précédemment bénéficié d'un prêt auront 1 comme valeur pour la variable *partition*. Ces clients sont utilisés pour créer le modèle. Les autres clients ayant précédemment bénéficié d'un prêt auront une valeur de *partition* égale à −1, et seront utilisés pour valider les résultats du modèle.

Exécution de l'analyse

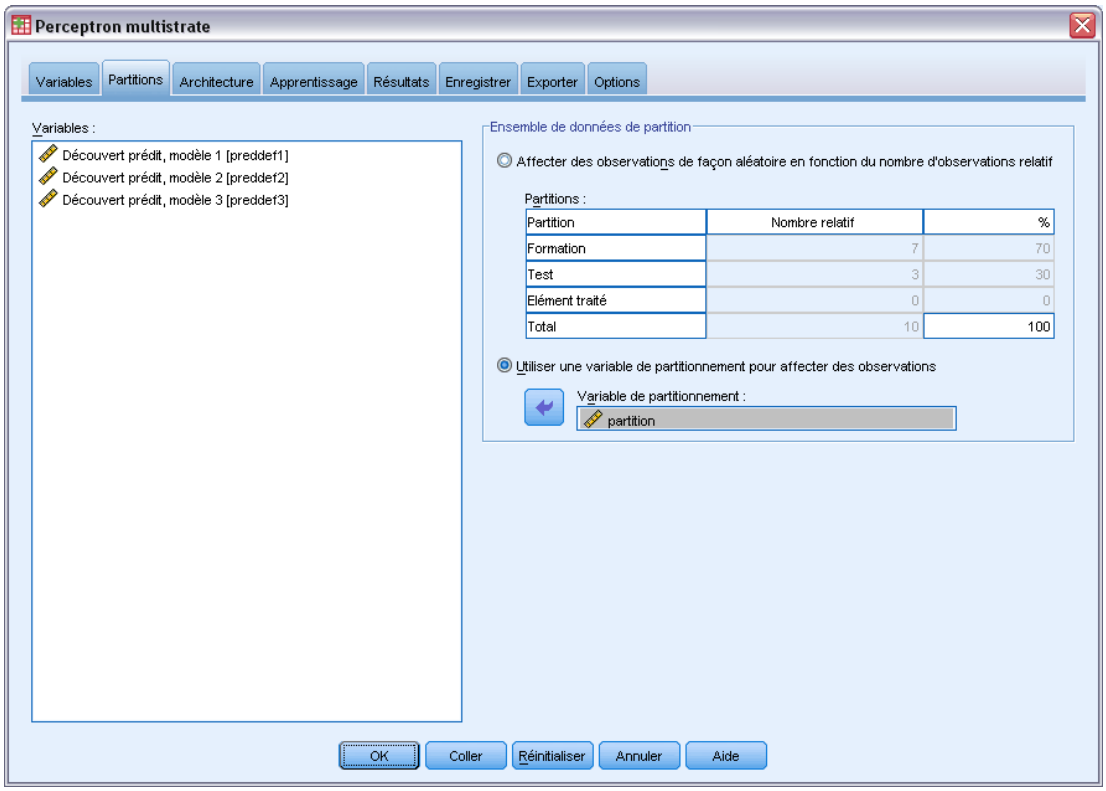
- Pour lancer une analyse Perceptron multistrate, choisissez les options suivantes dans les menus : Analyse > Réseaux neuronaux : > Perceptron multistrate...

Figure 4-3
Perceptron multistrate : l'onglet Variables



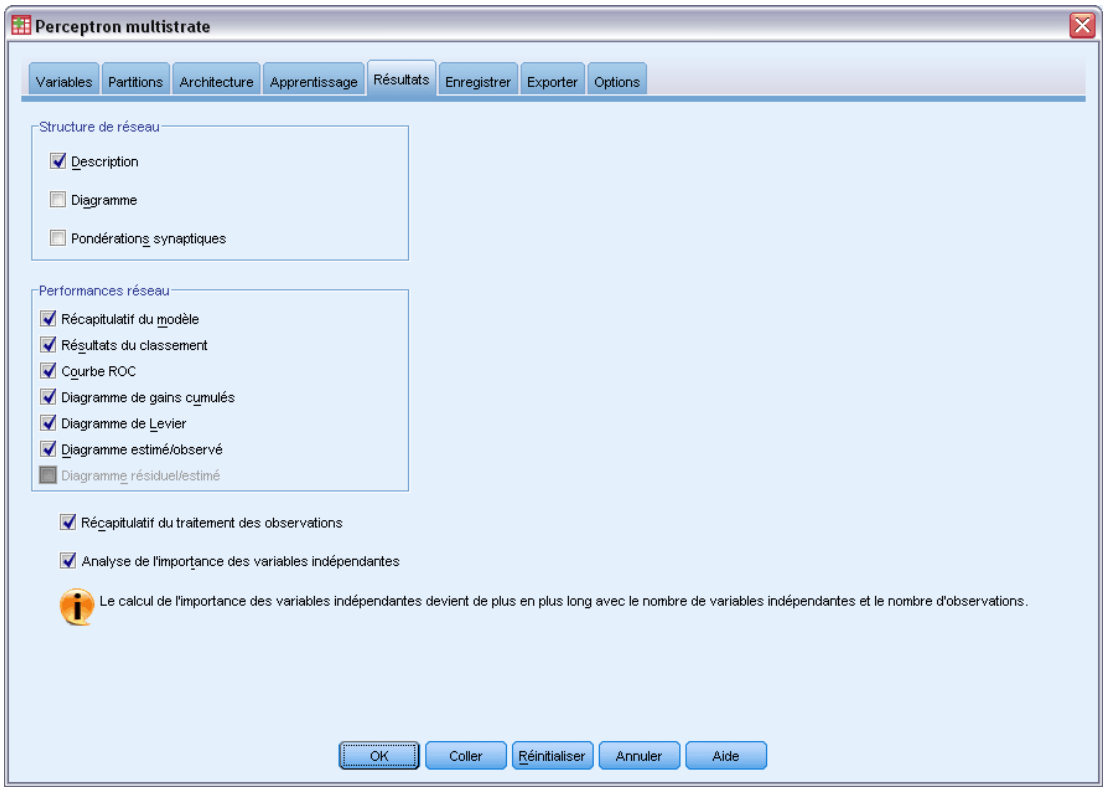
- Sélectionnez *Manquement précédent [défaut]* comme variable dépendante.
- Sélectionnez *Niveau d'éducation [ne]* comme facteur.
- Sélectionnez les options de *Age en années [age]* à *Autres dettes en milliers [autrdettes]* comme covariables.
- Cliquez sur l'onglet Partitions.

Figure 4-4
Perceptron multistrate : Onglet Partitions



- Sélectionnez l'option Utiliser une variable de partitionnement pour affecter des observations.
- Sélectionnez l'option *partition* comme variable de ligne.
- Cliquez sur l'onglet Résultats.

Figure 4-5
Perceptron multistrate : Onglet Résultats



- Désélectionnez l'option Diagramme dans le groupe Structure de réseau.
- Sélectionnez les options Courbe ROC, Diagramme de gains cumulés, Diagramme de Levier et Diagramme estimé/observé dans le groupe Performances réseau. Le diagramme Valeurs résiduelles par prévisions n'est pas disponible, car la variable dépendante n'est pas une variable d'échelle.
- Sélectionnez l'option Analyse de l'importance des variables indépendantes.
- Cliquez sur OK.

Récapitulatif de traitement des observations

Figure 4-6
Récapitulatif du traitement des observations

		N	Pourcentage
Echantillon	Formation	499	71,3%
	Traitement	201	28,7%
	Valide	700	100,0%
	Exclue	150	
	Total	850	

Le récapitulatif du traitement des observations montre que 499 observations ont été attribuées à l'échantillon d'apprentissage et 201 à l'échantillon traité. Les 150 observations exclues de l'analyse correspondent aux clients potentiels.

Informations réseau

Figure 4-7
Informations sur le réseau

Strate d'entrée	Facteurs	1	Niveau d'éducation
	Covariables	1	Age en années
		2	Autre dette en milliers de $
	Nombre d'unités[a]		7
	Méthode de rééchelonnement des covariables		Standardisé
Strate(s) masquée(s)	Nombre de strates masquées		1
	Nombre d'unités dans la strate masquée 1[a]		3
	Fonction d'activation		Tangente hyperbolique
Strate de sortie	Variables dépendantes	1	Antécédent découvert
	Nombre d'unités		2
	Fonction d'activation		MaxMou
	Fonction d'erreur		Entropie croisée

a. Exclusion de l'unité biaisée

Le tableau d'informations sur le réseau affiche des informations sur le réseau neuronal et permet de vérifier que les spécifications sont correctes. En l'occurrence, notez les points suivants :

- Le nombre d'unités dans la strate d'entrée correspond au nombre de covariables plus le nombre total de niveaux de facteur ; une unité spécifique est créée pour chaque modalité de *niveau d'éducation* et aucune des modalités n'est considérée comme une unité « redondante », comme cela est courant dans de nombreuses procédures de modélisation.
- De même, une unité de résultat spécifique est créée pour chaque modalité de *manquement précédent*, pour un total de 2 unités dans la strate de résultat.
- La sélection automatique de l'architecture a choisi 4 unités dans la strate masquée.
- Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Récapitulatif des modèles

Figure 4-8
Récapitulatif du modèle

Formation	Erreur d'entropie croisée	251,7%
	Prévisions de pourcentage incorrectes	25,2%
	Arrêt de la règle utilisée	Modification relative dans le critère d'erreur de formation (,0001) atteinte%
	Durée de formation	0:00:00.651%
Traitement	Prévisions de pourcentage incorrectes	23,1%
Variable dépendante : Antécédent découvert		

Le récapitulatif du modèle affiche des informations sur les résultats de l'apprentissage du réseau final et de son application à l'échantillon traité.

- Une erreur d'entropie croisée apparaît, car la strate de résultat utilise la fonction d'activation softmax. Il s'agit de la fonction d'erreur que le réseau essaie de minimiser pendant l'apprentissage.
- Le pourcentage de prévisions incorrectes provient du tableau de classement et sera abordé plus loin dans cette section.
- L'algorithme d'estimation s'est arrêté, car le nombre maximum de périodes a été atteint. Normalement, l'apprentissage s'arrête lorsque l'erreur a convergé. Cela soulève des questions quant à un dysfonctionnement éventuel pendant l'apprentissage et doit être pris en compte lors de l'examen du résultat.

Classification

Figure 4-9
Classification

		Estimé		
Echantillon	Observations	Non	Oui	Pourcentage d'éléments corrects
Formation	Non	328	22	93,7%
	Oui	101	37	26,8%
	Pourcentage global	87,9%	12,1%	74,8%
Traitement	Non	155	12	92,8%
	Oui	37	8	17,8%
	Pourcentage global	90,6%	9,4%	76,9%
Variable dépendante : Antécédent découvert				

Le tableau de classement affiche les résultats pratiques de l'utilisation du réseau. Pour chaque observation, la réponse prévue est *Oui* si la pseudo-probabilité prévue de cette observation est supérieure à 0,5. Pour chaque échantillon :

- Les cellules situées sur la diagonale de la classification croisée des observations sont des prévisions correctes.
- Les cellules hors de la diagonale de la classification croisée des observations sont des prévisions incorrectes.

Des observations utilisées pour créer le modèle, 74 des 124 personnes qui ont précédemment manqué à leurs engagements ont été classées correctement. 347 des 375 personnes n'ayant pas manqué à leurs engagements ont été classées correctement. Au total, 84,4 % des observations d'apprentissage ont été classées correctement, ce qui correspond à la proportion de 15,6 % indiquée dans le tableau récapitulatif des modèles. Un meilleur modèle doit correctement identifier un pourcentage supérieur des observations.

Les classements basés sur les observations utilisées pour créer le modèle tendent à être trop « optimistes » dans le sens où leur taux de classification est augmenté. L'échantillon traité permet de valider le modèle ; en l'occurrence, le modèle a correctement classé 74.6% de ces observations. Ceci suggère qu'en général votre modèle est en fait correct environ trois fois sur quatre.

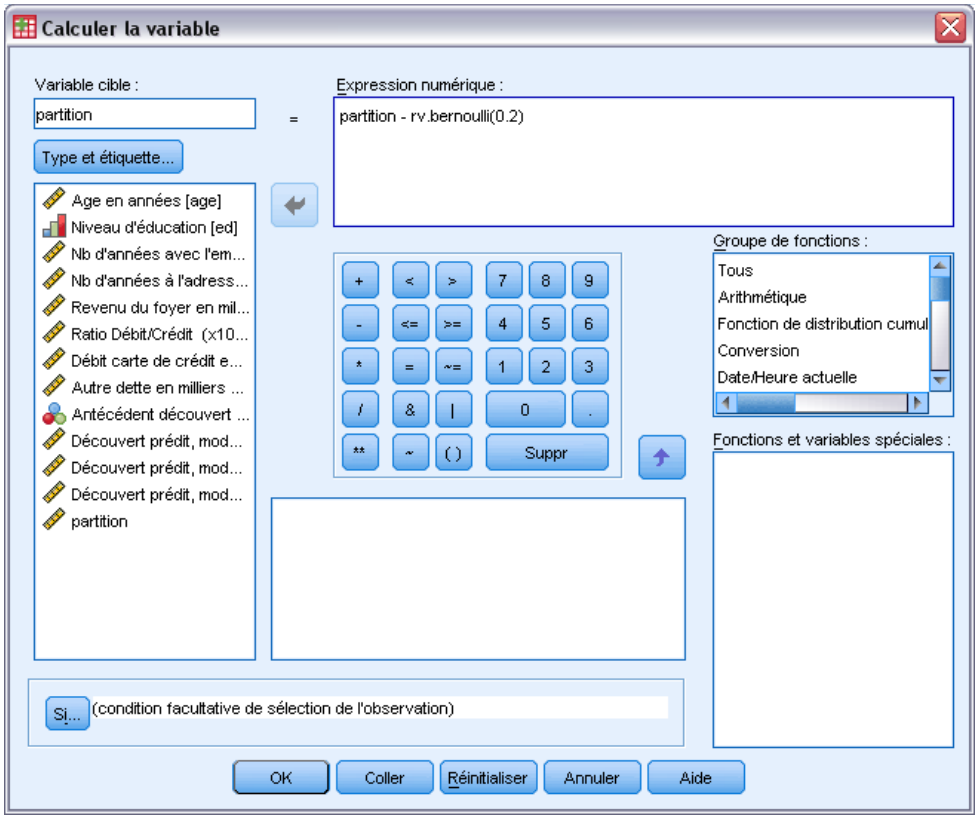
Correction du surapprentissage

En repensant à l'analyse de régression logistique précédemment réalisée, le responsable des prêts se souvient que l'échantillon d'apprentissage et l'échantillon traité ont correctement prévu un pourcentage similaire d'observations, environ 80 %. En revanche, le réseau neuronal avait un pourcentage d'observations correctes plus élevé dans l'échantillon d'apprentissage, l'échantillon traité ayant beaucoup moins bien prévu les clients ayant effectivement manqué à leurs engagements (correct à 45,8 % pour l'échantillon traité et 59,7 % pour l'échantillon d'apprentissage). Compte tenu de la règle d'arrêt indiquée dans le tableau récapitulatif des modèles, vous êtes amené à penser que le réseau est peut-être soumis à un **surapprentissage** ; c'est-à-dire qu'il recherche les modèles faux apparaissant dans les données d'apprentissage par variation aléatoire.

