REVUE DE STATISTIQUE APPLIQUÉE

ALICE GUÉGUEN
JAVIER NICOLAU
JEAN-PIERRE NAKACHE

Utilisation des réseaux probabilistes en analyse discriminante sur variables qualitatives

Revue de statistique appliquée, tome 44, n° 1 (1996), p. 55-75
<http://www.numdam.org/item?id=RSA_1996__44_1_55_0>

Rev. Statistique Appliquée, 1996, **XLIV** (1), 55-75

UTILISATION DES RÉSEAUX PROBABILISTES EN ANALYSE DISCRIMINANTE SUR VARIABLES QUALITATIVES

Alice Guéguen, Javier Nicolau, Jean-Pierre Nakache

INSERM Unité 88
Hôpital National de Saint-Maurice.
14, rue du Val d'Osne. 94410 Saint-Maurice

1. Introduction

La méthode de discrimination présentée ici est une méthode adaptée au cas où les variables explicatives sont nominales et en nombre important. Plus précisément cette méthode utilise l'approche qui consiste à estimer les densités de probabilité par groupe, puis à l'aide de la formule de Bayes, d'expliciter les probabilités *a posteriori* d'appartenance aux différents groupes.

Dans cette approche, un modèle paramétrique apparaît très naturellement. Il consiste à supposer que, pour chaque groupe, les fréquences observées dans les différents états possibles sont les réalisations d'une loi multinomiale : il s'agit du modèle multinomial complet. En pratique, dès que le nombre de variables est élevé, la construction d'un modèle s'impose : le modèle le plus simple est le modèle d'indépendance conditionnelle qui consiste à estimer la densité de probabilité dans un groupe donné par le produit des fréquences des différentes variables explicatives. On suppose ainsi que dans chacun des groupes à discriminer il n'existe pas de liaison entre les variables explicatives. Plusieurs modèles se situant entre le modèle d'indépendance et le modèle multinomial complet ont été proposés parmi lesquels le modèle de Parzen, le modèle de Bahadur, le modèle log-linéaire et le modèle de Lancaster. Cependant pour ces modèles intermédiaires la difficulté de la modélisation persiste si le nombre de variables est élevé.

Les réseaux probabilistes fournissent aisément des estimations des densités de probabilité quelle que soit l'importance du nombre de variables explicatives. Ils constituent un formalisme de représentation des connaissances associant un graphe et une distribution de probabilité; les sommets du graphe représentent des variables aléatoires et la structure de la densité de probabilité conjointe des variables aléatoires se déduit sans ambiguïté de la structure du graphe. Le réseau probabiliste le plus simple, constitué uniquement de sommets, conduit au modèle d'indépendance; le réseau probabiliste dans lequel les sommets sont reliés par des arêtes, permet la prise en compte des associations de variables deux à deux alors que le réseau probabiliste constitué de triangles permet la prise en compte des associations de variables trois à trois.

Notations

Les groupes *a priori* sont notés $G_1, \cdots, G_k$ et ont pour probabilité *a priori* $\pi_1, \cdots, \pi_k$ qui vérifient $\pi_r > 0$ pour $r = 1, \cdots, k$ et $\sum \pi_r = 1$. Les probabilités a priori π_r, $r = 1, \cdots, k$ étant rarement connues, elles peuvent être soit estimées par les fréquences des groupes si le schéma d'échantillonnage le permet, soit évaluées à partir de connaissances antérieures. Les coûts de mauvais classement résultant de l'affectation d'une observation du groupe G_r au groupe G_s sont notés $c_{s|r}$ pour $s = 1, \cdots, k$ et $r = 1, \cdots, k$. Les coûts $c_{s|r}$ sont positifs si $s \neq r$ et nuls si $s = r$. En général $c_{s|r}$ est différent de $c_{r|s}$.

On note $\mathbf{X} = (X_1, \cdots, X_p)$ le vecteur de dimension p des variables aléatoires explicatives. La variable X_j, $(1 \leqslant j \leqslant p)$ prend les modalités $1, \cdots, m_j, \cdots, M_j$. $M = M_1 \times \cdots \times M_p$ représente le nombre d'états ou de profils possibles du vecteur $\mathbf{X}$. On note $P(\mathbf{X} = \mathbf{x}|G_r)$ la densité de probabilité de $\mathbf{X}$ dans le groupe G_r, $P_{obs}(\mathbf{X} = \mathbf{x}|G_r)$ la densité de probabilité observée dans le groupe G_r à l'aide de l'échantillon d'apprentissage et $\widehat{P}(\mathbf{X} = \mathbf{x}|G_r)$ la densité de probabilité dans le groupe G_r estimée par les réseaux probabilistes avec arêtes ou triangles.

2. Modèles graphiques

Dans les modèles log-linéaires hiérarchiques décomposables, les estimations du maximum de vraisemblance peuvent s'exprimer sous forme explicite, sans avoir recours à une procédure itérative. Ces modèles peuvent se représenter graphiquement, et à tout modèle graphique correspond un modèle log-linéaire hiérarchique décomposable [Agresti 1990].

Prenons un exemple dans lequel on a 4 variables explicatives. Le modèle log-linéaire saturé consiste à écrire le logarithme de la densité de probabilité dans un groupe donné sous la forme suivante :

$$\ln P(X_1 = m_1, X_2 = m_2, X_3 = m_3, X_4 = m_4) =$$

$$u + \sum_{j_1=1}^{4} u_{j_1}^{(m_{j_1})} + \sum_{j_1<j_2} u_{j_1 j_2}^{(m_{j_1} m_{j_2})} + \sum_{j_1<j_2<j_3} u_{j_1 j_2 j_3}^{(m_{j_1} m_{j_2} m_{j_3})} + u_{1\cdots,4}^{(m_1 \cdots, m_4)}, \quad (1)$$

où u représente l'effet global, $u_{j_1}^{(m_{j_1})}$ les effets principaux dûs aux variables, $u_{j_1 j_2}^{(m_{j_1} m_{j_2})}$ les effets des interactions entre les variables deux à deux, $u_{j_1 j_2 j_3}^{(m_{j_1} m_{j_2} m_{j_3})}$ les effets des interactions entre les variables trois à trois et $u_{1\cdots,4}^{(m_1 \cdots, m_4)}$ l'effet de l'interaction entre les quatre variables.

A ce modèle correspond le modèle graphique représenté dans la figure 1. Les sommets du graphe représentent les variables, les arêtes représentent les liaisons entre deux variables, les triangles représentent les liaisons entre trois variables et le volume de la pyramide représente l'interaction entre les quatre variables.

