Offre de stage de Master 2
Étude et implémentation d'un modèle de classification de variables en fonction de leur comportement groupant

Vincent Vandewalle

25 janvier 2017

Durée : 4 à 5 mois.
Profils du candidat : Master 2 ou équivalent disposant de compétences en classification supervisée, non supervisée, optimisation et programmation R.
Contact : Vincent Vandewalle (vincent.vandewalle@inria.fr), maître de conférence à l'Université Lille 2 et membre de l'équipe Modal d'Inria Lille Nord-Europe.
Lieu du stage : Équipe Modal de l'Inria Lille Nord-Europe, 40 Avenue du Halley, 59650 Villeneuve-d'Ascq, France.
Rémunération : Environ 500 euros par mois.

1 Présentation du stage

En classification non supervisée on suppose souvent l'existence d'une seule variable de classe sensée expliquer l'hétérogénéité de l'ensemble des données. Cependant quand les variables proviennent de diverses sources cette hypothèse est souvent irréaliste.

Dans ce stage on propose d'étudier un modèle de classification supposant l'existence de plusieurs variables de classe, comme dans [6], chacune d'entre-elle expliquant l'hétérogénéité d'une partie des variables observées. Cette approche permet alors de regrouper ensembles des variables expliquées par la même variable de classe. Contrairement à l'approche habituelle, ce regroupement des variables n'est pas réalisé par rapport à une distance entre variables, mais par rapport à un même comportement groupant vis à vis des individus.

Le stage proposé consistera dans un premier temps à implementer un algorithme de type EM [2] pour estimer les paramètres du modèle proposé dans un cadre similaire à celui présenté dans [1]. Dans un second temps, il consistera à proposer une approche de choix de modèle permettant de choisir le nombre de variables de classes, ainsi que le nombre de classes pour chacune d'entre-elles. Ce choix pourra être réalisé à partir du critère BIC [5, 3], ou à partir d'une extension du critère MICL proposé dans [4] dans le contexte de la selection de variables. Les méthodes développées seront testées sur données simulées et réelles. Enfin, en fonction de son avancement le stage pourra donner lieu à la création d'un package R.

2 Détails sur le modèle proposé

2.1 Modèle

Soit $\boldsymbol{X} = (X_1, \ldots, X_d)$ un vecteur aléatoire où X_1, $\ldots$, X_d peuvent être de nature variée (quantitatives continues, quantitatives discontinues, qualitatives, ...). On suppose que les variables de $\boldsymbol{X}$ peuvent être séparées en B blocs indépendants les uns des autres. On notera par $\mathcal{B}_b$ l'ensemble des numéros de variables associées au bloc b et par $\boldsymbol{X}^b$ les variables du bloc b : $\boldsymbol{X}^b = \{X_j | j \in \mathcal{B}_b\}$. On suppose que dans chaque bloc b les variables sont indépendantes les unes des autres sachant une variable de classe Z^b à valeur dans $\{1, \ldots, K_b\}$. En pratique les variables de classe $Z^1, \ldots, Z^B$ ne sont pas observées.

La densité de probabilité de $\boldsymbol{x}$, une réalisation de $\boldsymbol{X}$, s'écrit :

$$p(\boldsymbol{x}) = \prod_{b=1}^{B} p(\boldsymbol{x}^b) = \prod_{b=1}^{B} \left(\sum_{k_b=1}^{K_b} p(\boldsymbol{x}^b | Z^b = k_b) p(Z^b = k_b) \right) = \prod_{b=1}^{B} \left(\sum_{k_b=1}^{K_b} \prod_{j \in \mathcal{B}_b} p(x_j | Z^b = k_b) p(Z^b = k_b) \right).$$

Afin de spécifier totalement le modèle on définit un modèle paramétrique de paramètres $\boldsymbol{\alpha}^b_{jk_b}$ sur la distribution de $X_j | Z^b = k_b$ pour $j \in \mathcal{B}_b$; si la variable X_j est quantitative le modèle considéré sera une distribution normale, si la variable est qualitative on considérera un modèle multinomial. Les quantités $p(Z^b = k_b)$ notées $\pi^b_{k_b}$ seront considérées comme des paramètres du modèle.

La figure 1 illustre le modèle proposé dans le cas de $d = 4$ variables quantitatives séparées en $B = 2$ blocs et chacun constitués de deux classes ($K_1 = K_2 = 2$: noir ou rouge, et croix ou triangle) ; on voit que les variables X_1 et X_2 permettent de bien séparer les points rouges des points noirs ($\mathcal{B}_1 = \{1, 2\}$), tandis que les variables X_3 et X_4 permettent de bien séparer les croix des triangles ($\mathcal{B}_2 = \{3, 4\}$).