Heureusement, la solution est relativement simple : indiquez un échantillon de test pour aider le réseau à rester « sur la bonne voie ». Nous avons créé la variable de partitionnement de manière à récréer exactement l'échantillon d'apprentissage et l'échantillon traité utilisés dans l'analyse de régression logistique ; toutefois, le concept d'échantillon de « test » est étranger à la régression logistique. Prenons une partie de l'échantillon d'apprentissage et réaffectons-la à un échantillon de test.

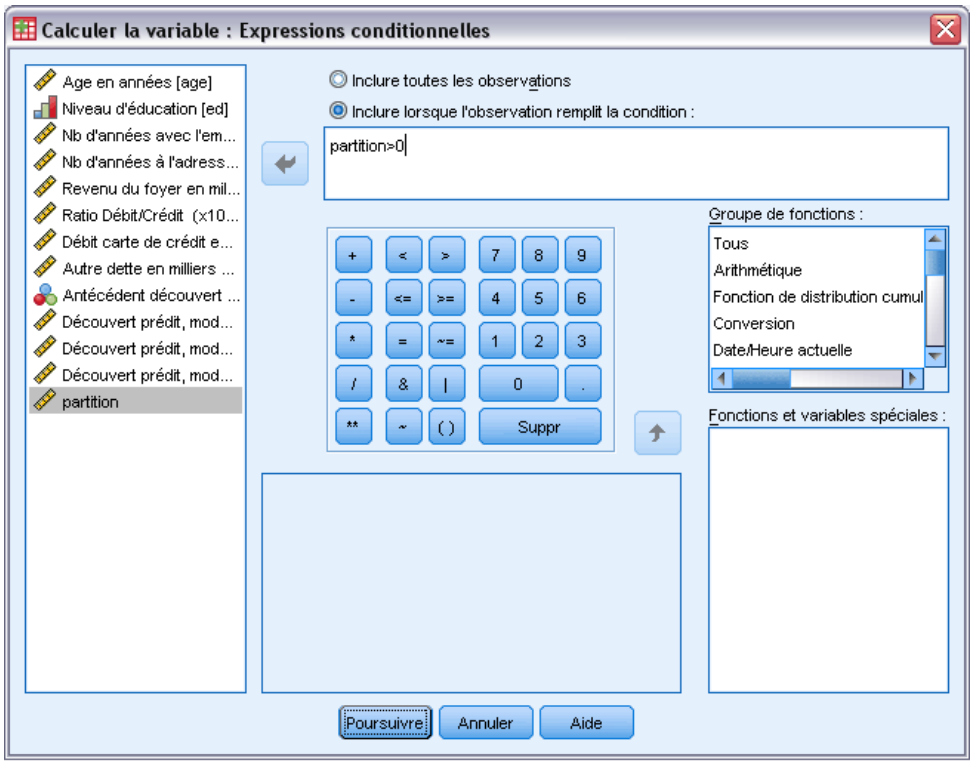
Création de l'échantillon de test

Figure 4-10
Boîte de dialogue Calculer la variable



- Rappelez la boîte de dialogue Calculer la variable.
- Tapez partition - rv.bernoulli(0,2) dans la zone Expression numérique.
- Cliquez sur Si.

Figure 4-11
Calculer la variable : Boîte de dialogue Calculer la variable : si les observations



- Sélectionnez Inclure si l'observation remplit la condition :
- Saisissez partition>0 dans la zone de texte.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Calculer la variable.

Au terme de cette opération, les valeurs de *partition* qui étaient supérieures à 0 sont redéfinies si bien qu'environ 20 % ont pour valeur 0 et 80 % conservent la valeur 1. Au total, environ 100*(0,7*0,8)=56 % des clients ayant précédemment bénéficié d'un prêt figureront dans l'échantillon d'apprentissage et 14 % dans l'échantillon de test. Les clients initialement attribués à l'échantillon traité y demeurent.

Exécution de l'analyse

- Dans la boîte de dialogue Perceptron multistrate, cliquez sur l'onglet Enregistrer.
- Sélectionnez l'option Enregistrer la pseudo-probabilité prévue pour chaque variable dépendante.
- Cliquez sur OK.

Récapitulatif de traitement des observations

Figure 4-12
Récapitulatif du traitement des observations pour un modèle avec échantillon de test

		N	Pourcentage
Echantillon	Formation	395	56,4%
	Test	93	13,3%
	Traitement	212	30,3%
	Valide	700	100,0%
	Exclue	150	
	Total	850	

Parmi les 499 observations initialement attribuées à l'échantillon d'apprentissage, 101 ont été réaffectées à l'échantillon de test.

Informations réseau

Figure 4-13
Informations sur le réseau

Strate d'entrée	Facteurs	1	Niveau d'éducation
	Covariables	1	Age en années
		2	Autre dette en milliers de $
	Nombre d'unités[a]		7
	Méthode de rééchelonnement des covariables		Standardisé
Strate(s) masquée(s)	Nombre de strates masquées		1
	Nombre d'unités dans la strate masquée 1[a]		3
	Fonction d'activation		Tangente hyperbolique
Strate de sortie	Variables dépendantes	1	Antécédent découvert
	Nombre d'unités		2
	Fonction d'activation		MaxMou
	Fonction d'erreur		Entropie croisée

a. Exclusion de l'unité biaisée

La seule modification apportée au tableau d'informations sur le réseau est le fait que la sélection automatique de l'architecture ait choisi 7 unités dans la strate masquée.

Récapitulatif des modèles

Figure 4-14
Récapitulatif du modèle

Formation	Erreur d'entropie croisée	211,1%
	Prévisions de pourcentage incorrectes	28,4%
	Arrêt de la règle utilisée	1 étape(s) consécutive(s) sans diminution dans l'erreur%[a]
	Durée de formation	0:00:00.882%
Test	Erreur d'entropie croisée	51,4%
	Prévisions de pourcentage incorrectes	26,9%
Traitement	Prévisions de pourcentage incorrectes	21,7%

Variable dépendante : Antécédent découvert

a. Les calculs d'erreurs sont basés sur l'échantillon de test.

Le récapitulatif du modèle montre quelques signes positifs :

- Le pourcentage de prévisions incorrectes est approximativement égal dans les échantillons d'apprentissage, de test et traité.
- L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme.

Cela renforce l'hypothèse d'un surapprentissage du modèle d'origine et le problème a été résolu par l'ajout d'un échantillon de test. Bien sûr, les tailles des échantillons sont relativement petites et peut-être serait-il préférable de ne pas tirer de conclusion hâtive de la variation de quelques points de pourcentage.

Classification

Figure 4-15
Classification

		Estimé		
Echantillon	Observations	Non	Oui	Pourcentage d'éléments corrects
Formation	Non	276	6	97,9%
	Oui	106	7	6,2%
	Pourcentage global	96,7%	3,3%	71,6%
Test	Non	66	2	97,1%
	Oui	23	2	8,0%
	Pourcentage global	95,7%	4,3%	73,1%
Traitement	Non	164	3	98,2%
	Oui	43	2	4,4%
	Pourcentage global	97,6%	2,4%	78,3%

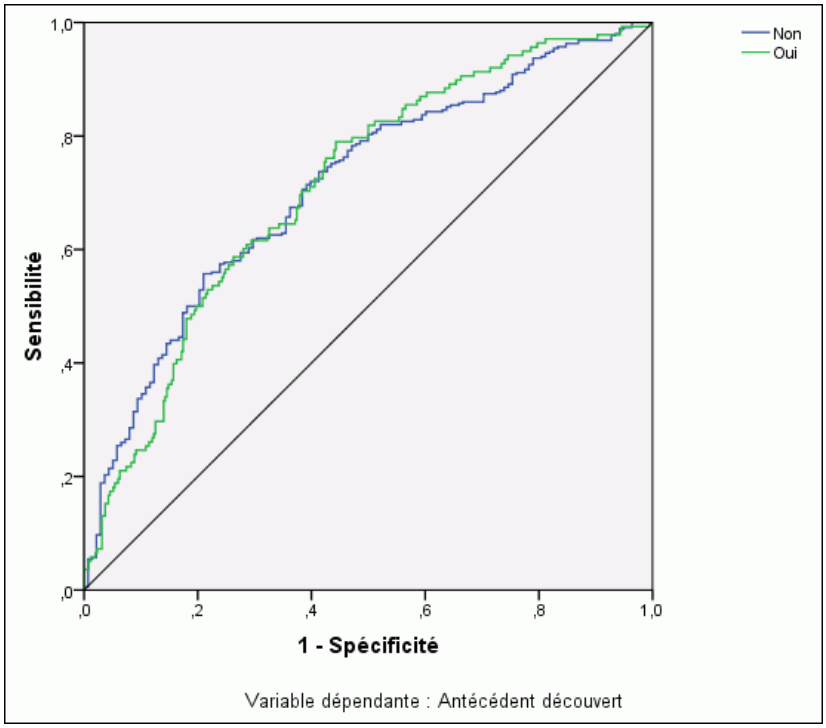
Variable dépendante : Antécédent découvert

Le tableau de classement montre que, avec 0,5 comme césure de pseudo-probabilité pour le classement, le réseau effectue des prévisions nettement meilleures pour les personnes ne manquant pas à leurs engagements que pour celles manquant à leurs engagements. Malheureusement, comme la valeur de césure unique donne un aperçu très limité de la capacité de prévision du

réseau, elle n'est pas nécessairement très utile pour la comparaison de réseaux concurrents. Observez plutôt la courbe ROC.

Courbe ROC

Figure 4-16
Courbe ROC



La courbe ROC présente un affichage visuel de **sensibilité** et **spécificité** pour toutes les césures possibles dans un diagramme unique, ce qui constitue un outil plus clair et plus puissant qu'une série de tableaux. Le diagramme proposé ici affiche deux courbes, l'une pour la modalité *Non*, l'autre pour la modalité *Oui*. Dans la mesure où il n'y a que deux modalités, les courbes sont symétriques par rapport à une ligne inclinée à 45 degrés (non affichée) allant de l'angle supérieur gauche du diagramme à l'angle inférieur droit.

Ce diagramme repose sur la combinaison de l'échantillon d'apprentissage et de l'échantillon de test. Pour obtenir un diagramme ROC pour l'échantillon traité, scindez le fichier au niveau de la variable de partitionnement, puis exécutez la procédure Courbe ROC sur les pseudo-probabilités prévues enregistrées.

Figure 4-17
Zone inférieure à la courbe

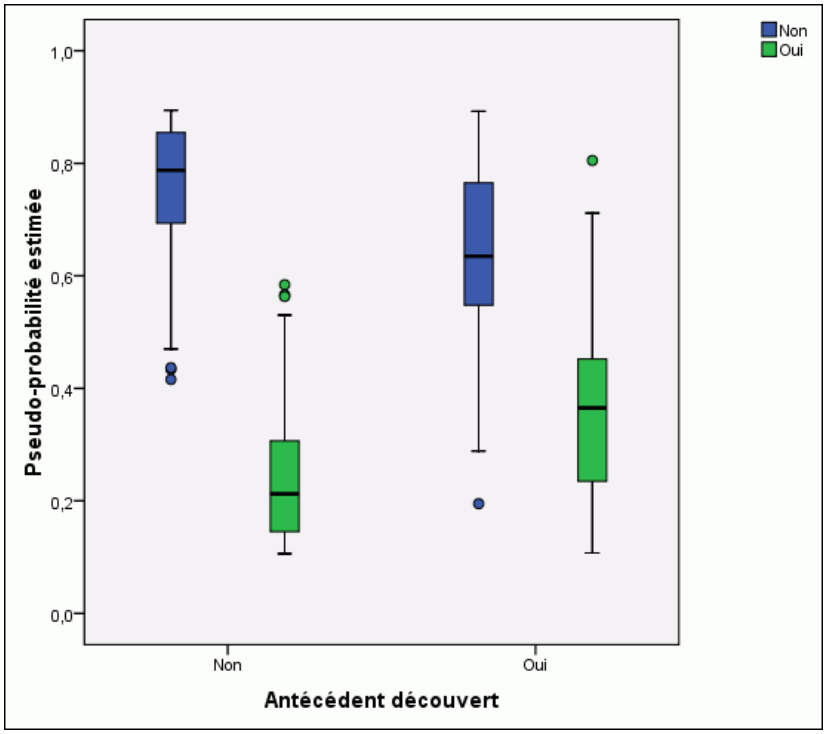
		Zone
Antécédent découvert	Non	,712
	Oui	,712

La zone inférieure à la courbe est un récapitulatif numérique de la courbe ROC, tandis que les valeurs du tableau représentent, pour chaque modalité, la probabilité que la présence de la pseudo-probabilité prévue dans cette modalité soit supérieure pour une observation choisie aléatoirement appartenant à cette modalité que pour une observation choisie aléatoirement n'appartenant pas à cette modalité. Par exemple, pour une personne manquant à ses engagements sélectionnée aléatoirement et une personne ne manquant pas à ses engagements sélectionnée aléatoirement, il existe une probabilité de 0,853 que la pseudo-probabilité de manquement prévue par le modèle soit plus élevée pour la personne manquant à ses engagements que pour la personne ne manquant pas à ses engagements.

Bien que la zone inférieure à la courbe constitue un récapitulatif à statistique unique utile de la précision du réseau, vous devez être en mesure de choisir un critère spécifique pour classer les clients. Pour ce faire, vous pouvez vous appuyer sur le diagramme estimé/observé.

Diagramme estimé/observé

Figure 4-18
Diagramme estimé/observé



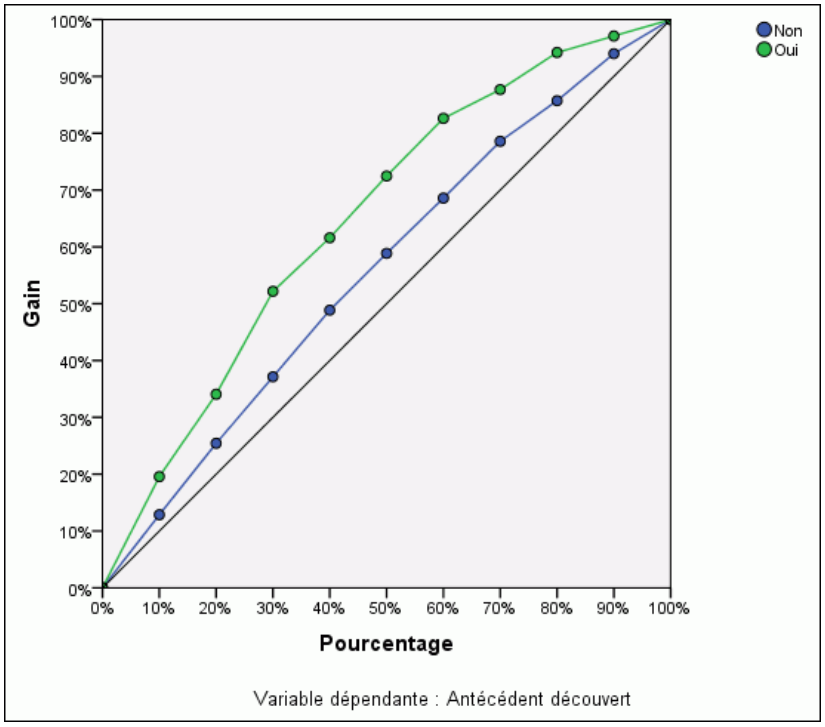
Dans le cas des variables dépendantes qualitatives, le diagramme estimé/observé affiche des boîtes à moustaches juxtaposées de pseudo-probabilités prévues pour les échantillons d'apprentissage et de test combinés. L'axe des *X* correspond aux modalités de réponses observées, et la légende aux modalités estimées.