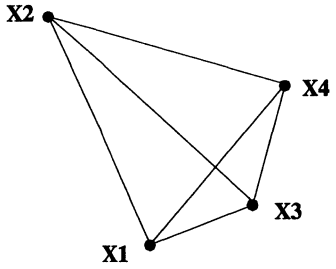


FIGURE 1
Modèle graphique associé au modèle log-linéaire saturé

Au modèle log-linéaire non saturé hiérarchique et décomposable donné par :

$$\begin{aligned}\ln P(X_1 = m_1, X_2 = m_2, X_3 = m_3, X_4 = m_4) = \\ u + u_1^{(m_1)} + u_2^{(m_2)} + u_3^{(m_3)} + u_4^{(m_4)} \\ + u_{13}^{(m_1 m_3)} + u_{14}^{(m_1 m_4)} + u_{23}^{(m_2 m_3)} + u_{24}^{(m_2 m_4)} + u_{34}^{(m_3 m_4)} \\ + u_{134}^{(m_1 m_3 m_4)} + u_{234}^{(m_2 m_3 m_4)},\end{aligned} \tag{2}$$

correspond le modèle graphique représenté dans la figure 2.

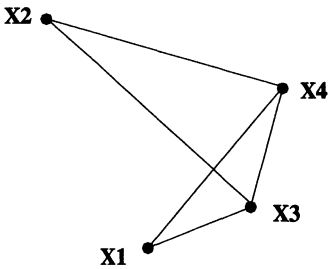


FIGURE 2
Modèle graphique associé au modèle log-linéaire non saturé défini en (2)

Le modèle log-linéaire non décomposable défini par :

$$\begin{aligned}\ln P(X_1 = m_1, X_2 = m_2, X_3 = m_3, X_4 = m_4) = \\ u + u_1^{(m_1)} + u_2^{(m_2)} + u_3^{(m_3)} + u_4^{(m_4)} \\ + u_{12}^{(m_1 m_2)} + u_{13}^{(m_1 m_3)} + u_{14}^{(m_1 m_4)} \\ + u_{23}^{(m_2 m_3)} + u_{24}^{(m_2 m_4)} + u_{34}^{(m_3 m_4)},\end{aligned} \tag{3}$$

ne peut pas être représenté graphiquement. En effet, les arêtes $u_{13}^{(m_1 m_3)}$, $u_{14}^{(m_1 m_4)}$ et $u_{34}^{(m_3 m_4)}$ définissent un triangle auquel ne correspond pas d'interaction dans le modèle; l'effet $u_{134}^{(m_1 m_3 m_4)}$ n'y est pas inclus.

De manière générale, un modèle graphique est défini par le couple (G, P^G) où G est un graphe fini non orienté associé à une distribution de probabilité P^G, dans lequel les sommets représentent des variables aléatoires $X_1, \cdots, X_p$.

L'intérêt des modèles graphiques est qu'ils permettent d'obtenir directement les estimateurs du maximum de vraisemblance sans utiliser de procédure itérative. L'utilisation de tels modèles présentent cependant des inconvénients quand le nombre de variables est important, ce qui explique l'intérêt que nous avons porté aux *réseaux probabilistes*, modèles graphiques particuliers définis dans les paragraphes ci-dessous.

2.1. Réseaux probabilistes avec prise en compte des liaisons d'ordre 1 : graphe avec arêtes

Un réseau probabiliste avec arêtes est un modèle graphique dans lequel les arêtes – représentant les liaisons entre variables – ne forment pas de cycle. Dans ce cas, le graphe qui lui est associé est un graphe connexe; la distribution de probabilité P^G peut s'écrire [Pearl 1988] comme un produit de distributions de probabilités conditionnelles :

$$\begin{aligned} P^G(\mathbf{X} = \mathbf{x}) &= P^G(X_1 = m_1, X_2 = m_2, \cdots, X_p = m_p) \\ &= P(X_1 = m_1) \prod_{j=2}^{p} P(X_j = m_j | X_{i(j)} = m_i), \end{aligned} \tag{4}$$

où $X_{i(j)}$ est la variable qui désigne le parent de X_j. La racine X_1 du graphe peut être arbitrairement choisie et, n'ayant pas de parent, elle est caractérisée par les probabilités $P(X_1 = m_1), m_1 = 1, \cdots, M_1$.

En pratique, on cherche à déterminer la distribution $\widehat{P}$ – appartenant à la famille décrite en (4) – qui approche au mieux la distribution P_{obs} observée dans l'échantillon d'apprentissage dans un groupe donné. En d'autres termes, parmi tous les graphes qu'on peut construire à partir de p variables explicatives, chacun d'entre eux conduisant à une densité $\widehat{P}$, quelle est la densité $\widehat{P}$ la plus proche de P_{obs} ? Chow et Liu (1968) ont fourni une réponse utilisant la distance de Kullback-Leibler (1951) :

$$D(P, P') = \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}) \ln \frac{P(\mathbf{X} = \mathbf{x})}{P'(\mathbf{X} = \mathbf{x})}. \tag{5}$$

Cette distance est non négative et prend la valeur 0 si et seulement si P' coïncide avec P. A partir de la définition de l'information mutuelle, notée $I(X_{j_1}, X_{j_2})$, entre deux

variables :

$$I(X_{j_1}, X_{j_2}) = \sum_{m_{j_1}=1}^{M_{j_1}} \sum_{m_{j_2}=1}^{M_{j_2}} P(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2}) \ln \frac{P(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2})}{P(X_{j_1} = m_{j_1}) P(X_{j_2} = m_{j_2})},$$

Chow et Liu ont proposé l'algorithme suivant qui permet de construire le graphe et d'en déduire la densité de probabilité qui lui est associée :

1. Calculer, à partir de l'échantillon d'apprentissage, les probabilités $P_{obs}(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2})$ pour tous les couples de variables X_{j_1} et X_{j_2}.
2. Calculer les informations mutuelles observées $I_{obs}(X_{j_1}, X_{j_2})$ des $p(p-1)/2$ arêtes (X_{j_1}, X_{j_2}) et les ranger par ordre décroissant.
3. Retenir les deux arêtes correspondant aux deux plus fortes informations mutuelles observées.
4. Puis, parmi les suivantes, retenir celle qui ne crée pas de cycle avec les arêtes déjà retenues.
5. Répéter (4) jusqu'à retenir $(p-1)$ arêtes.
6. $\widehat{P}(\mathbf{X})$ est ensuite obtenue comme le produit des distributions des probabilités conditionnelles à partir d'une racine que l'on choisit arbitrairement.

$$\widehat{P}(\mathbf{X} = \mathbf{x}) = P_{obs}(X_1 = m_1) \prod_{j=2}^{p} P_{obs}(X_j = m_j | X_{i(j)} = m_i). \quad (6)$$

On obtient ainsi le réseau probabiliste avec arêtes qui approche au mieux la densité de probabilité observée P_{obs} d'un groupe donné.