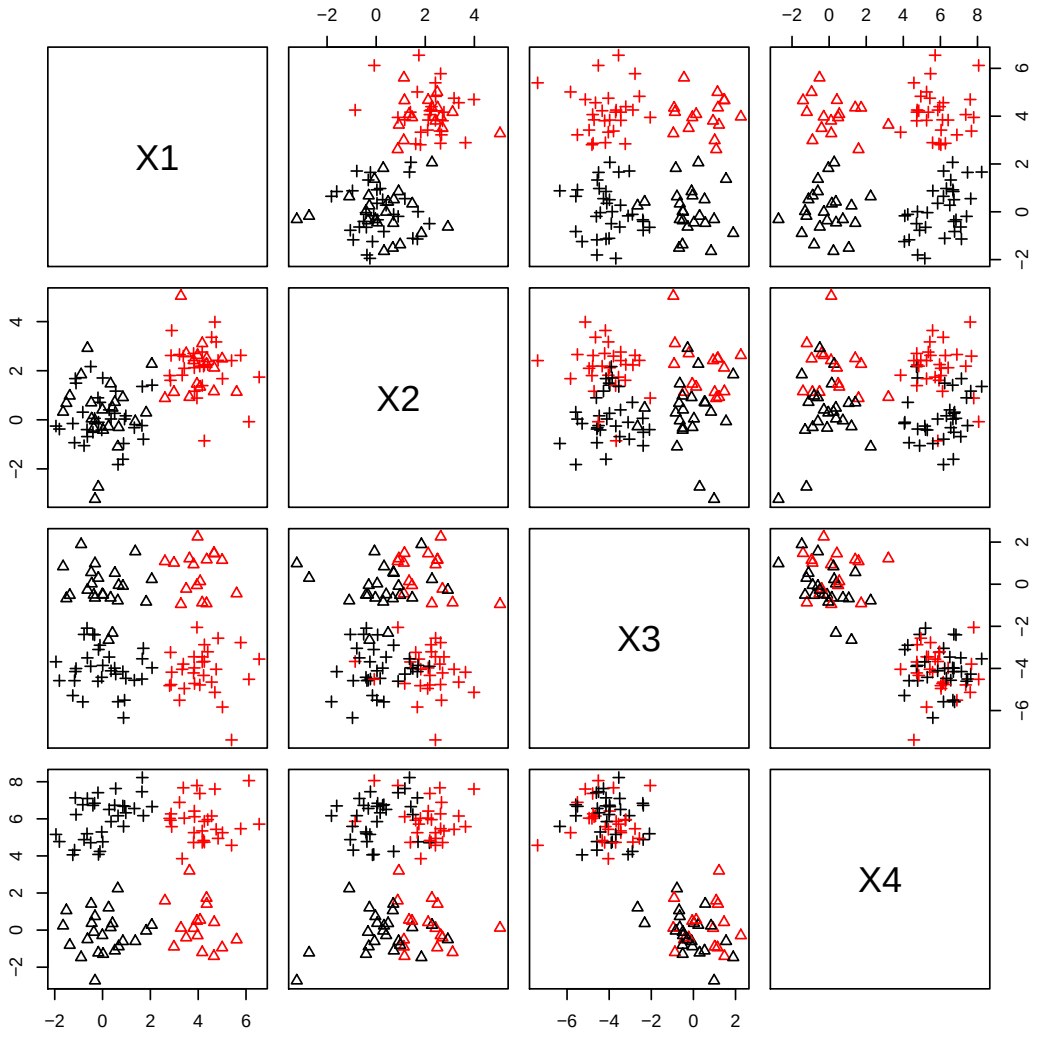

FIGURE 1 – Nuage de points 2 à 2 colorés en fonction de la valeur de la première variable classifiante et de formes différentes selon la valeur de la seconde variable classifiante.

2.2 Inférence

On dispose d'un n-échantillon i.d.d. issu de la distribution de $\boldsymbol{X}$. On souhaite estimer les paramètres du modèle. On travaille dans cette partie à nombre B de blocs fixé, à composition des blocs $\mathcal{B}_1, \ldots, \mathcal{B}_B$ fixée, et à nombre de classes dans chaque bloc $K_1, \ldots, K_B$ fixé. Dans ce cas le problème d'inférence se résume à estimer les $\boldsymbol{\alpha}^b_{jk_b}$ et les $\pi^b_{k_b}$. Ceux-ci peuvent être estimés par maximum de vraisemblance à l'aide de l'algorithme EM [2, 1].

2.3 Choix de modèle

En pratique B, $\mathcal{B}_1, \ldots, \mathcal{B}_B$, $K_1, \ldots, K_B$ sont inconnus et doivent être déterminés à partir des données. Le choix de ces paramètres peut être vu comme un problème de choix de modèle, qui est ici un réel challenge puisque le nombre possible de modèles est excessivement grand. Ce choix pourra être réalisé à partir d'un critère de choix de modèle de type BIC [5, 3] ou à partir d'une extension du critère MICL proposé dans [4] dans le contexte de la selection de variables.

Références

[1] Christophe Biernacki. Pourquoi les modèles de mélange pour la classification. *Revue de MODULAD*, 40 :1–22, 2009.

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[3] Chris Fraley and Adrian E Raftery. How many clusters ? which clustering method ? answers via model-based cluster analysis. *The computer journal*, 41(8) :578–588, 1998.

[4] Matthieu Marbac and Mohammed Sedki. Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, pages 1–15.

[5] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[6] Vincent Vandewalle. Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis. CMStatistics 2016, December 2016.