- La boîte à moustaches le plus à gauche montre, pour les observations ayant comme modalité observée Non, la pseudo-probabilité prévue de la modalité Non. La partie de la boîte à moustaches au-dessus du repère 0,5 sur l'axe des *Y* représente les prévisions correctes montrées dans le tableau de classement. La partie au-dessous du repère 0,5 représente les prévisions incorrectes. D'après le tableau de classement, le réseau est très performant pour la prévision des observations ayant pour modalité Non avec la césure 0,5 ; par conséquent, seule une partie de la moustache inférieure et certaines observations éloignées sont mal classées.
- La boîte à moustaches suivante vers la droite montre, pour les observations ayant comme modalité observée Non, la pseudo-probabilité prévue de la modalité Oui. Dans la mesure où la variable cible ne comporte que deux modalités, les deux premières boîtes à moustaches sont symétriques par rapport à la ligne horizontale au niveau de 0,5.
- La troisième boîte à moustaches montre, pour les observations ayant comme modalité observée Oui, la pseudo-probabilité prévue de la modalité Non. Cette boîte à moustaches et la dernière sont symétriques par rapport à la ligne horizontale au niveau de 0,5.
- La dernière boîte à moustaches montre, pour les observations ayant comme modalité observée Oui, la pseudo-probabilité prévue de la modalité Oui. La partie de la boîte à moustaches au-dessus du repère 0,5 sur l'axe des *Y* représente les prévisions correctes montrées dans le tableau de classement. La partie au-dessous du repère 0,5 représente les prévisions incorrectes. D'après le tableau de classement, le réseau prévoit légèrement plus de la moitié des observations ayant pour modalité Oui avec la césure 0,5 ; par conséquent, une bonne partie de la boîte est mal classée.

Le diagramme indique que le fait d'abaisser la césure de classement d'une observation de modalité Oui de 0,5 à approximativement 0.3—qui représente plus ou moins la valeur à laquelle se trouvent le sommet de la deuxième boîte et la base de la quatrième, augmente la probabilité de repérer correctement les personnes susceptibles de manquer à leurs engagements sans perdre de nombreux bons clients potentiels. En d'autres termes, le passage de 0,5 à 0,3 dans la deuxième boîte aboutit au classement incorrect de relativement peu de clients ne manquant pas à leurs engagements le long de la moustache en tant que personnes manquant à leurs engagements prévues, tandis que dans la quatrième boîte, ce passage aboutit au classement correct de nombreux clients manquant à leurs engagements dans la boîte en tant que personnes manquant à leurs engagements prévues.

Diagrammes de gains cumulés et de Levier

Figure 4-19
Diagramme de gains cumulés

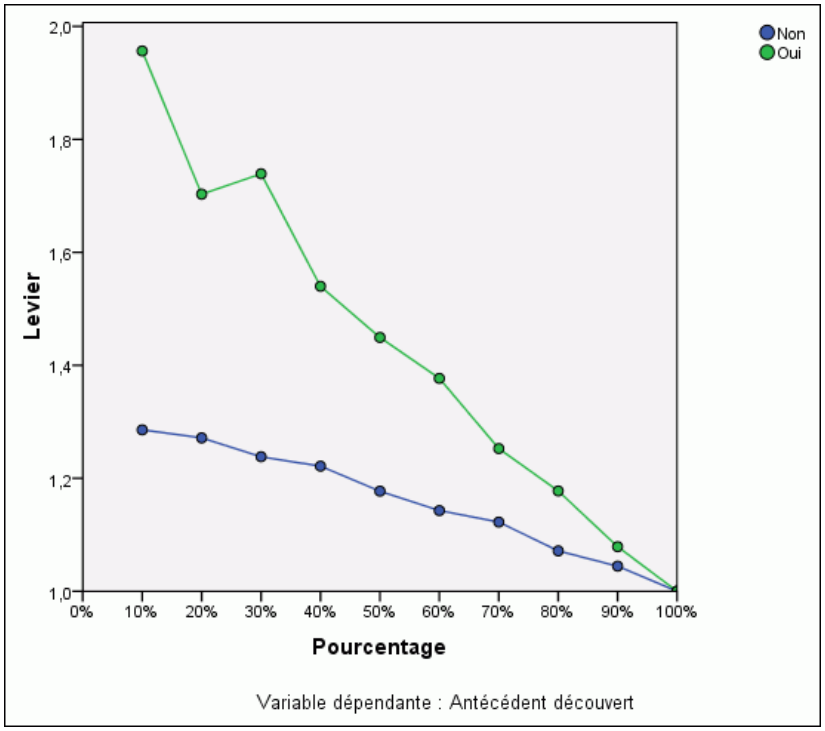


Le diagramme de gains cumulés montre le pourcentage du nombre total d'observations dans une modalité donnée obtenu en ciblant un pourcentage du nombre total d'observations. Par exemple, le premier point de la courbe pour la modalité *Oui* se situe à (10 %, 30 %), ce qui signifie que si vous évaluez un ensemble de données avec le réseau et que vous triez toutes les observations en fonction de la pseudo-probabilité prévue de la modalité *Oui*, vous pouvez vous attendre à ce que la tranche supérieure de 10 % contienne approximativement 30 % de la totalité des observations qui ont véritablement la modalité *Oui* (personnes manquant à leurs engagements). De même, la tranche supérieure de 20 % contiendrait approximativement 50 % des personnes manquant à leurs engagements, la tranche supérieure de 30 % des observations comporterait 70 % des personnes manquant à leurs engagements, et ainsi de suite. Si vous sélectionnez 100 % de l'ensemble de données évalué, vous obtenez la totalité des personnes manquant à leurs engagements dans l'ensemble de données.

La diagonale correspond à la courbe « de référence » ; si vous sélectionnez aléatoirement 10 % des observations dans l'ensemble de données évalué, vous pouvez espérer « obtenir » approximativement 10 % de la totalité des observations qui ont véritablement la modalité Oui. Plus une courbe se situe au-dessus de la ligne de base, plus le gain est élevé. Vous pouvez utiliser le diagramme de gains cumulés pour sélectionner une césure de classement en choisissant un pourcentage correspondant à un gain souhaitable, puis en associant ce pourcentage à la valeur de césure appropriée.

Ce qui constitue un gain « souhaitable » dépend du coût des erreurs de type I et de type II. En fait, quel est le coût du classement d'une personne manquant à ses engagements dans la catégorie des personnes ne manquant pas à leurs engagements (type I) ? Quel est le coût du classement d'une personne ne manquant pas à ses engagements dans la catégorie des personnes manquant à leurs engagements (type II) ? Si les mauvaises dettes sont votre préoccupation principale, alors minimisez votre erreur de type I ; dans le diagramme de gains cumulés, cela peut correspondre au rejet des prêts pour les demandeurs dans la tranche supérieure de 40 % de la pseudo-probabilité prévue de la modalité *Oui*, avec pour conséquence la capture de presque 90 % des personnes susceptibles de manquer à leurs engagements, mais la suppression de pratiquement la moitié de votre groupe de demandeurs. Si le développement de votre base client est la priorité, abaissez alors votre erreur de type II. Dans le diagramme, cela peut correspondre au rejet de la tranche supérieure de 10 %, avec pour conséquence la capture de 30 % des personnes manquant à leurs engagements et la conservation de votre groupe de demandeurs pratiquement tel quel. Habituellement, les deux sont des préoccupations majeures, vous devez donc choisir une règle de décision optimisant à la fois la sensibilité et la spécificité pour classer les clients.

Figure 4-20
Diagramme de Levier



Le diagramme de Levier est issu du diagramme de gains cumulés ; les valeurs de l'axe des *Y* correspondent au ratio du gain cumulé pour chaque courbe par rapport à la ligne de base. Par conséquent, le levier à 10 % pour la modalité Oui est 30 %/10 % = 3,0. Il permet d'observer différemment les informations du diagramme de gains cumulés.

Remarque : Le diagramme de gains cumulés et le diagramme de Levier reposent sur la combinaison de l'échantillon d'apprentissage et de l'échantillon de test.

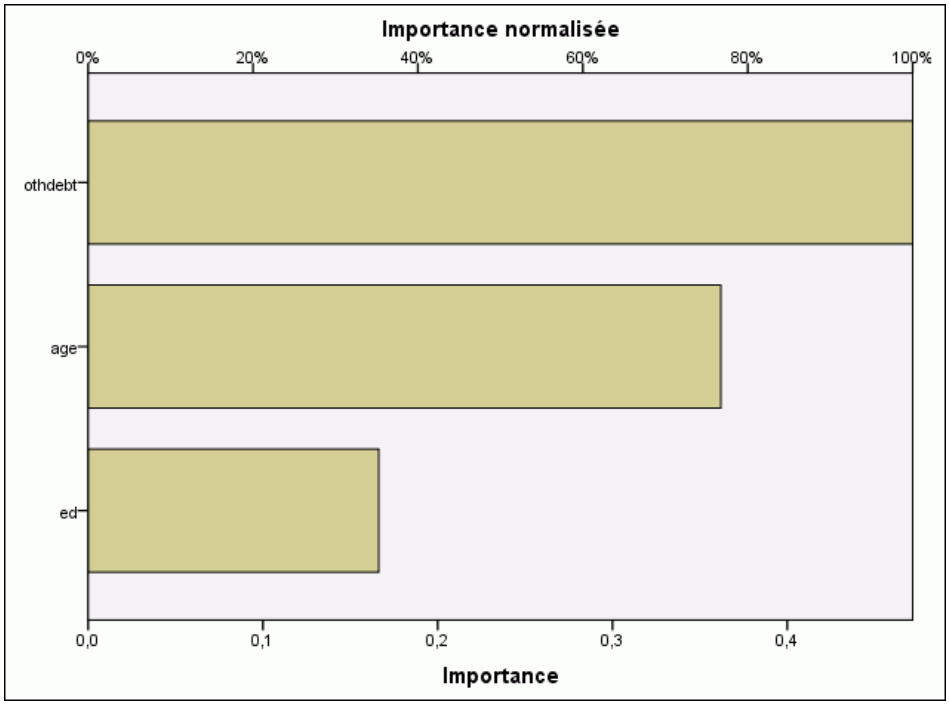
Importance des variables indépendantes

Figure 4-21
Importance de la variable indépendante

	Importance	Importance normalisée
Niveau d'éducation	,166	35,2%
Age en années	,362	76,7%
Autre dette en milliers de $	,472	100,0%

L'importance d'une variable indépendante mesure l'évolution de la valeur du réseau prévue par le modèle pour différentes valeurs de la variable indépendante. L'importance normalisée correspond simplement aux valeurs d'importance divisées par les valeurs d'importance les plus élevées et exprimées en pourcentages.

Figure 4-22
Diagramme de l'importance de la variable indépendante



Le diagramme d'importance est simplement un diagramme en bâtons des valeurs du tableau d'importance, triées par ordre décroissant de la valeur d'importance. Il apparaît que les variables liées à la stabilité (*emploi*, *adresse*) et aux dettes (*dettcred*, *dettrev*) d'un client ont la plus forte incidence sur la façon dont le réseau classe les clients ; toutefois, vous ne pouvez pas déterminer la « direction » de la relation entre ces variables et la probabilité de manquement prévue. Il vous semblerait que plus les dettes sont élevées, plus la probabilité de manquement est forte, mais vous devriez utiliser un modèle avec des paramètres plus faciles à interpréter pour confirmer cette impression.

Récapitulatif

A l'aide de la procédure de perceptron multistrate, vous avez construit un réseau pour prévoir la probabilité qu'un client donné manque à son prêt. Les résultats du modèle étant comparables à ceux obtenus à l'aide de la régression logistique ou de l'analyse discriminante, vous pouvez être raisonnablement assuré que les données ne contiennent pas de relations ne pouvant pas être capturées par ces modèles et, par conséquent, vous pouvez les utiliser pour déterminer avec précision la nature de la relation entre les variables dépendantes et indépendantes.

Utilisation d'un perceptron multistrate permettant d'évaluer les coûts liés aux soins et les durées de séjour

Un hôpital souhaite effectuer un suivi des coûts et des durées de séjour des patients admis pour soigner un infarctus du myocarde (crise cardiaque). Des estimations précises de ces mesures permettent à l'administration de gérer correctement le nombre de lits disponibles lors du traitement des patients.

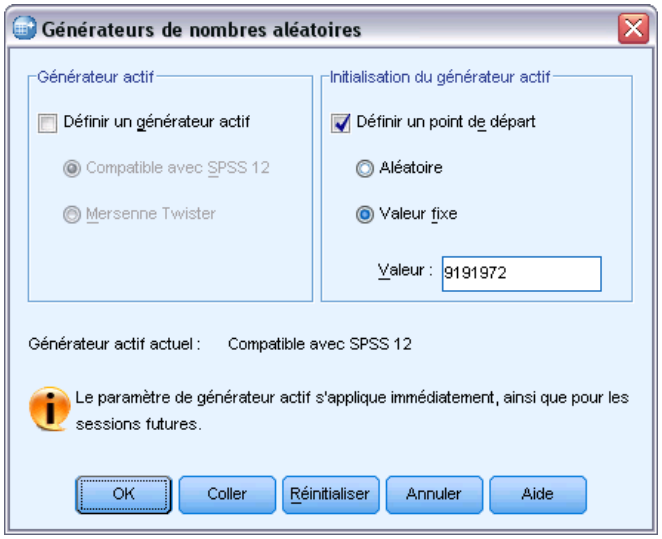
Le fichier de données *patient_los.sav* contient les données de traitement d'un échantillon de patients qui ont reçu un traitement pour l'infarctus du myocarde.Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 87. Utilisez la procédure du perceptron multistrate afin de mettre en place un réseau permettant de prévoir les coûts et la durée du séjour.

Préparation des données pour l'analyse

Définir le générateur aléatoire vous permet de reproduire l'analyse exactement.