2.2. Réseaux probabilistes avec prise en compte des liaisons d'ordre 2 : graphe avec triangles

Les réseaux probabilistes avec triangles sont des extensions des réseaux probabilistes avec arêtes prenant en compte des liaisons entre trois variables; ils sont tels que deux triangles ont au plus un sommet commun et ils ne forment pas de cycle. La généralisation de la démonstration de Chow et Liu (Annexe A) montre que la recherche de la distribution $\widehat{P}$, associée à un graphe avec triangles, la plus proche d'une distribution observée (au sens de la distance de Kullback-Leibler), consiste à chercher les t triangles pour lesquels la somme des informations mutuelles observées soit minimum.

Le nombre maximum de triangles t qui puissent être construits avec p variables est égal à la partie entière de $(p-1)/2$. (Si p est pair, il est donc possible de créer une variable artificielle X_{p+1} à valeur constante, de manière à ce que $t = p/2$ triangles soient construits.)

La mesure de l'information mutuelle entre trois variables s'écrit :

$$I(X_{j_1}, X_{j_2}, X_{j_3}) = \sum_{m_{j_1}=1}^{M_{j_1}} \sum_{m_{j_2}=1}^{M_{j_2}} \sum_{m_{j_3}=1}^{M_{j_3}} [\, P(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2}, X_{j_3} = m_{j_3}) \cdot \ln \frac{P(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2}, X_{j_3} = m_{j_3})}{P(X_{j_1} = m_{j_1}) P(X_{j_2} = m_{j_2}) P(X_{j_3} = m_{j_3})} \,],$$

et l'algorithme qui permet de construire le graphe et d'obtenir la densité de probabilité associée est le suivant :

1. Calculer, à partir de la distribution observée, les probabilités $P_{obs}(X_{j_1} = m_{j_1}, X_{j_2} = m_{j_2}, X_{j_3} = m_{j_3})$ pour tous les triplets de variables X_{j_1}, X_{j_2} et X_{j_3}.
2. Calculer les informations mutuelles observées $I_{obs}(X_{j_1}, X_{j_2}, X_{j_3})$ des p $(p-1)(p-2)/6$ triplets et les ranger par ordre décroissant.
3. Retenir le triangle correspondant à la plus forte information mutuelle observée.
4. Puis, parmi les suivants, retenir le premier triangle qui ne crée pas de cycle avec les triangles déjà retenus et qui ne possède pas non plus d'arête commune avec ces derniers.
5. Répéter (4) jusqu'à obtenir t triangles.
6. La densité de probabilité s'écrit alors comme le produit de t probabilités conditionnelles et de la probabilité de la racine choisie arbitrairement.

$$\widehat{P}(\mathbf{X} = \mathbf{x}) =$$

$$P_{obs}(X_1 = m_1) \prod_{v=1}^{t} P_{obs}(X_{1(v)} = m_{1(v)}, X_{2(v)} = m_{2(v)} | X_{0(v)} = m_{0(v)}), \quad (7)$$

où les sommets $X_{0(v)}$, $X_{1(v)}$ et $X_{2(v)}$ forment le $v^{ème}$ triangle dans lequel $X_{0(v)}$ est le parent de $X_{1(v)}$ et $X_{2(v)}$, et où $X_{0(1)} = X_1$.

On obtient ainsi le réseau probabiliste avec triangles qui approche au mieux la densité de probabilité P_{obs} d'un groupe donné. (Si p est pair, la variable artificielle X_{p+1} n'appartient qu'à un seul triangle v^* ($X_{p+1} = X_{1(v^*)}$) et la probabilité conditionnelle $P_{obs}(X_{1(v^*)} = cte, X_{2(v^*)} = m_{2(v^*)} | X_{0(v^*)} = m_{0(v^*)})$, s'écrit $P_{obs}(X_{2(v^*)} = m_{2(v^*)} | X_{0(v^*)} = m_{0(v^*)})$.)

3. Discrimination à l'aide des réseaux probabilistes

3.1. Règle de décision

Dans un contexte décisionnel, la discrimination consiste à construire une règle de décision, à partir de l'échantillon d'apprentissage. Une règle de décision δ est une

application de l'ensemble $\mathbf{X}$ des états dans l'ensemble G des groupes.

$$\mathbf{X} \xrightarrow{\delta} \{1, \cdots, k\}$$

$$\mathbf{x} \longrightarrow \delta(\mathbf{x}) = s.$$

La règle de décision obtenue à l'aide des réseaux probabilistes consiste à affecter l'état x au groupe pour lequel le coût entraîné par cette affectation est minimum.

$$\delta(\mathbf{x}) = \arg \min_{s} \left\{ \sum_{r=1}^{k} c_{s|r} \ \widehat{P}(G_r|\mathbf{X} = \mathbf{x}) \right\},$$

où $c_{s|r}$ est le coût de mauvais classement et $\widehat{P}(G_r|\mathbf{X} = \mathbf{x})$ est l'estimation de la probabilité *a posteriori* d'appartenance au groupe G_r. Elle est obtenue à l'aide de la formule de Bayes :

$$\widehat{P}(G_r|\mathbf{x}) = \frac{\pi_r \, \widehat{P}(\mathbf{x}|G_r)}{\sum_{s=1}^{k} \pi_s \, \widehat{P}(\mathbf{x}|G_s)} \qquad r = 1, \cdots, k,$$

où π_r est la probabilité *a priori* d'appartenance au groupe G_r.

Le risque associé à la règle de décision δ, noté $R^*(\delta)$ et permettant d'évaluer la qualité de la règle de décision, est égal à :

$$R^*(\delta) = \sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r} \, Q^*_{s|r}, \tag{8}$$

où $Q^*_{s|r}$ est la probabilité pour qu'une observation $\mathbf{x}$ du groupe G_r soit affectée au groupe G_s :

$$Q^*_{s|r} = P(\delta(\mathbf{x}) = s|\mathbf{x} \in G_r).$$

3.2.Évaluation d'une règle de décision

Une fois la règle de décision construite, il est important d'en évaluer la performance c'est-à-dire d'estimer le risque associé à la règle, l'erreur qu'elle produira lorsqu'elle sera appliquée à de nouvelles observations. Plusieurs procédures peuvent être utilisées pour estimer ce risque : les estimations par resubstitution et par échantillon test qui sont présentées dans les paragraphes suivants, ainsi que l'estimation par validation croisée et par bootstrap [Celeux et Nakache, 1994].

3.2.1. Estimation du risque par resubstitution

On note N_r le nombre d'observations de l'échantillon d'apprentissage appartenant au groupe G_r et $N_{s|r}$ le nombre d'observations du groupe G_r affectées au groupe G_s par la règle δ. L'estimation du risque par resubstitution s'obtient en estimant les probabilités $Q^*_{s|r}$ par les proportions observées $N_{s|r}/N_r$ d'observations du groupe G_r affectées par la règle δ au groupe G_s :

$$R^{ea}(\delta) = \sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r} \frac{N_{s|r}}{N_r}. \tag{9}$$

L'utilisation de cette valeur $R^{ea}(\delta)$ donne une idée optimiste de la performance de la règle de décision puisque le même échantillon sert à la fois pour construire la règle de décision et pour fournir l'estimation du risque qui lui est associé. Ce biais d'optimisme est d'autant plus fort que la taille de l'échantillon est petite et que la règle est complexe.

3.2.2. Estimation du risque par échantillon test

L'estimation du risque par la méthode de l'échantillon test est obtenue à partir d'un échantillon test indépendant de l'échantillon d'apprentissage. Soit δ la règle obtenue à l'aide de l'échantillon d'apprentissage. On note N^{et}_r le nombre d'observations de l'échantillon test appartenant au groupe G_r et $N^{et}_{s|r}$ le nombre d'observations de l'échantillon test appartenant au groupe G_r qui sont affectées au groupe G_s par la règle δ. L'estimation par échantillon test s'obtient en estimant les probabilités $Q^*_{s|r}$ par les proportions observées $N^{et}_{s|r}/N^{et}_r$ d'observations dans l'échantillon test appartenant au groupe G_r et affectées au groupe G_s par la règle δ :

$$R^{et}(\delta) = \sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r} \frac{N^{et}_{s|r}}{N^{et}_r}. \tag{10}$$

L'estimation ainsi obtenue est sans biais car l'échantillon test est indépendant de l'échantillon d'apprentissage. La variance de cette estimation est fournie dans l'Annexe B.

Remarque

Si les probabilités *a priori* π_r sont estimées par les fréquences des groupes, le risque estimé par resubstitution est alors égal à

$$R^{ea}(\delta) = \frac{1}{N} \sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}\, N_{s|r},$$

où N est le nombre d'observations de l'échantillon d'apprentissage. Pour l'estimation par échantillon test, il est licite d'estimer les probabilités *a priori* par N_r^{et}/N^{et} car l'échantillon test est indépendant de l'échantillon d'apprentissage. L'estimation du risque par échantillon test est égale à :

$$R^{et}(\delta) = \frac{1}{N^{et}} \sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}\, N_{s|r}^{et},$$

où N^{et} est le nombre d'observations de l'échantillon test.

Si de plus, les coûts de mauvais classement sont unitaires, l'estimation du risque par resubstitution est égale à :

$$R^{ea}(\delta) = \frac{1}{N} \sum_{r=1}^{k} \sum_{s=1, s\neq r}^{k} N_{s|r},$$

et l'estimation du risque par échantillon test est égale à :

$$R^{et}(\delta) = \frac{1}{N^{et}} \sum_{r=1}^{k} \sum_{s=1, s\neq r}^{k} N_{s|r}^{et}.$$

Dans ce cas l'estimation du risque est égale au pourcentage d'observations de l'échantillon mal classées par la règle δ.

3.3. Discrimination fondée sur l'adéquation des probabilités a posteriori

Adopter un point de vue purement décisionnel n'est pas toujours nécessaire, ni même recommandé; par exemple dans un problème de diagnostic médical, au lieu d'affecter un patient à une classe diagnostique, il peut être préférable de fournir uniquement les estimations des probabilités a posteriori d'appartenance aux différentes classes diagnostiques. Dans ce cas, il n'y a pas de règle de décision, la discrimination consiste à fournir pour chaque état $\mathbf{x} \in \mathbf{X}$ un vecteur $d(\mathbf{x})$

$$d(\mathbf{x}) = \{\widehat{P}(G_1|\mathbf{x}), \cdots, \widehat{P}(G_k|\mathbf{x})\},$$

décrivant les différentes probabilités *a posteriori* d'appartenance aux groupes. Breiman *et al.* [Breiman 1984] proposent de prendre l'écart quadratique moyen du vecteur $d(\mathbf{X})$ comme mesure de la qualité de la discrimination :

$$R^*(d) = E\left\{\sum_{r=1}^{k} \left(Z_r - \widehat{P}(G_r|\mathbf{X})\right)^2\right\}, \tag{11}$$

où Z_r est une variable aléatoire prenant la valeur 1 si il y a appartenance au groupe G_r et la valeur 0 sinon.

L'estimation de cette mesure par resubstitution est égale à

$$R^{ea}(d) = \sum_{r=1}^{k} \frac{\pi_r}{N_r} \sum_{i \in I_r} \sum_{s=1}^{k} \left(z_{is} - \widehat{P}(G_s|\mathbf{x}_i) \right)^2, \tag{12}$$

où I_r représente l'ensemble des indices des observations de l'échantillon d'apprentissage appartenant au groupe G_r, et où $z_{is} = 1$ si l'observation i appartient au groupe G_s et $z_{is} = 0$ sinon.

L'estimation de cette mesure par échantillon test est égale à

$$R^{et}(d) = \sum_{r=1}^{k} \frac{\pi_r}{N_r^{et}} \sum_{i \in I_r^{et}} \sum_{s=1}^{k} \left(z_{is} - \widehat{P}(G_s|\mathbf{x}_i) \right)^2, \tag{13}$$

où I_r^{et} est l'ensemble des observations des individus de l'échantillon test appartenant au groupe G_r. La variance de cette estimation est fournie dans l'Annexe B.