- Pour définir le générateur aléatoire, à partir des menus, sélectionnez :
  Transformer > Générateurs de nombres aléatoires…

Figure 4-23
Boîte de dialogue Générateurs de nombres aléatoires

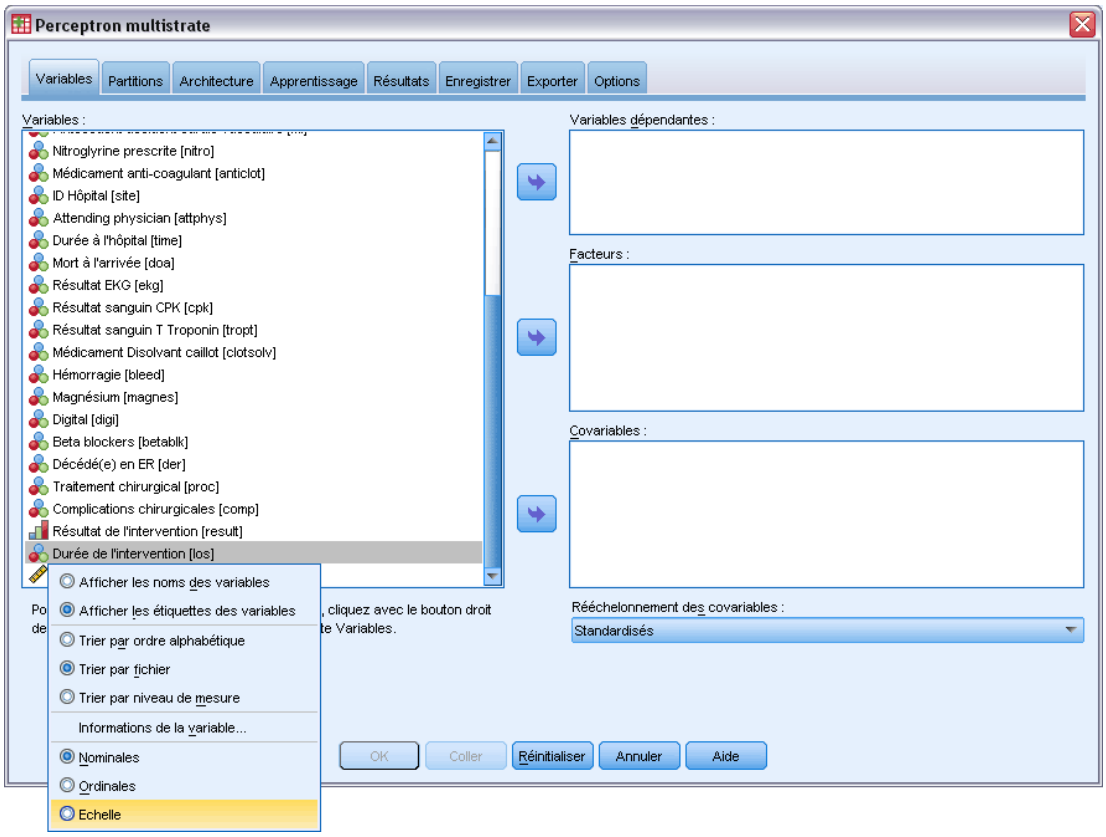


- Sélectionnez Définir un point de départ.
- Sélectionnez Valeur fixe et tapez la valeur 9 191 972.
- Cliquez sur OK.

Exécution de l'analyse

- Pour lancer une analyse Perceptron multistrate, choisissez les options suivantes dans les menus : Analyse > Réseaux neuronaux : > Perceptron multistrate...

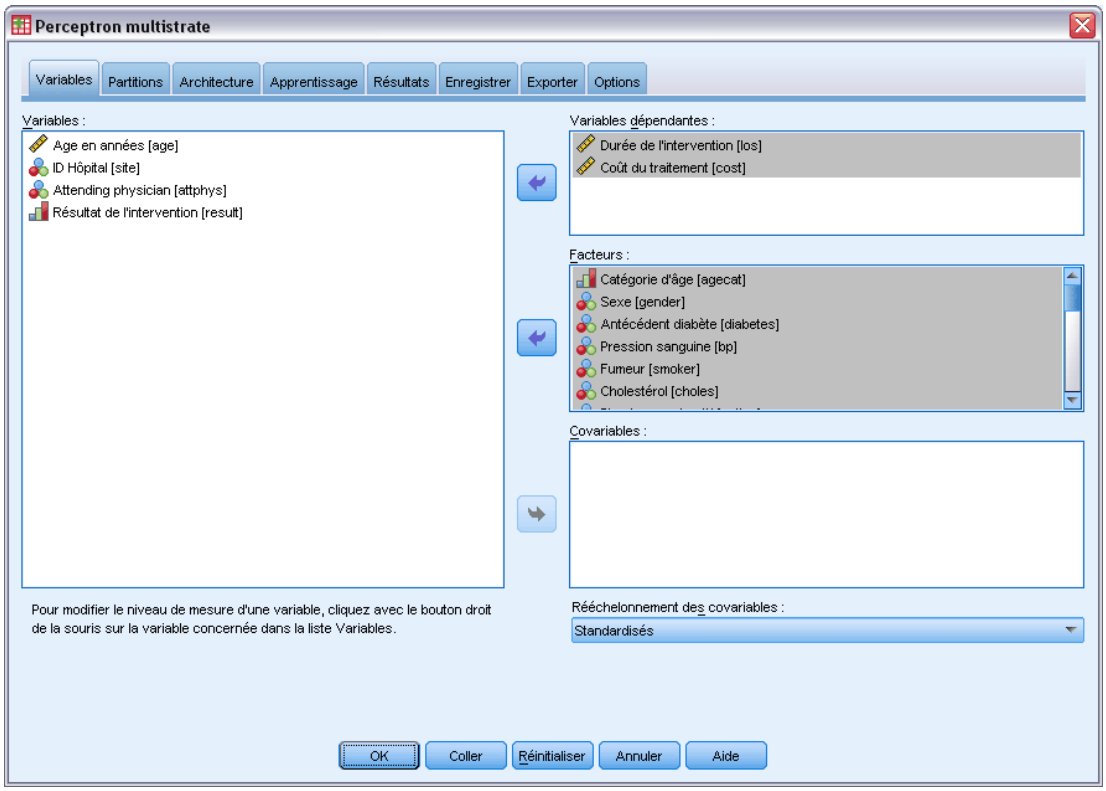
Figure 4-24
Perceptron multistrate : Onglet Variables - Menu contextuel pour la durée du séjour



Durée du séjour [los] a un niveau de mesure ordinal, mais vous voulez que le réseau la traite en tant qu'échelle.

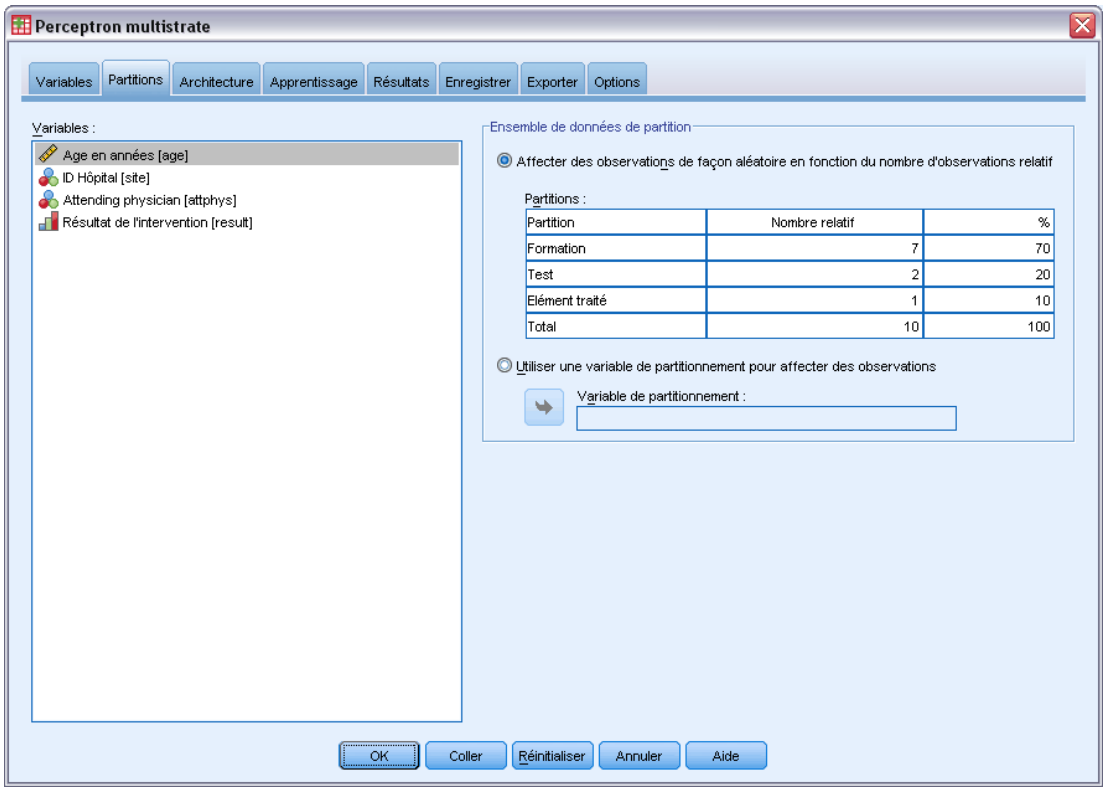
- Cliquez avec le bouton droit de la souris sur *Durée du séjour [dds]* et sélectionnez Echelle dans le menu contextuel.

Figure 4-25
Perceptron multistrate : Onglet Variables avec variables dépendantes et facteurs sélectionnés



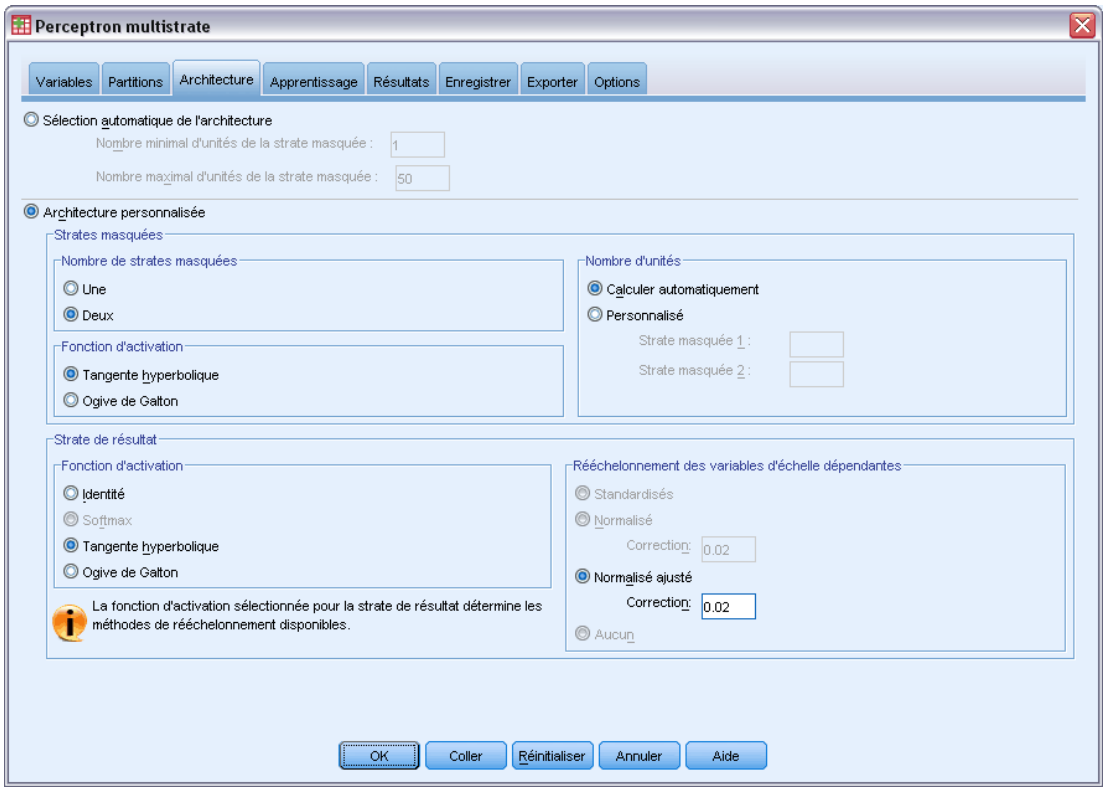
- Sélectionnez *Durée du séjour [los]* et *Coûts du traitement [cost]* en tant que variables dépendantes.
- Sélectionnez *Tranche d'âge [agecat]* dans la catégorie *Prise de médicaments anti-coagulation [anticlot]* et *Durée de l'hospitalisation [time]* dans la catégorie *Complications chirurgicales [comp]* en tant que facteurs. Afin de vous assurer une reproduction exacte des résultats du modèle ci-dessous, conservez bien l'ordre des variables dans la liste de facteurs. A cette fin, vous pouvez trouver utile de sélectionner chaque ensemble de variables indépendantes et d'utiliser le bouton pour les insérer dans la liste de facteurs, plutôt que de les oublier et les mettre de côté. Changer l'ordre des variables vous aide également à évaluer la stabilité de la solution.
- Cliquez sur l'onglet Partitions.

Figure 4-26
Perceptron multistrate : Onglet Partitions



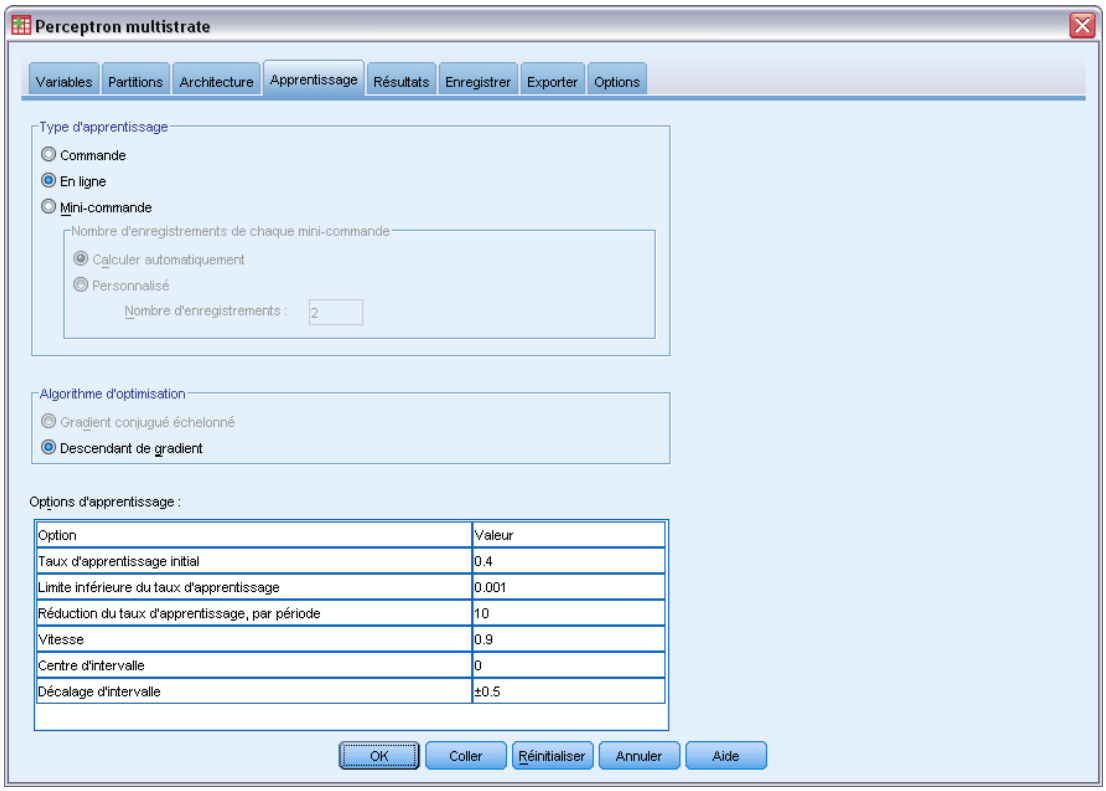
- Tapez 2 en tant que nombre de cas relatif à assigner à l'échantillon test.
- Tapez 1 en tant que nombre de cas relatif à assigner à l'échantillon traité.
- Cliquez sur l'onglet Architecture.

Figure 4-27
Perceptron multistrate : Onglet Architecture.



- Sélectionnez Architecture personnalisée.
- Sélectionnez Deux pour le nombre de strates masquées.
- Sélectionnez Tangente hyperbolique en tant que fonction d'activation de la strate de résultats. Veuillez remarquer que cette opération déclenche automatiquement la méthode de rééchelonnement des variables dépendantes en les définissant sur Normalisé ajusté.
- Cliquez sur l'onglet Formation.