Remarque

Si les probabilités *a priori* π_r sont estimées par la fréquence des groupes, $R^{ea}(d)$ et $R^{et}(d)$ s'écrivent :

$$R^{ea}(d) = \frac{1}{N} \sum_{i \in I} \sum_{s=1}^{k} \left(z_{is} - \widehat{P}(G_s|\mathbf{x}_i) \right)^2,$$

$$R^{et}(d) \frac{1}{N^{et}} \sum_{i \in I^{et}} \sum_{s=1}^{k} \left(z_{is} - \widehat{P}(G_s|\mathbf{x}_i) \right)^2,$$

où I et I^{et} indicent les observations de l'échantillon d'apprentissage et de l'échantillon test. On parle alors de *score quadratique moyen* estimé par resubstitution ou par échantillon test.

4. Application

L'exemple utilisé pour illustrer cette méthode est constitué de 6000 observations simulées réparties en 1800 observations pour l'échantillon d'apprentissage et 4200 observations pour l'échantillon test. Trois groupes sont à discriminer; chaque observation est caractérisée par 9 variables binaires et 2 variables à 3 modalités. Les probabilités *a priori* sont estimées par les fréquences des groupes dans l'échantillon et les coûts de mauvais classement sont égaux.

Les probabilités *a posteriori* d'appartenance aux différents groupes ont été estimées à l'aide du système MAID [Nicolau 1994] pour les trois modèles envisagés :

indépendance conditionnelle, réseaux probabilistes avec arêtes et réseaux probabilistes avec triangles. Le système MAID a également permis de construire les règles de décision et d'estimer leurs risques (pourcentages d'observations mal classées) par resubstitution et échantillon test. Par contre, le calcul du score quadratique moyen (non implémenté pour l'instant dans le système MAID) a été effectué grâce au logiciel SAS.

Evaluation par le pourcentage d'observations mal classées

Le tableau 1 fournit le classement des observations des deux échantillons, apprentissage et test, par la règle issue du modèle avec indépendance conditionnelle. Le risque (pourcentage d'observations mal classées – MC –) estimé par resubstitution est égal à 0,619 et celui estimé par échantillon test est égal à 0,659, avec un écart-type de 7,32 10^{-3}.

TABLEAU 1

Classement des observations (modèle avec indépendance conditionnelle)

Affectation	*Origine* G_1	G_2	G_3
G_1	233	163	187
G_2	169	232	192
G_3	198	205	221

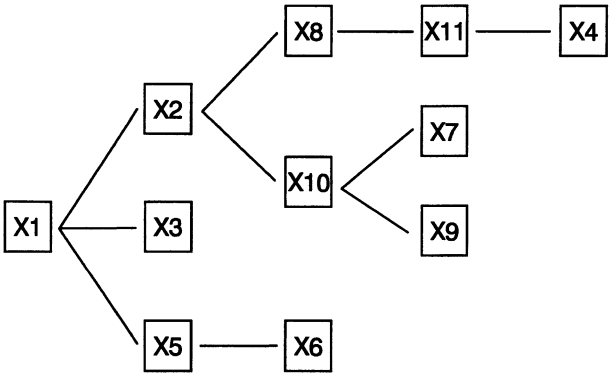
Échantillon d'apprentissage
MC : 61, 9%

Affectation	*Origine* G_1	G_2	G_3
G_1	485	396	492
G_2	432	528	487
G_3	483	476	421

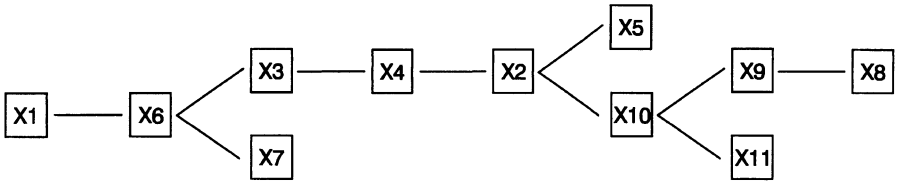
Échantillon test
MC : $65, 9 \pm 0, 7\%$

La figure 3 fournit, d'une part, la représentation graphique des réseaux probabilistes avec arêtes associés aux trois groupes G_1, G_2 et G_3, et d'autre part, les estimations des densités de probabilités correspondantes. Les probabilités *a posteriori* d'appartenance aux groupes sont ensuite obtenues grâce à la formule de Bayes. Chaque observation de l'échantillon d'apprentissage est alors affectée au groupe pour lequel la probabilité *a posteriori* est maximum. La même règle d'affectation est appliquée à l'échantillon test. Le tableau 2 fournit le classement des observations des deux échantillons – apprentissage et test – par la règle issue des réseaux probabilistes avec arêtes. Le risque (pourcentage d'observations mal classées) estimé par resubstitution est égal à 0,334 et celui estimé par échantillon test est égal à 0,378, avec un écart-type de 7,48 10^{-3}.

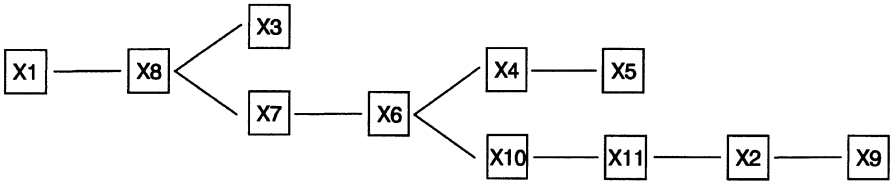
La figure 4 fournit, d'une part, la représentation graphique des réseaux probabilistes avec triangles associés aux trois groupes G_1, G_2 et G_3, et d'autre part les estimations des densités de probabilités correspondantes. Les probabilités *a posteriori* d'appartenance aux groupes sont ensuite obtenues grâce à la formule de Bayes. Chaque observation de l'échantillon d'apprentissage est alors affectée au groupe pour lequel la probabilité *a posteriori* est maximum. La même règle d'affectation est appliquée à l'échantillon test. Le tableau 3 fournit le classement des observations des deux échantillons, apprentissage et test, par la règle issue des réseaux probabilistes



$$P(\mathbf{X}|G_1) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_5|X_1)P(X_8|X_2)P(X_{10}|X_2)$$
$$P(X_{11}|X_8)P(X_4|X_{11})P(X_7|X_{10})P(X_9|X_{10})P(X_6|X_5)$$



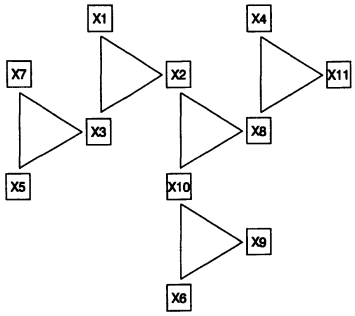
$$P(\mathbf{X}|G_2) = P(X_1)P(X_6|X_1)P(X_3|X_6)P(X_7|X_6)P(X_4|X_3)P(X_2|X_4)$$
$$P(X_5|X_2)P(X_{10}|X_2)P(X_9|X_{10})P(X_{11}|X_{10})P(X_8|X_9)$$



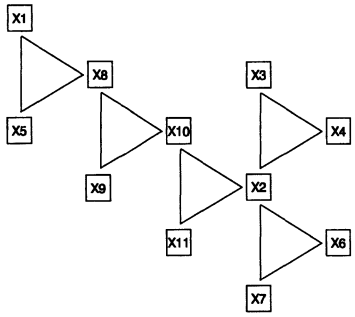
$$P(\mathbf{X}|G_3) = P(X_1)P(X_8|X_1)P(X_3|X_8)P(X_7|X_8)P(X_6|X_7)P(X_4|X_6)$$
$$P(X_{10}|X_6)P(X_5|X_4)P(X_{11}|X_{10})P(X_2|X_{11})P(X_9|X_2)$$

FIGURE 3

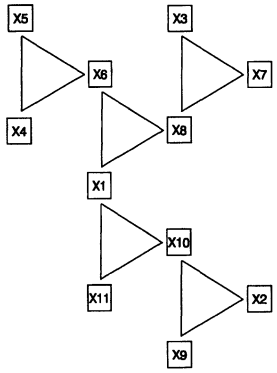
Réseaux probabilistes avec arêtes associés aux différents groupes



$$P(\mathbf{X}|G_1) = P(X_1)P(X_2, X_3|X_1)P(X_5, X_7|X_3)P(X_8, X_{10}|X_2)$$
$$P(X_4, X_{11}|X_8)P(X_6, X_9|X_{10})$$



$$P(\mathbf{X}|G_2) = P(X_1)P(X_5, X_8|X_1)P(X_9, X_{10}|X_8)P(X_2, X_{11}|X_{10})$$
$$P(X_3, X_4|X_2)P(X_6, X_7|X_2)$$



$$P(\mathbf{X}|G_3) = P(X_1)P(X_6, X_8|X_1)P(X_{10}, X_{11}|X_1)P(X_4, X_5|X_6)$$
$$P(X_3, X_7|X_8)P(X_2, X_9|X_{10})$$

FIGURE 4

Réseaux probabilistes avec triangles associés aux différents groupes

TABLEAU 2

Classement des observations (réseaux probabilistes avec arêtes)

Affectation	*Origine* G_1	G_2	G_3
G_1	370	108	87
G_2	137	401	85
G_3	93	91	428

Échantillon d'apprentissage
MC : 33, 4%

Affectation	*Origine* G_1	G_2	G_3
G_1	786	313	205
G_2	353	869	236
G_3	261	218	959

Échantillon test
MC : $37,8 \pm 0,7\%$

avec triangles. Le risque (pourcentage d'observations mal classées) estimé par resubstitution est égal à 0,322 et le risque estimé par échantillon test est égal à 0,380, avec un écart-type de 7,49 10^{-3}.

TABLEAU 3

Classement des observations (réseaux probabilistes avec triangles)

Affectation	*Origine* G_1	G_2	G_3
G_1	387	112	91
G_2	128	398	73
G_3	85	90	436

Échantillon d'apprentissage
MC : 32, 2%

Affectation	*Origine* G_1	G_2	G_3
G_1	828	341	213
G_2	337	816	228
G_3	235	243	959

Échantillon test
MC : $38,0 \pm 0,8\%$

Évaluation par le score quadratique moyen

L'utilisation du modèle avec indépendance conditionnelle fournit un score quadratique moyen calculé sur les 1800 observations de l'échantillon d'apprentissage (estimation par resubstitution) égal à 0,662. Son estimation par échantillon test est égale à 0,671, avec un écart-type de 0,02 10^{-3}. Les réseaux probabilistes avec arêtes conduisent à un score quadratique moyen estimé par resubstitution égal à 0,432. Son estimation par échantillon test est égale à 0,450, avec un écart-type de 0,10 10^{-3}. L'utilisation des réseaux probabilistes avec triangles fournit une estimation du score quadratique moyen estimé par resubstitution égale à 0,423. Son estimation par échantillon test est égale à 0,416, avec un écart-type de 0,10 10^{-3}.

Que le critère d'évaluation de la qualité de la discrimination soit le pourcentage d'observations mal classés ou le score quadratique moyen, on observe – comme on s'y attend – une amélioration du critère estimé par resubstitution quand on passe du modèle d'indépendance conditionnelle, au modèle avec arêtes et au modèle avec triangles. Il est clair que dans cet exemple, les réseaux probabilistes avec arêtes ou triangles fournissent une discrimination de bien meilleure qualité que celle obtenue

à l'aide du modèle d'indépendance conditionnelle. Par contre la prise en compte de liaisons entre trois variables n'améliore pas la qualité de la discrimination.

5. Discussion

L'intérêt des réseaux probabilistes réside dans leur facilité et rapidité d'obtention des estimations de densités de probabilité par groupe. Pour l'instant la technique proposée permet de retenir un modèle parmi le modèle d'indépendance conditionnelle, le modèle «arêtes» et le modèle «triangles», au moyen d'estimations des risques par échantillon test.

Il est possible d'étendre cette méthode de deux façons : la première consiste à évaluer la qualité de la discrimination à l'aide de techniques plus puissantes que l'estimation par échantillon test, comme la validation croisée ou le bootstrap. La seconde extension consiste à proposer d'autres modèles tenant compte du fait que les dernières arêtes ou les derniers triangles construits par l'algorithme peuvent apporter une information peu pertinente car ils dépendent trop des fluctuations d'échantillonnage.

Bibliographie

Agresti A. (1990), *Categorical data analysis*, New York : Wiley.

Breiman L., Freidman J.H., Olshen R.A., and Stone C.J. (1984), *Classification And Regression Trees*, Wadsworth International group, Belmont, California.

Celeux G., Nakache J.P. (1994), *Analyse discriminante sur variables qualitatives*, Polytechnica.

Chow C.K. and Liu C.N. (1968), "Approximating discrete probability distributions with dependence trees", *IEEE Trans. on Info. Theory*, **IT-14**, 462–467.