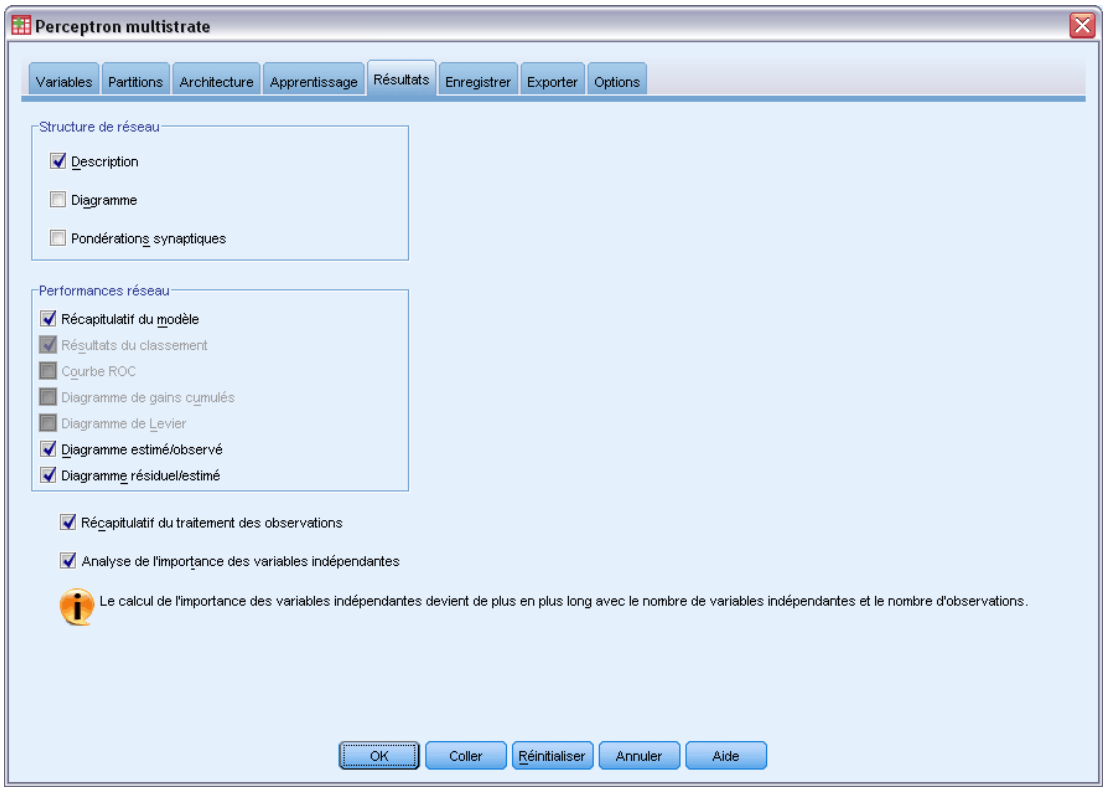
Figure 4-28
Perceptron multistrate : Onglet formation



- Sélectionnez En ligne en tant que type de formation. On suppose que la formation en ligne fonctionne bien sur des ensembles de données « plus vastes » avec des variables indépendantes corrélées. Notez que cela définit Descendant de gradient comme l'algorithme d'optimisation avec les options de défaut correspondantes.

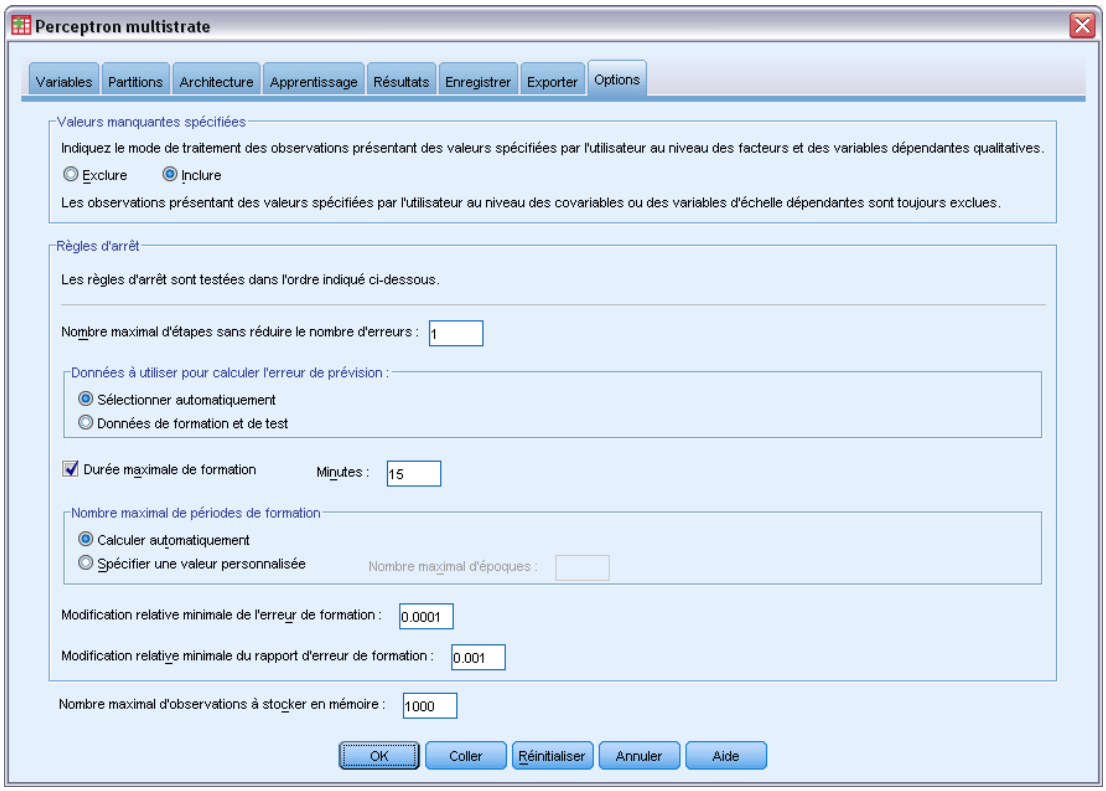
- Cliquez sur l'onglet Résultats.

Figure 4-29
Perceptron multistrate : Onglet Résultats



- Désélectionnez Diagramme. De nombreuses données apparaissent et le diagramme obtenu est peu pratique.
- Sélectionnez Diagramme estimé/observé et Valeurs résiduelles par prévisions dans le groupe de performances réseau. Les résultats de classification, les courbes ROC, le diagramme de gains cumulés et le diagramme de Levier ne sont pas disponibles parce qu'aucune des variables dépendantes n'est traitée en tant que catégorie (nominale ou ordinale).
- Sélectionnez l'option Analyse de l'importance des variables indépendantes.
- Cliquez sur l'onglet Options.

Figure 4-30
Onglet Options



- Choisissez d'inclure les variables manquantes spécifiées par l'utilisateur. Les patients qui n'ont pas subi de chirurgie ont des valeurs manquantes spécifiées par l'utilisateur dans la variable *Complications chirurgicales*. On a donc la confirmation que ces patients sont bien compris dans l'analyse.
- Cliquez sur OK.

Avertissements

Figure 4-31
Avertissements

Les variables indépendantes suivantes sont constantes dans l'échantillon d'apprentissage et sont exclues de l'analyse: doa, der.

Le tableau des avertissements indique que les variables *doa* et *der* sont constantes dans l'échantillon de formation. Les patients décédés au moment de l'arrivée ou qui sont décédés au service des urgences ont des valeurs manquantes définies par l'utilisateur pour la variable *Durée du séjour*. Etant donné que l'on traite la variable *Durée de séjour* comme variable d'échelle pour cette analyse et que les cas présentant des valeurs manquantes définies par l'utilisateur sur les variables d'échelle sont exclus, seuls sont inclus les patients encore vivants après être sortis du service des urgences.

Récapitulatif de traitement des observations

Figure 4-32
Récapitulatif du traitement des observations

		N	Pourcentage
Echantillon	Formation	5647	70,6%
	Test	1570	19,6%
	Traitement	781	9,8%
	Valide	7998	100,0%
	Exclue	2002	
	Total	10000	

Le récapitulatif du traitement des observations montre que 5 647 observations ont reçu l'échantillon de formation, 1 570 l'échantillon test et 781 l'échantillon traité. Les 2 002 observations exclues de l'analyse concernent des patients décédés sur le chemin de l'hôpital ou au service des urgences.

Informations réseau

Figure 4-33
Informations sur le réseau

Strate d'entrée	Factors	1	Catégorie d'âge
		2	Sexe
		3	Antécédent diabète
		4	Pression sanguine
		5	Fumeur
		6	Cholestérol
		7	Physiquement actif
		8	Obésité
		9	Antécédent angine
		10	Antécédent accident cardio vasculaire
		11	Nitroglyrine prescrite
		12	Médicament anti-coagulant
		13	Durée à l'hôpital
		14	Résultat EKG
		15	Résultat sanguin CPK
		16	Résultat sanguin T Troponin
		17	Médicament Disolvant caillot
		18	Hémorragie
		19	Magnésium
		20	Digital
		21	Beta blockers
		22	Traitement chirurgical
		23	Complications chirurgicales
	Nombre d'unités[a]		63
Strate(s) masquée(s)	Nombre de strates masquées		2
	Nombre d'unités dans la strate masquée 1[a]		12
	Number of Units in Hidden Layer 2[a]		9
	Fonction d'activation		Tangente hyperbolique
Strate de sortie	Dependent Variables	1	Durée de l'intervention
		2	Coût du traitement
	Nombre d'unités		2
	Rescaling Method for Scale Dependents		Adjusted Normalized
	Fonction d'activation		Tangente hyperbolique
	Fonction d'erreur		Somme des carrés

a. Exclusion de l'unité biaisée

Le tableau d'informations sur le réseau affiche des informations sur le réseau neuronal et permet de vérifier que les spécifications sont correctes. En l'occurrence, notez les points suivants :

- Le nombre d'unités dans la strate d'entrée correspond au nombre total de niveaux de facteur (il n'y a pas de covariables).

- Deux strates masquées ont été sollicitées, la procédure a choisi 12 unités dans la première strate masquée et 9 dans la seconde.
- Une unité de résultat séparée est créée pour chacune des variables d'échelle dépendantes. Elles sont rééchelonnées par la méthode normalisée ajustée, ce qui nécessite l'emploi de la fonction d'activation de la tangente hyperbolique pour la strate de résultats.
- L'erreur de la somme des carrés est signalée car les variables dépendantes sont des variables d'échelle.

Récapitulatif des modèles

Figure 4-34
Récapitulatif du modèle

Formation	Erreur de somme des carrés		183.920
	Erreur relative générale moyenne		.166
	Erreur relative pour les variables d'échelle dépendantes	Durée de l'intervention	.210
		Coût du traitement	.120
	Arrêt de la règle utilisée		1 étape(s) consécutive(s) sans diminution dans l'erreur[a]
	Durée de formation		00:00:02.987
Test	Erreur de somme des carrés		52.369
	Erreur relative générale moyenne		.171
	Erreur relative pour les variables d'échelle dépendantes	Durée de l'intervention	.218
		Coût du traitement	.124
Traitement	Erreur relative générale moyenne		.180
	Erreur relative pour les variables d'échelle dépendantes	Durée de l'intervention	.225
		Coût du traitement	.132

a. Les calculs d'erreurs sont basés sur l'échantillon de test.

Le récapitulatif du modèle affiche des informations sur les résultats de l'apprentissage du réseau final et de son application à l'échantillon traité.

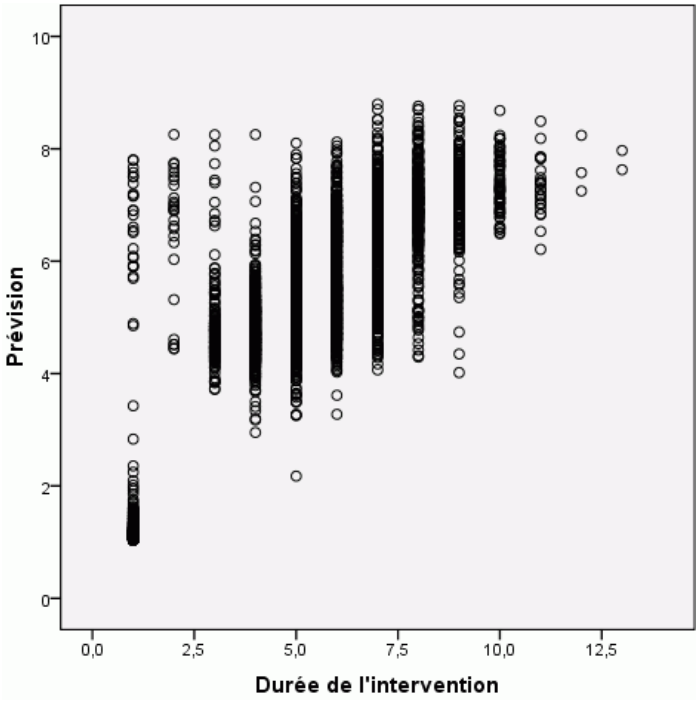
- L'erreur de la somme des carrés apparaît car la strate de résultat comporte des variables d'échelle dépendantes. Il s'agit de la fonction d'erreur que le réseau essaie de minimiser pendant l'apprentissage. Notez que les sommes des carrés et toutes les valeurs d'erreur qui en découlent sont calculées pour les valeurs rééchelonnées des variables dépendantes.
- L'erreur relative pour chaque variable d'échelle dépendante est le ratio de l'erreur de la somme des carrés pour la variable dépendante ajouté à l'erreur de la somme des carrés pour le modèle « nul » dans lequel on utilise la valeur moyenne de la variable dépendante en tant que valeur de prédiction pour chaque observation. Il semble qu'il y ait davantage d'erreurs pour les prédictions de *durée de séjour* que pour les *coûts de traitement*.
- L'erreur d'ensemble moyenne représente le ratio de l'erreur de la somme des carrés pour toutes les variables dépendantes ajouté à l'erreur de la somme des carrés pour le modèle « nul » dans lequel on utilise les valeurs moyennes des variables dépendantes en tant que valeurs de prédiction pour chaque observation. Dans cet exemple, il se trouve que l'erreur globale moyenne est proche de la moyenne des erreurs relatives, mais cela ne sera pas toujours le cas.

L'erreur relative globale moyenne et les erreurs relatives sont assez constantes lors de la formation, des tests et des échantillons traités, ce qui garantit que le modèle n'est pas surentraîné et qu'à l'avenir, l'erreur indiquée par le réseau sera proche de l'erreur mentionnée dans ce tableau.

- L'algorithme d'estimation s'est arrêté, car l'erreur n'a pas diminué après un pas dans l'algorithme.