Kullback S. and Leibler R.A. (1951), "Information and sufficiency", *Ann. Math. Statist.*, **22**, 79–86.

Nicolau J. (1994), *Méthodes d'aide à la décision. Le système MAID : aspects statistiques et informatiques*, Thèse de 3ème cycle, Université Paris V.

Pearl J. (1988), *Probabilistic reasoning in intelligent systems : networks of plausible inference*, Morgan Kaufmann publishers, San Mateo, California.

Annexe A

La recherche de la distribution $\widehat{P}$ (associée à un graphe G avec triangles), la plus proche d'une distribution observée P_{obs} se fait en deux étapes. Dans un premier temps, on suppose que la structure du graphe G est fixée et on cherche quelles probabilités $P^G(X_{1(v)} = m_{1(v)}, X_{2(v)} = m_{2(v)} | X_{0(v)} = m_{0(v)})$ pour $v = 1, \cdots, t$ rendent la distribution P^G la plus proche possible de la distribution P_{obs} au sens de la distance de Kullback-Leibler. La distribution P^G est appelée *projection* de P_{obs} sur le graphe

G. Dans un deuxième temps, on fait varier G parmi tous les graphes possibles et, parmi toutes les projections de P_{obs} sur ces graphes, on cherche celle qui est la plus proche de P_{obs}.

Première étape

Considérons deux distributions P et P^G, P^G étant associée à un graphe avec triangles. P^G s'écrit :

$$P^G(\mathbf{X}) = P^G(X_1) \prod_{v=1}^{t} P^G(X_{1(v)}, X_{2(v)}|X_{0(v)}).$$

Le graphe G est défini par t triangles, le $v^{ème}$ triangle étant composé des 3 sommets $X_{1(v)}$, $X_{2(v)}$ et $X_{0(v)}$. La distance de Kullback-Leibler entre P et P^G s'écrit :

$$D(P, P^G) = -\sum_{\mathbf{x}} P(\mathbf{X}) \ln P^G(\mathbf{X}) + \sum_{\mathbf{x}} P(\mathbf{X}) \ln P(\mathbf{X}). \tag{14}$$

En posant $H(\mathbf{X}) = \sum_{\mathbf{x}} P(\mathbf{X}) \ln P(\mathbf{X})$ on obtient :

$$D(P, P^G)$$

$$= -\sum_{m_1,\cdots,m_p} P(X_1, \cdots, X_p) \ln P^G(X_1, \cdots, X_p) + H(\mathbf{X}) \tag{15}$$

$$= -\sum_{m_1,\cdots,m_p} P(X_1, \cdots, X_p) \left[\ln P^G(X_1) + \sum_{v=1}^{t} \ln P^G(X_{1(v)}, X_{2(v)}|X_{0(v)}) \right] + H(\mathbf{X}) \tag{16}$$

$$= -\sum_{m_1} \left[\sum_{m_2,\cdots,m_p} P(X_2, \cdots, X_p|X_1) \right] P(X_1) \ln P^G(X_1) - \sum_{v=1}^{t} \sum_{m_{1(v)}} \sum_{m_{2(v)}} \sum_{m_{0(v)}} P(X_{1(v)}, X_{2(v)}, X_{0(v)}) \ln P^G(X_{1(v)}, X_{2(v)}|X_{0(v)}) + H(\mathbf{X}) \tag{17}$$

$$= -\sum_{m_1} P(X_1) \ln P^G(X_1) - \sum_{v=1}^{t} \sum_{m_{0(v)}} P(X_{0(v)}) \sum_{m_{1(v)}} \sum_{m_{2(v)}} P(X_{1(v)}, X_{2(v)}|X_{0(v)}) \ln P^G(X_{1(v)}, X_{2(v)}|X_{0(v)}) + H(\mathbf{X}) \tag{18}$$

On sait que si la distribution P est fixe, l'expression :

$$\sum_{m} P(X = m) \ln P'(X = m),$$

est maximum pour $P'(X = m) = P(X = m)$, et que l'expression :

$$\sum_{m_1} \sum_{m_2} P(X_1 = m_1, X_2 = m_2) \ln P'(X_1 = m_1, X_2 = m_2),$$

est maximum pour $P'(X_1 = m_1, X_2 = m_2) = P(X_1 = m_1, X_2 = m_2)$. Donc, le graphe G étant fixé, la distribution P^G qui minimise la distance de Kullback-Leibler est telle que :

$$P^G(X_1) = P(X_1) \quad \text{et}$$
$$P^G(X_{1(v)}, X_{2(v)} | X_{0(v)}) = P(X_{1(v)}, X_{2(v)} | X_{0(v)}) \quad \text{pour } v = 1, \cdots, t.$$

La distribution P_G qui minimise $D(P, P^G)$ est égale à :

$$P^G(\mathbf{X}) = P(X_1) \prod_{v=1}^{t} P(X_{1(v)}, X_{2(v)} | X_{0(v)}). \tag{19}$$

Deuxième étape

On remplace $P^G(\mathbf{X})$ par son expression donnée en (19) dans l'expression (15) :

$$\begin{aligned} & D(P, P^G) \\ & = - \sum_{m_1} P(X_1) \ln P(X_1) \\ & \quad - \sum_{v=1}^{t} \sum_{m_{1(v)}} \sum_{m_{2(v)}} \sum_{m_{0(v)}} P(X_{1(v)}, X_{2(v)}, X_{0(v)}) \ln P(X_{1(v)}, X_{2(v)} | X_{0(v)}) \\ & \quad + H(\mathbf{X}) \qquad (20) \\ & = - \sum_{m_1} P(X_1) \ln P(X_1) \\ & \quad - \sum_{v=1}^{t} \sum_{m_{1(v)}} \sum_{m_{2(v)}} \sum_{m_{0(v)}} P(X_{1(v)}, X_{2(v)}, X_{0(v)}) \\ & \qquad \left[\ln \frac{P(X_{1(v)}, X_{2(v)}, X_{0(v)})}{P(X_{1(v)}) P(X_{2(v)}) P(X_{0(v)})} + \ln P(X_{1(v)}) + \ln P(X_{2(v)}) \right] \\ & \quad + H(\mathbf{X}) \qquad (21) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{m_1} P(X_1)\ln P(X_1) \\
&\quad -\sum_{v=1}^{t} I(X_{1(v)}, X_{2(v)}, X_{0(v)}) \\
&\quad -\sum_{v=1}^{t}\sum_{m_{1(v)}} P(X_{1(v)})\ln P(X_{1(v)}) - \sum_{v=1}^{t}\sum_{m_{2(v)}} P(X_{2(v)})\ln P(X_{2(v)}) \\
&\quad + H(\mathbf{X}) \qquad (22)
\end{aligned}
$$

où $I(X_{1(v)}, X_{2(v)}, X_{0(v)}) =$

$$
\sum_{m_{1(v)}}\sum_{m_{2(v)}}\sum_{m_{0(v)}} P(X_{1(v)}, X_{2(v)}, X_{0(v)})\ln \frac{P(X_{1(v)}, X_{2(v)}, X_{0(v)})}{P(X_{1(v)})P(X_{2(v)})P(X_{0(v)})}
$$

$$
D(P, P^G) = -\sum_{v=1}^{t} I(X_{1(v)}, X_{2(v)}, X_{0(v)}) - \sum_{j=1}^{p} P(X_j)\ln P(X_j) + H(\mathbf{X}) \quad (23)
$$