Diagrammes estimés/observés

Figure 4-35
Diagramme estimé/observé pour la durée de séjour

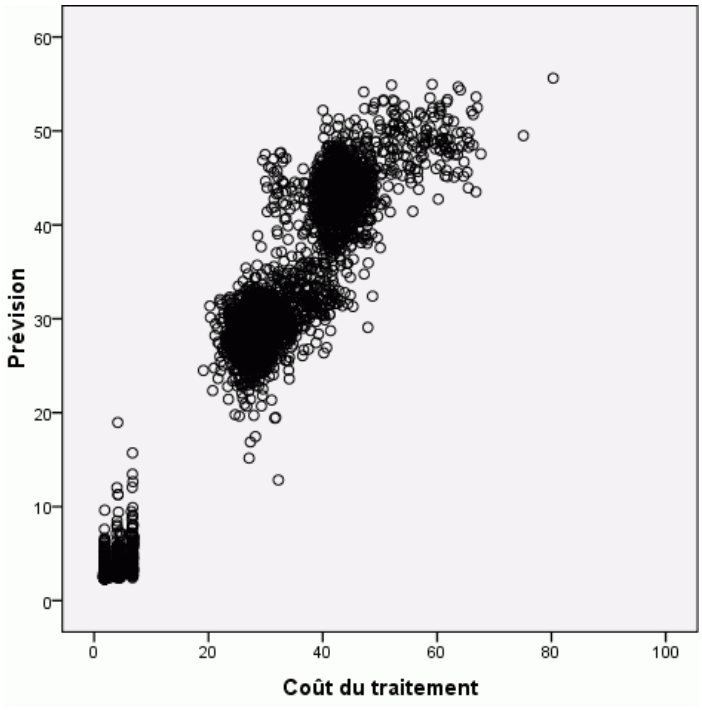


Pour les variables d'échelle dépendantes, le diagramme estimé/observé affiche un diagramme de dispersion de valeurs de prédiction sur *l'axe y*et des valeurs observées sur *l'axe x*, pour les échantillons de formation et de test. Dans l'idéal, les valeurs devraient se trouver plus ou moins le long d'une ligne de 45 degrés, qui part du point d'origine. Les points situés sur ce graphique forment des lignes verticales sur lesquelles on trouve le nombre de jours correspondant à la variable *Durée du séjour*.

En regardant le graphique, on remarque que la prévision effectuée par le réseau concernant la *durée du séjour* est plutôt efficace. La tendance générale du graphique se situe en dehors de la ligne idéale de 45 degrés dans la mesure où les prédictions pour les durées de séjour observées inférieures à 5 jours ont tendance à surestimer la durée de séjour, alors que les prédictions pour les durées de séjour observées supérieures à 6 jours ont tendance à sous-estimer cette durée.

La catégorie de patients dans la partie située en bas à gauche du graphique est susceptible de représenter les patients qui n'ont pas subi d'intervention chirurgicale. On trouve aussi une catégorie de patients dans la partie située en haut à gauche du graphique, pour lesquels la durée de séjour observée est de 1 à 3 jours et les valeurs de prédiction sont bien supérieures. Il est probable que ces observations représentent des patients décédés dans l'hôpital, après une intervention chirurgicale.

Figure 4-36
Diagramme estimé/observé pour les coûts de traitement



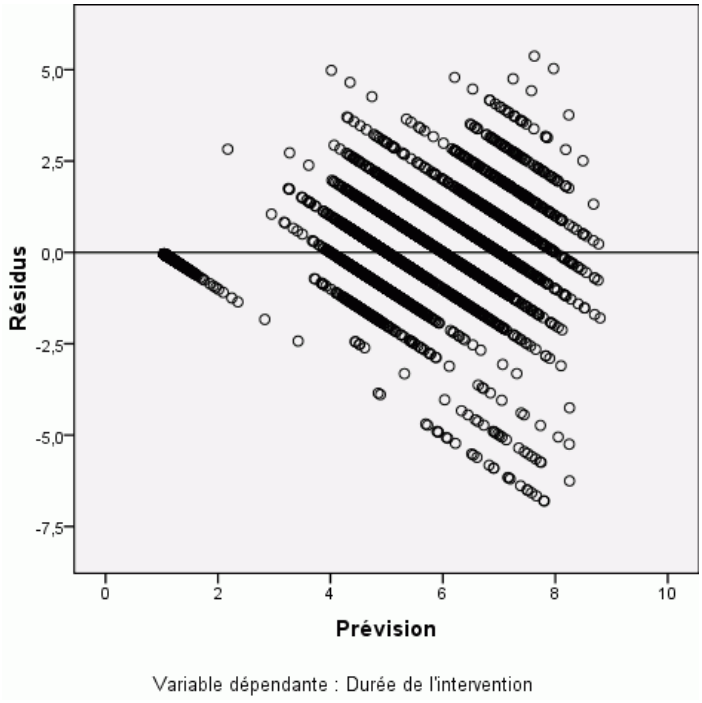
Le réseau semble aussi être raisonnablement efficace pour prédire les *coûts de traitement*. Trois catégories de patients semblent se distinguer :

- Dans la partie située en en bas à gauche figurent les patients qui n'ont pas subi d'intervention chirurgicale. Leurs coûts sont relativement faibles et différenciés par le type d'*anticoagulants [anti-coagulation]* administrés au service des urgences.
- Le coût des traitements administrés à la catégorie de patients suivante est d'environ 30 000 dollars. Il s'agit de patients qui ont subi une angiosplastie coronaire transluminale percutanée (ACTP).
- Le coût des traitements administrés à la dernière catégorie de patients dépasse les 40 000 dollars. Il s'agit de patients qui ont subi un pontage aortocoronarien (PAC). Cette opération chirurgicale est un peu plus chère que l'ACTP et la période de rétablissement chez les patients est plus longue (ce qui augmente encore un peu plus les coûts).

Il existe aussi un nombre de cas entraînant des surcoûts de 50 000 dollars qui ne sont pas bien prévus par le réseau. Il s'agit de patients qui ont connu des complications lors de l'opération chirurgicale, ce qui peut augmenter les coûts et la durée de séjour.

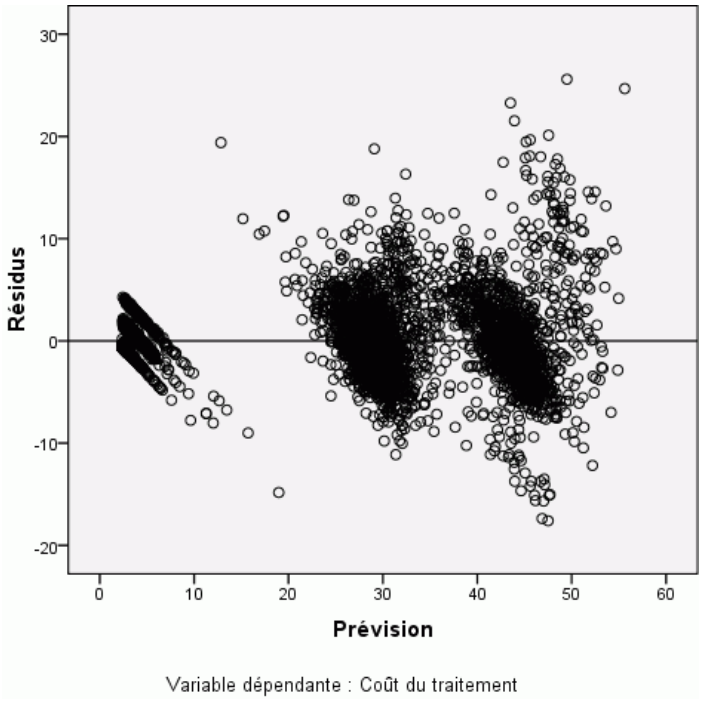
Diagrammes résiduels/estimés

Figure 4-37
Diagramme résiduel/estimé pour la durée de séjour



Le diagramme résiduel/estimé affiche un diagramme de dispersion des résidus (valeur observée moins valeur de prédiction) sur *l'axe y* et la valeur de prédiction sur *l'axe x*. Chaque ligne diagonale du graphique correspond à une ligne verticale du diagramme estimé/observé et vous pouvez davantage vous rendre compte de la progression de la surprédiction à la sous-prédiction de la durée de séjour au fur et à mesure que la durée de séjour observée augmente.

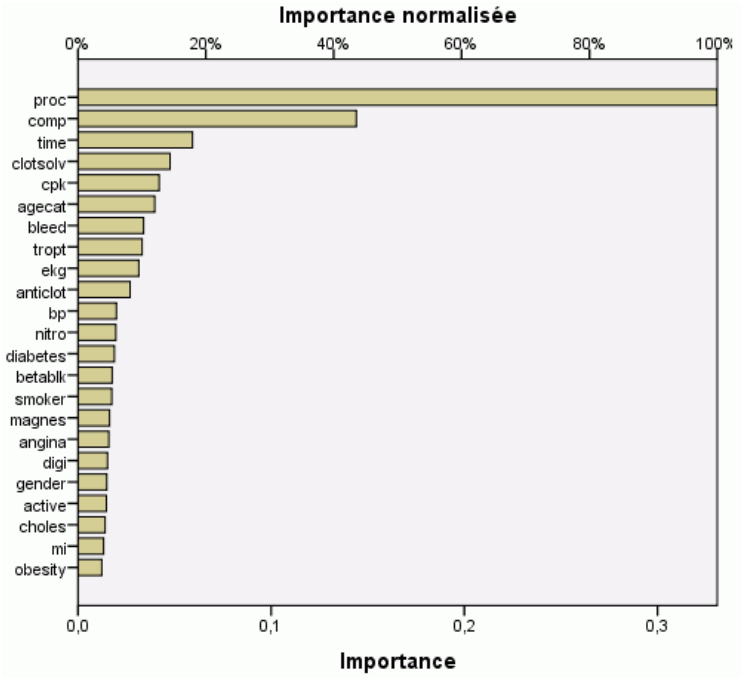
Figure 4-38
Diagramme résiduel/estimé pour les coûts de traitement



En outre, pour chacune des trois catégories de patients observés dans le diagramme estimé/observé pour la variable *Coûts de traitement*, le diagramme estimé/observé montre une progression partant d'une surprédiction à une sous-prédiction des coûts au fur et à mesure que les coûts augmentent. Les patients qui subissent des complications lors du PAC sont encore très visibles, mais il est encore plus facile de visualiser les patients qui ont subi des complications lors de l'ACTP ; ils apparaissent sous la forme d'un sous-groupe légèrement en haut à droite du groupe principal des patients qui ont subi une ACTP aux alentours de la marque des 30 000 dollars sur *l'axe x*.

Importance des variables indépendantes

Figure 4-39
Diagramme de l'importance de la variable indépendante



Le diagramme d'importance montre que les résultats sont déterminés par la procédure chirurgicale employée, suivie de l'apparition (ou non) de complications, puis par d'autres variables indépendantes. L'importance de la procédure chirurgicale est clairement visible dans les graphiques pour les *coûts de traitement*, un peu moins pour la *durée de séjour*, bien que l'effet des complications sur la *durée de séjour* soit visible chez les patients présentant les plus longues durées de séjour observées.

Récapitulatif

Le réseau semble bien fonctionner lorsqu'il prévoit des valeurs pour des patients « typiques », mais ne prend pas en compte les patients décédés après l'opération chirurgicale. Il serait possible de traiter cela en créant plusieurs réseaux. Un réseau pourrait prévoir le résultat du patient, peut-être juste pour avancer si le patient va survivre ou non. Ensuite, des réseaux séparés pourraient prévoir les *coûts de traitement* et la *durée de séjour* à condition que le patient survive. Vous pourrez ensuite combiner les résultats des réseaux et obtenir de meilleures prédictions. Vous pourrez aborder de la même manière le problème de la sous-prédiction des coûts et des durées de séjour pour les patients qui ont subi des complications lors d'une opération chirurgicale.

Lectures recommandées

Pour plus d'informations sur les réseaux neuronaux et sur les perceptrons multistrates, reportez-vous aux textes suivants :

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, 3rd éd. Oxford: Oxford University Press.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd éd. New York: Springer-Verlag.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd éd. New York: Macmillan College Publishing.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Chapitre
5

Fonction de base radiale

La procédure de fonction à base radiale (RBF) produit un modèle de prévision pour une ou plusieurs variables dépendantes (cibles) en fonction des valeurs des variables indépendantes.

Utilisation de la procédure Fonction à base radiale pour classer les clients d'un service de télécommunications

Un fournisseur de services de télécommunication a segmenté sa base de clients par type d'utilisation des services en catégorisant les clients en quatre groupes. Si les données démographiques peuvent être utilisées pour prévoir les groupes d'affectation, vous pouvez personnaliser les offres pour chaque client éventuel.

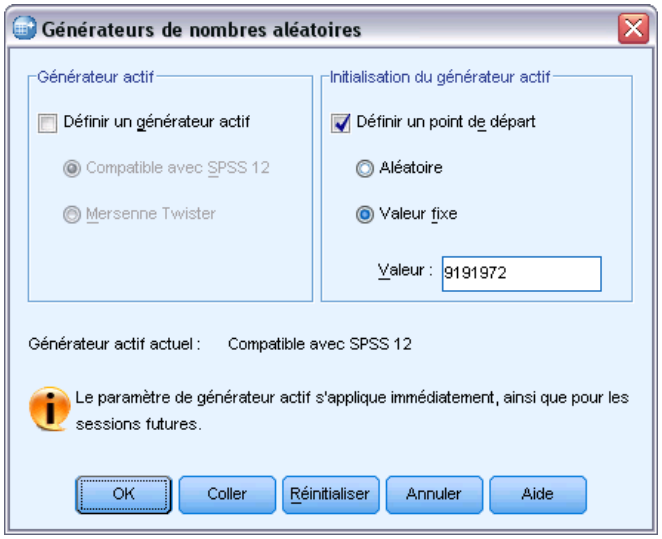
Supposez que des informations sur les clients actuels sont contenues dans le fichier *telco.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 87. Utilisez la procédure Fonction à base radiale pour classer les clients.

Préparation des données pour l'analyse

Définir le générateur aléatoire vous permet de reproduire l'analyse exactement.

- Pour définir le générateur aléatoire, à partir des menus, sélectionnez :
  Transformer > Générateurs de nombres aléatoires…

Figure 5-1
Boîte de dialogue Générateurs de nombres aléatoires

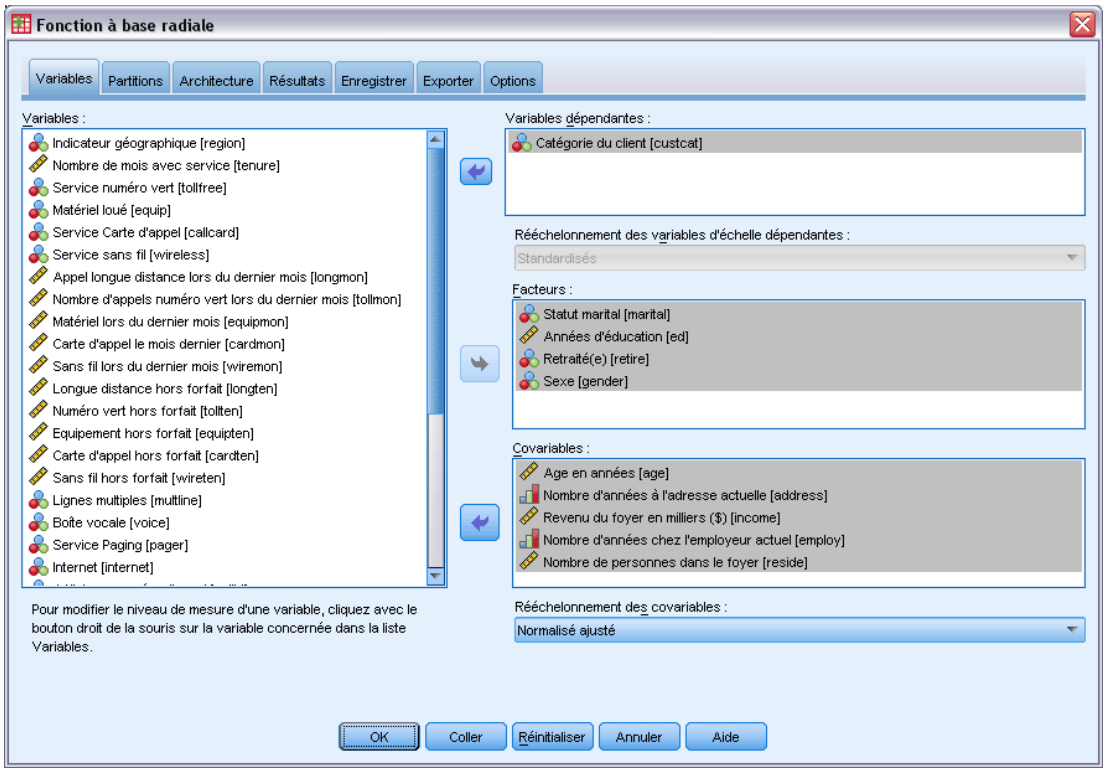


- Sélectionnez Définir un point de départ.
- Sélectionnez Valeur fixe et tapez la valeur 9 191 972.
- Cliquez sur OK.