Le deuxième et le troisième terme étant indépendants du graphe G, la distribution P^G qui rend minimum la distance de Kullback-Leibler est celle qui maximise la somme des informations mutuelles :

$$
\sum_{v=1}^{t} I(X_{1(v)}, X_{2(v)}, X_{0(v)}).
$$

Annexe B

Discrimination décisionnelle

L'estimation du risque associé à une règle δ par la méthode de l'échantillon test est égale à :

$$
R^{et}(\delta) = \sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r} \frac{N_{s|r}^{et}}{N_r^{et}}.
$$

- Dans le cas où les probabilités *a priori* sont estimées par les fréquences observées des groupes dans l'échantillon test, on note $p_{s|r}^{et_1} = N_{s|r}^{et}/N^{et}$ la proportion dans l'ensemble de l'échantillon test des observations du groupe G_r affectées au groupe G_s. L'estimation du risque peut également s'écrire de la

façon suivante :

$$R^{et}(\delta) = \sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}\, p_{s|r}^{et_1}.$$

La variance (estimée) de l'estimation du risque s'écrit :

$$\hat{var}\left(R^{et}(\delta)\right) = \hat{var}\left(\sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}\, p_{s|r}^{et_1}\right).$$

En indiçant par t les différentes combinaisons des indices r et s, on écrit :

$$\hat{var}\left(R^{et}(\delta)\right) = \hat{var}\left(\sum_{t} c_t\, p_t^{et_1}\right),$$

et la variance s'écrit :

$$\hat{var}\left(R^{et}(\delta)\right) = \sum_{t} c_t^2 \frac{p_t^{et_1}(1 - p_t^{et_1})}{N^{et}} - 2 \sum_{t \neq t'} c_t c_{t'} \frac{p_t^{et_1} p_{t'}^{et_1}}{N^{et}},$$

$$\hat{var}\left(R^{et}(\delta)\right) = \frac{1}{N^{et}} \left[\sum_{t} c_t^2\, p_t^{et_1} - \left(\sum_{t} c_t\, p_t^{et_1}\right)^2\right],$$

c'est-à-dire :

$$\hat{var}\left(R^{et}(\delta)\right) = \frac{1}{N^{et}} \left[\sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}^2\, p_{s|r}^{et_1} - \left(\sum_{r=1}^{k} \sum_{s=1}^{k} c_{s|r}\, p_{s|r}^{et_1}\right)^2\right].$$

- Dans le cas où les probabilités *a priori* π_r sont données par l'utilisateur, on note $p_{s|r}^{et_2}$ la proportion des observations dans l'échantillon test appartenant au groupe G_r affectées au groupe G_s, $p_{s|r}^{et_2} = N_{s|r}^{et} / N_r^{et}$. L'estimation du risque s'écrit :

$$R^{et}(\delta) = \sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r}\, p_{s|r}^{et_2},$$

et la variance (estimée) de l'estimation du risque :

$$\hat{var}\left(R^{et}(\delta)\right) = \hat{var}\left(\sum_{r=1}^{k} \pi_r \sum_{s=1}^{k} c_{s|r}\, p_{s|r}^{et_2}\right).$$

En utilisant une démonstration semblable à celle utilisée ci-dessus, on montre que :

$$\hat{var}\left(R^{et}(\delta)\right) = \sum_{r=1}^{k} \frac{\pi_r^2}{N_r^{et}} \left[\sum_{s=1}^{k} c_{s|r}^2 \, p_{s|r}^{et2} - \left(\sum_{s=1}^{k} c_{s|r} \, p_{s|r}^{et2} \right)^2 \right].$$

Discrimination non décisionnelle, fondée sur l'adéquation des probabilités a posteriori

L'estimation du risque associé à la discrimination d, par la méthode de l'échantillon test, est égale à :

$$R^{et}(d) = \sum_{r=1}^{k} \pi_r R_r^{et}(d),$$

où $R_r^{et}(d) = \dfrac{1}{N_r^{et}} \sum_{i \in I_r^{et}} L_i$ et $L_i = \sum_{s=1}^{k} \left(z_{is} - \widehat{P}(G_s|\mathbf{x}_i) \right)^2$ dans lequel z_{is} prend la valeur 1 si l'individu i appartient au groupe G_s et 0 sinon.

- Si les probabilités *a priori* sont estimées par les fréquences observées des groupes dans l'échantillon test, l'estimation du risque s'écrit :

$$R^{et}(d) = \frac{1}{N^{et}} \sum_{i \in I^{et}} L_i,$$

et la variance (estimée) de l'estimation du risque est égale à :

$$\hat{var}\left(R^{et}(d)\right) = \frac{1}{N^{et}} \left(\frac{\sum_{i \in I^{et}} L_i^2}{N^{et}} - R^{et}(d)^2 \right).$$

- Si les probabilités *a priori* sont données, l'estimation du risque s'écrit :

$$R^{et}(d) = \sum_{r=1}^{k} \pi_r R_r^{et}.$$

La variance de R^{et} est égale à :

$$\hat{var}\left(R^{et}(d)\right) = \sum_{r=1}^{k} \pi_r^2 \, \hat{var}\left(R_r^{et}(d)\right).$$

On obtient une estimation de la variance en remplaçant $\hat{var}\left(R_r^{et}(d)\right)$ par son estimation :

$$\hat{var}\left(R_r^{et}(d)\right) = \frac{1}{N_r^{et}} \left(\frac{\sum_{i \in I_r^{et}} L_i^2}{N_r^{et}} - R_r^{et}(d)^2 \right).$$