Exécution de l'analyse

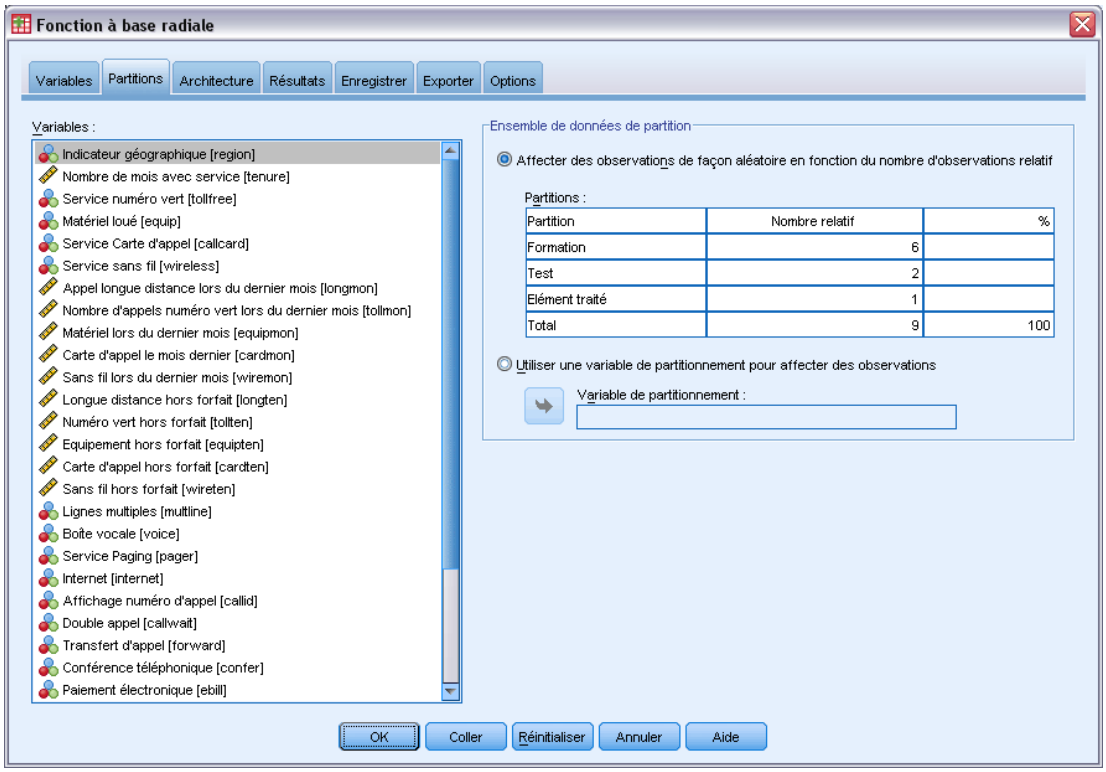
- Pour exécuter une analyse Fonction à base radiale, à partir des menus, sélectionnez : Analyse > Réseaux neuronaux : > Fonction à base radiale...

Figure 5-2
Fonction à base radiale : l'onglet Variables



- Sélectionnez *Catégorie de client [catclient]* comme variable dépendante.
- Sélectionnez *Marital status [marital]*, *Level of education [ed]*, *Retired [retire]* et *Gender* comme facteurs.
- Sélectionnez *Age in years [age]* et *Number of people in household [reside]* comme covariables.
- Sélectionnez Adjusted Normalized comme méthode pour redimensionner les covariables.
- Cliquez sur l'onglet Partitions.

Figure 5-3
Fonction à base radiale : Onglet Partitions



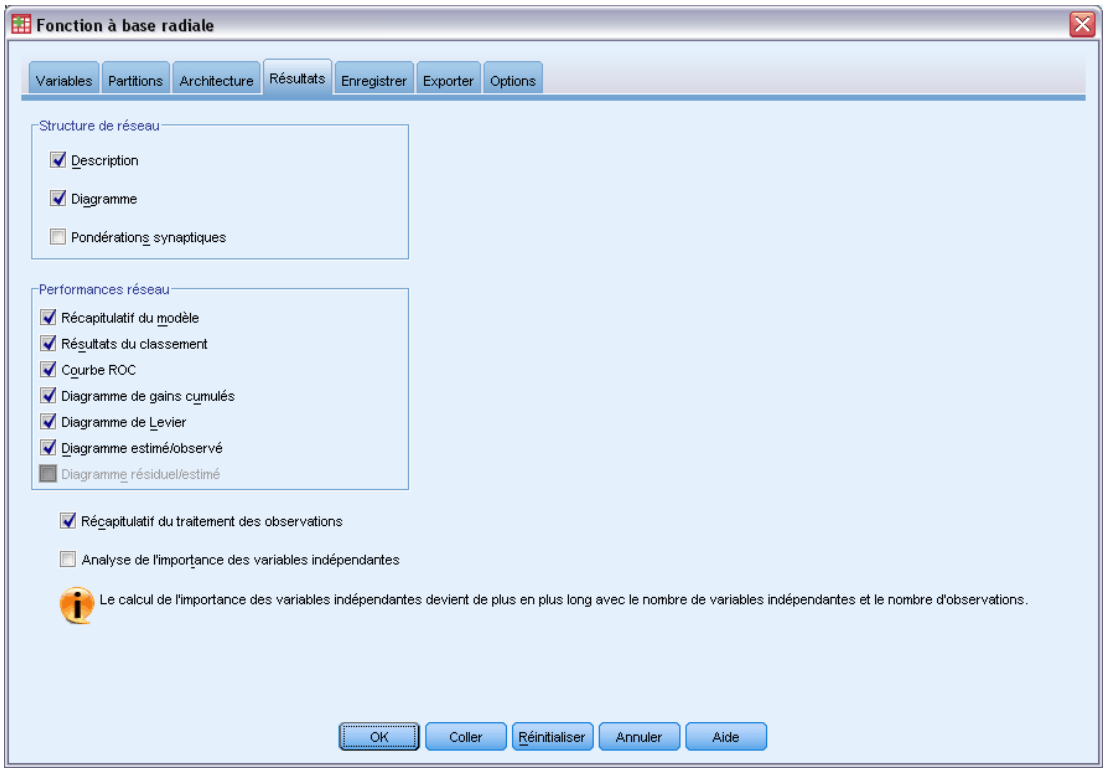
En indiquant le nombre d'observations relatif, il est facile de créer des partitions fractionnelles pour lesquelles il serait difficile d'indiquer des pourcentages. Supposons que vous voulez attribuer les 2/3 de l'ensemble de données à l'échantillon de formation, et les 2/3 des observations restantes aux tests.

- Tapez 6 comme nombre relatif pour l'échantillon d'apprentissage.
- Tapez 2 comme nombre relatif pour l'échantillon de test.
- Tapez 1 comme nombre relatif pour l'échantillon traité.

Un total de 9 observations relatives a été indiqué. 6/9 = 2/3, ou environ 66,67 %, sont attribués à l'échantillon d'apprentissage ; 2/9, ou environ 22,22 %, sont attribués à l'échantillon de test ; 1/9, ou environ 11,11 % sont attribués à l'échantillon traité.

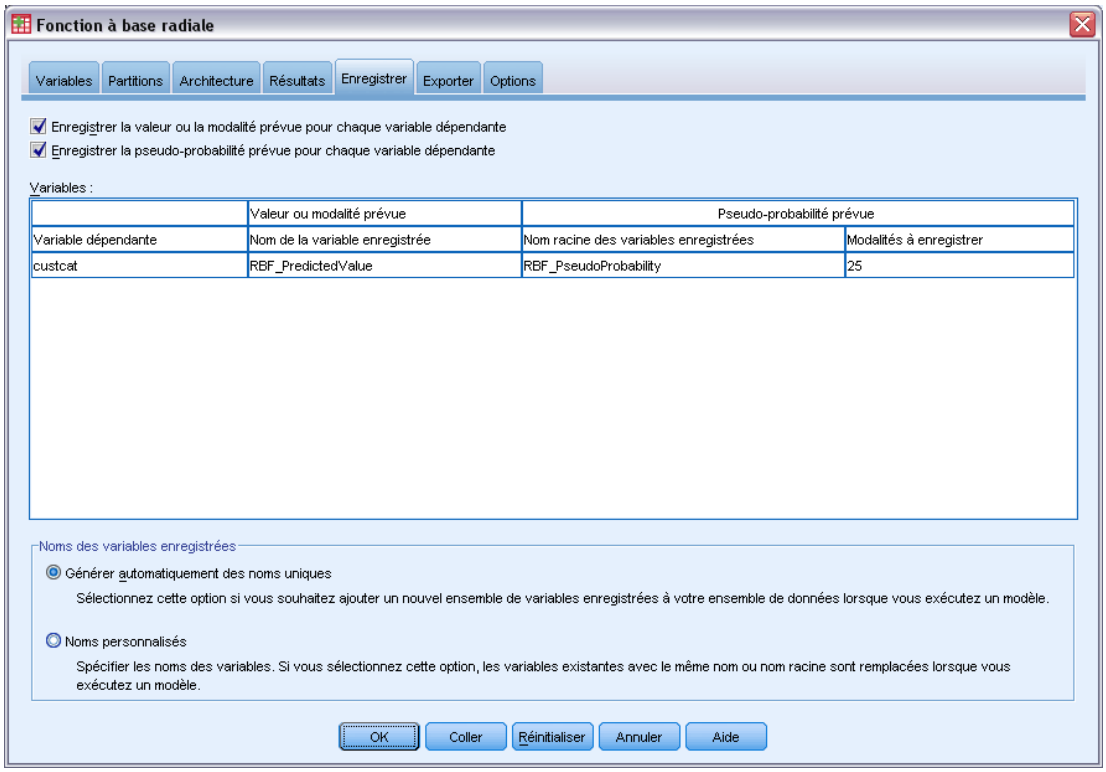
- Cliquez sur l'onglet Résultats.

Figure 5-4
Fonction à base radiale : Onglet Résultats



- Désélectionnez l'option Diagramme dans le groupe Structure de réseau.
- Sélectionnez les options Courbe ROC, Diagramme de gains cumulés, Diagramme de Levier et Diagramme estimé/observé dans le groupe Performances réseau.
- Cliquez sur l'onglet Enregistrer.

Figure 5-5
Fonction à base radiale : Onglet Enregistrer



- Sélectionnez Save predicted value or category for each dependent variable et Save predicted pseudo-probability for each dependent variable.
- Cliquez sur OK.

Récapitulatif de traitement des observations

Figure 5-6
Récapitulatif du traitement des observations

		N	Pourcentage
Echantillon	Formation	665	66,5%
	Test	224	22,4%
	Traitement	111	11,1%
	Valide	1000	100,0%
	Exclue	0	
	Total	1000	

Le récapitulatif du traitement des observations montre que 665 observations ont reçu l'échantillon de formation, 224 l'échantillon test et 111 l'échantillon traité. Ces observations sont exclues de l'analyse.

Informations réseau

Figure 5-7
Informations sur le réseau

Factors	1	Statut marital
	2	Années d'éducation
	3	Retraité(e)
	4	Sexe
Covariates	1	Age en années
	2	Nombre de personnes dans le foyer
	3	Nombre d'années à l'adresse actuelle
	4	Revenu du foyer en milliers ($)
	5	Nombre d'années chez l'employeur actuel
5	Nombre d'unités	16
	Rescaling Method for Covariates	Adjusted Normalized
	Nombre d'unités	9[a]
	Fonction d'activation	MaxMou
Dependent Variables	1	Catégorie du client
7	Nombre d'unités	4
	Fonction d'activation	Identité
	Fonction d'erreur	Somme des carrés

a. Déterminé par le critère des données de test : le "meilleur" nombre d'unités masquées est celui qui présente la plus petite erreur dans les données de test.

Le tableau d'informations sur le réseau affiche des informations sur le réseau neuronal et permet de vérifier que les spécifications sont correctes. En l'occurrence, notez les points suivants :

- Le nombre d'unités dans la strate d'entrée correspond au nombre de covariables plus le nombre total de niveaux de facteur ; une unité spécifique est créée pour chaque modalité de *Marital status*, *Level of education*, *Retired* et *Gender* et aucune des modalités n'est considérée comme une unité « redondante », comme cela est courant dans de nombreuses procédures de modélisation.
- De même, une unité de résultat spécifique est créée pour chaque modalité de *client*, pour un total de 4 unités dans la strate de résultat.
- Les covariables sont rééchelonnées grâce à la méthode normalisée ajustée.
- La sélection automatique de l'architecture a choisi 9 unités dans la strate masquée.
- Toutes les autres informations sur le réseau correspondent aux valeurs par défaut pour la procédure.

Récapitulatif des modèles

Figure 5-8
Récapitulatif du modèle

Test	Prévisions de pourcentage incorrectes	67,4%

Le récapitulatif du modèle affiche des informations sur les résultats de l'apprentissage du réseau final, des tests et de son application à l'échantillon traité.

- L'erreur de la somme des carrés apparaît car elle est toujours utilisée pour les réseaux RBF. Il s'agit de la fonction d'erreur que le réseau essaie de réduire pendant l'apprentissage et les tests.
- Le pourcentage de prévisions incorrectes provient du tableau de classement et sera abordé plus loin dans cette section.

Classification

Figure 5-9
Classification

Echantillon	Observations	Estimé				
		Service basic	Service électronique	Service plus	Service Total	Percent Correct
Formation	Service basic	64	0	66	45	36,6%
	Service électronique	22	1	57	61	,7%
	Service plus	47	0	104	34	56,2%
	Service Total	29	1	49	85	51,8%
	Overall Percent	24,4%	,3%	41,5%	33,8%	38,2%
Test	Service basic	18	0	26	15	30,5%
	Service électronique	15	0	16	22	,0%
	Service plus	11	0	39	15	60,0%
	Service Total	4	0	17	26	55,3%
	Overall Percent	21,4%	,0%	43,8%	34,8%	37,1%
Traitement	Service basic	11	0	11	10	34,4%
	Service électronique	4	0	9	10	,0%
	Service plus	10	0	19	2	61,3%
	Service Total	5	0	5	15	60,0%
	Overall Percent	27,0%	,0%	39,6%	33,3%	40,5%

Variable dépendante : Catégorie du client

Le tableau de classement affiche les résultats pratiques de l'utilisation du réseau. Pour chaque observation, la réponse prévue est la modalité dotée de la pseudo-probabilité prévue la plus élevée du modèle.

- Les cellules de la diagonale sont des prévisions correctes.
- Les cellules hors de la diagonale sont des prévisions incorrectes.

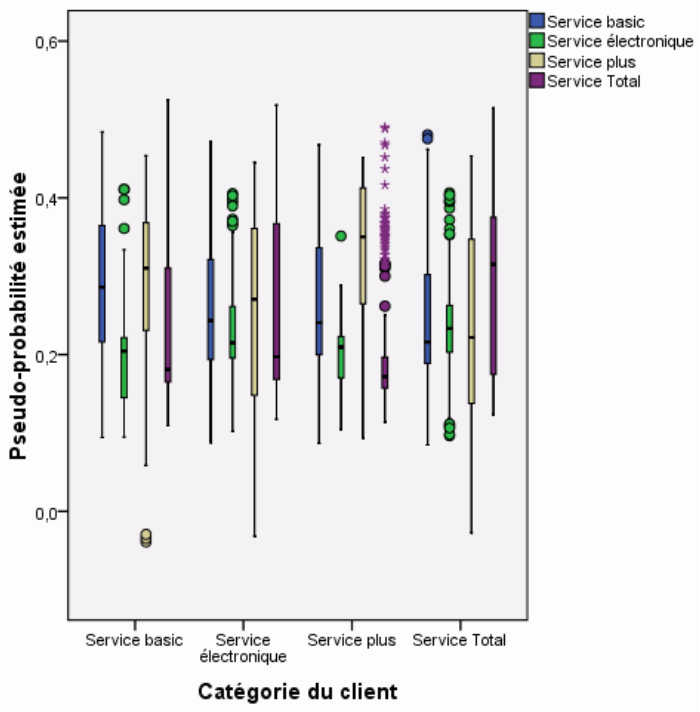
Etant donné les données observées, le modèle « nul » (qui est un modèle sans variable indépendante) classerait tous les clients dans le groupe modal *Service plus*. Ainsi, le modèle nul serait correct pour 281/1000 = 28,1 % des observations. Le réseau RBF obtient 10,1 % de plus, ou 38,2 % des clients. Votre modèle excelle particulièrement dans l'identification des clients *Plus service* et *Total service*. Cependant, il n'est pas très adapté au classement des clients *E-service*. Vous devez peut-être trouver une autre variable indépendante afin de distinguer ces clients ; ou bien, étant donné que ces clients sont la plupart du temps classés à tort comme clients *Plus service* et *Total service*, la société pourrait simplement essayer de surclasser les clients potentiels qui appartiennent normalement à la modalité *E-service*.

Les classements basés sur les observations utilisées pour créer le modèle tendent à être trop « optimistes » dans le sens où leur taux de classification est augmenté. L'échantillon traité permet de valider le modèle ; en l'occurrence, le modèle a correctement classé 40,2 % de ces

observations. Bien que l'échantillon traité soit plutôt réduit, il suggère que votre modèle est en fait correct environ deux fois sur cinq.

Diagramme estimé/observé

Figure 5-10
Diagramme estimé/observé



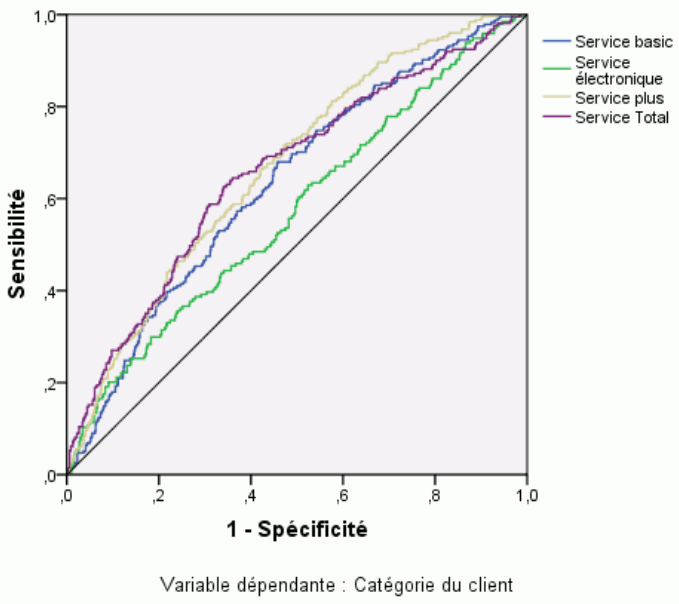
Dans le cas des variables dépendantes qualitatives, le diagramme estimé/observé affiche des boîtes à moustaches juxtaposées de pseudo-probabilités prévues pour les échantillons d'apprentissage et de test combinés. L'axe des *X* correspond aux modalités de réponses observées, et la légende aux modalités estimées. Ainsi :

- La boîte à moustaches le plus à gauche montre, pour les observations ayant comme modalité observée *Basic service*, la pseudo-probabilité prévue de la modalité *Basic service*.
- La boîte à moustaches suivante vers la droite montre, pour les observations ayant comme modalité observée *Basic service*, la pseudo-probabilité prévue de la modalité *E-service*.
- La troisième boîte à moustaches montre, pour les observations ayant comme modalité observée *Basic service*, la pseudo-probabilité prévue de la modalité *Plus service*. D'après le tableau de classement, presque autant de clients *Service de base* étaient classés de manière erronée comme clients *Service Plus* que correctement classés comme clients *Service de base* ; par conséquent, cette boîte à moustaches est presque équivalente à celle le plus à gauche.
- La quatrième boîte à moustaches montre, pour les observations ayant comme modalité observée *Basic service*, la pseudo-probabilité prévue de la modalité *Total service*.

Puisqu'il existe plus de deux modalités dans la variable cible, les quatre premières boîtes à moustaches ne sont pas symétriques par rapport à la ligne horizontale au niveau de 0,5, ni d'une quelconque autre façon. Par conséquent, l'interprétation de ce diagramme pour des cibles comportant plus de deux modalités peut s'avérer difficile car il est impossible de déterminer, à partir de l'observation d'une partie des observations dans une boîte à moustaches, l'emplacement correspondant de ces observations dans une autre boîte à moustaches.

Courbe ROC

Figure 5-11
Courbe ROC



La courbe ROC présente un affichage visuel de **sensibilité** par **spécificité** pour toutes les césures de classement possibles. Le diagramme présenté ci-après présente quatre courbes, une pour chaque modalité de la variable cible.

Ce diagramme repose sur la combinaison de l'échantillon d'apprentissage et de l'échantillon de test. Pour obtenir un diagramme ROC pour l'échantillon traité, scindez le fichier au niveau de la variable de partitionnement, puis exécutez la procédure Courbe ROC sur les pseudo-probabilités prévues.

Figure 5-12
Zone inférieure à la courbe

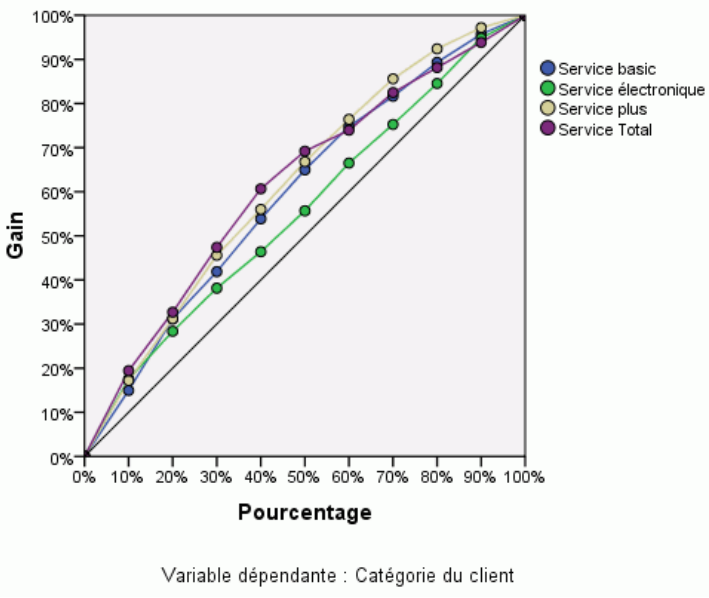
Zone sous la courbe

		Zone
Catégorie du client	Service basic	,635
	Service électronique	,573
	Service plus	,668
	Service Total	,659

La zone inférieure à la courbe est un récapitulatif numérique de la courbe ROC, tandis que les valeurs du tableau représentent, pour chaque modalité, la probabilité que la présence de la pseudo-probabilité prévue dans cette modalité soit supérieure pour une observation choisie aléatoirement appartenant à cette modalité que pour une observation choisie aléatoirement n'appartenant pas à cette modalité. Par exemple, pour un client sélectionné aléatoirement dans *Service Plus* et un client sélectionné aléatoirement dans *Service de base*, *Service en ligne* ou *Service Total*, il existe une probabilité de 0,668 que la pseudo-probabilité de manquement prévue par le modèle soit plus élevée pour le client dans *Service Plus*.

Diagrammes de gains cumulés et de Levier

Figure 5-13
Diagramme de gains cumulés

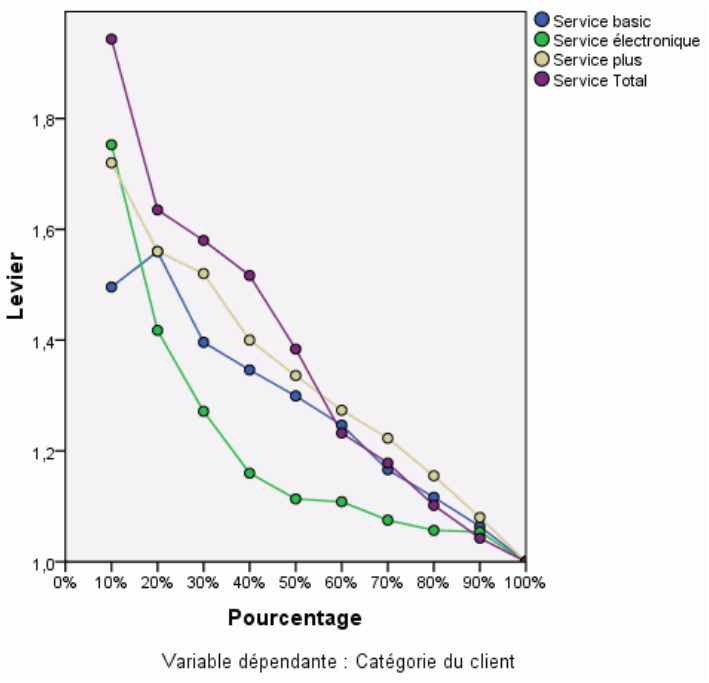


Le diagramme de gains cumulés montre le pourcentage du nombre total d'observations dans une modalité donnée obtenu en ciblant un pourcentage du nombre total d'observations. Par exemple, le premier point de la courbe pour la modalité *Service Total* se situe approximativement à (10 %, 20 %), ce qui signifie que si vous évaluez un ensemble de données avec le réseau et que vous triez toutes les observations en fonction de la pseudo-probabilité prévue de la modalité *Service Total*, vous pouvez vous attendre à ce que la tranche supérieure de 10 % contienne approximativement

20 % de la totalité des observations qui ont véritablement la modalité *Service Total*. De même, la tranche supérieure de 20 % contiendrait approximativement 30 % des personnes manquant à leurs engagements, la tranche supérieure de 30 % des observations comporterait 50 % des personnes manquant à leurs engagements, et ainsi de suite. Si vous sélectionnez 100 % de l'ensemble de données évalué, vous obtenez la totalité des personnes manquant à leurs engagements dans l'ensemble de données.

La diagonale correspond à la courbe « de référence » ; si vous sélectionnez aléatoirement 10 % des observations dans l'ensemble de données évalué, vous pouvez espérer « obtenir » approximativement 10 % de la totalité des observations qui ont véritablement une modalité donnée. Plus une courbe se situe au-dessus de la ligne de base, plus le gain est élevé.

Figure 5-14
Diagramme de Levier



Le diagramme de Levier est issu du diagramme de gains cumulés ; les valeurs de l'axe des *Y* correspondent au ratio du gain cumulé pour chaque courbe par rapport à la ligne de base. Par conséquent, le levier à 10 % pour la modalité *Total service* est 20 %/10 % = 2,0. Il permet d'observer différemment les informations du diagramme de gains cumulés.

Remarque : Le diagramme de gains cumulés et le diagramme de Levier reposent sur la combinaison de l'échantillon d'apprentissage et de l'échantillon de test.

Lectures recommandées

Pour plus d'informations sur la fonction à base radiale, reportez-vous aux textes suivants :

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, 3rd éd. Oxford: Oxford University Press.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd éd. New York: Springer-Verlag.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd éd. New York: Macmillan College Publishing.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. Dans : *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers,* A. Singh, éd. Los Alamitos, CA: IEEE Comput. Soc. Press.

Uykan, Z., C. Guzelis, M. E. Celebi, et H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. *IEEE Transactions on Neural Networks*, 11, .

Annexe

A

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espangol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.**Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs(Van der Ham, Meulman, Van Strien, et Van Engeland, 1997) ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique (Price et Bouffard, 1974), on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique (Green et Rao, 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.
- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et Nelder, 1989) qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.

- **carpet.sav.** Dans un exemple courant (Green et Wind, 1973), une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé (Kennedy, Riquier, et Sharp, 1996). Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le

montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.

- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.

- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman(Rickman, Mitchell, Dingman, et Dalen, 1974). Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases (Blake et Merz, 1998) de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui chercher à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell (Bell, 1961) a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman (Guttman, 1968) a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.

- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim (Rosenberg et Kim, 1975) se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér*(ation) et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.**Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents (Breiman et Friedman, 1985), (Hastie et Tibshirani, 1990), ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.
- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.

- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.**Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.**Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous.(Hartigan, 1975).
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et al., 1989) et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».

- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (*http://dx.doi.org/10.3886/ICPSR02934*) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stocks.sav** Ce fichier de données hypothétiques contient le cours et le volume des actions pour un an.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes

enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.

- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs (Collett et al., 2003).
- **verd1985.sav.** Ce fichier de données concerne une enquête (Verdegaal, 1985). Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants (Ware, Dockery, Spiro III, Speizer, et Ferris Jr., 1984). Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont

été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

- **worldsales.sav** Ce fichier de données hypothétiques contient les revenus des ventes par continent et par produit.

Annexe

B

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux Etats-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com/*.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Bibliographie

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*, 3rd éd. Oxford: Oxford University Press.

Blake, C. L., et C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., et J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 éd. Boca Raton: Chapman & Hall/CRC.

Fine, T. L. 1999. *Feedforward Neural Network Methodology*, 3rd éd. New York: Springer-Verlag.

Green, P. E., et V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., et Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., et R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Haykin, S. 1998. *Neural Networks: A Comprehensive Foundation*, 2nd éd. New York: Macmillan College Publishing.

Kennedy, R., C. Riquier, et B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

McCullagh, P., et J. A. Nelder. 1989. *Generalized Linear Models*, 2nd éd. Londres: Chapman & Hall.

Price, R. H., et D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, et J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenberg, S., et M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Tao, K. K. 1993. A closer look at the radial basis function (RBF) networks. Dans : *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers,* A. Singh, éd. Los Alamitos, CA: IEEE Comput. Soc. Press.

Uykan, Z., C. Guzelis, M. E. Celebi, et H. N. Koivo. 2000. Analysis of input-output clustering for determining centers of RBFN. *IEEE Transactions on Neural Networks*, 11, .

Van der Ham, T., J. J. Meulman, D. C. Van Strien, et H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en néerlandais)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, et B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Index