Scoring Appliqué à la Détection du Risque – Chapitre 2 – Méthodes probabilistes

Contexte et objectif

Les données. Elles sont constituées de n couples $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)$. Pour $i \in \{1, \ldots, n\}$: $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})' \in \mathcal{X}$ est la valeur des prédicteurs pour l'individu i et $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,G}) \in \{0,1\}^G$ indique sa classe : $\boldsymbol{z}_{i,g} = 1 \Leftrightarrow \boldsymbol{x}_i$ appartient à la classe g. L'ensemble $\mathcal{X}$ dépend de la nature des données prédictives (quantitatives continues/discrètes, qualitatives, mixtes etc).

L'objectif. Estimer la classe de $\boldsymbol{x}_{n+1} \in \mathcal{X}$.

La méthode. Définir G fonctions de score $\mathfrak{s}_g : \mathcal{X} \to \mathbb{R}$ permettant de décider : $\hat{z}_{n+1,g} = 1 \Leftrightarrow \mathfrak{s}_g(\boldsymbol{x}_{n+1}) = \max\{\mathfrak{s}_j(\boldsymbol{x}_{n+1}); j = 1, \ldots, G\}$.

Analyse discriminante probabiliste

L'analyse discriminante probabiliste s'inscrit dans une approche générative : elle propose un modèle paramétrique expliquant la génèse des observations $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}$.

Le modèle. On suppose que :
∘ les couples $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)$ sont des réalisations indépendantes de $(\boldsymbol{x}, \boldsymbol{z})$ où :
$\boldsymbol{x} = (x_1, \ldots, x_d)' \in \mathcal{X}$ est un vecteur aléatoire de prédicteurs,
$\boldsymbol{z} = (z_1, \ldots, z_G)' \in \{0,1\}^G$ est un vecteur aléatoire donnant la classe de $\boldsymbol{x}$ ($\sum_{g=1}^G z_g = 1$),
∘ $\boldsymbol{z}$ est distribué selon une loi multinomiale d'ordre G à un essai de paramètre : $(\pi_1, \ldots, \pi_G)$; $0 < \pi_g$ et $\sum_{g=1}^G \pi_g = 1$,
∘ le vecteur conditionnel $(\boldsymbol{x}|z_g = 1)$ est distribué selon une loi à support dans $\mathcal{X}$, de fonction de probabilté : $f_g(\bullet; \boldsymbol{\theta}_g)$.

Les fonctions $f_g(\bullet; \boldsymbol{\theta}_g)$ appartiennent généralement à une même famille de lois qui dépend de la nature des données.

La méthode.
∘ estimation du paramètre $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ par maximum de vraisemblance :

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}; \{\boldsymbol{x}_i, \boldsymbol{z}_i\}) = \log \left\{ \prod_{i=1}^n \prod_{g=1}^G [\pi_g f_g(\boldsymbol{x}_i; \boldsymbol{\theta}_g)]^{z_{i,g}} \right\} \right\}$$

∘ estimation des probabilités conditionnelles d'appartenance aux classes :

$$t_g(\boldsymbol{x}_{n+1}) = E(z_{n+1,g}|\boldsymbol{x}_{n+1}) = \hat{\pi}_g f_g(\boldsymbol{x}_{n+1}; \hat{\boldsymbol{\theta}}_g) / \sum_{j=1}^G \hat{\pi}_j f_j(\boldsymbol{x}_{n+1}; \hat{\boldsymbol{\theta}}_j)$$

∘ prédiction : en supposant les coûts de mauvais classement homogènes, la règle de Bayes visant à minimiser le coût moyen d'une erreur de classement revient à affecter $\boldsymbol{x}_{n+1}$ à la classe correspondant à la plus grande valeur des probabilités conditionnelles : $t_g(\boldsymbol{x}_{n+1})$; $g = 1, \ldots, G$ (règle du MAP[a]).

Choix d'un modèle/d'un classifieur.
Les fonctions de probabilité associées aux composantes : $f_g(\bullet, \boldsymbol{\theta}_g)$ déterminent entièrement le modèle/classifieur. Plusieurs modèles peuvent être en concurrence ; on les départage alors en s'appuyant sur (au choix) :

(i) un critère de vraisemblance comme : $AIC = \ell(\hat{\boldsymbol{\theta}}; \{\boldsymbol{x}_i, \boldsymbol{z}_i\}) - \eta$, $BIC = \ell(\hat{\boldsymbol{\theta}}; \{\boldsymbol{x}_i, \boldsymbol{z}_i\}) - (\eta/2) \times \log n$, où η est la dimension du paramètre

(ii) l'erreur de classement : $\epsilon = \sum_{g=1}^G \hat{\pi}_g \int_{\{\boldsymbol{x} \in \mathcal{X}; \hat{z}_g \neq 1\}} f_g(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_g) \mathrm{d}\boldsymbol{x}$ estimée par : resubstitution, validation croisée, ou bootstrap.

Les critères de vraisemblance (i) favorisent l'estimation de densité tandis que (ii) donne une importance prépondérante à la classification.

Remarque. En analyse discriminante, les fonctions de score sont les probabilités conditionnelles : $\mathfrak{s}_g(\boldsymbol{x}) = E(z_g|\boldsymbol{x})$.

L'analyse discriminante en pratique.

L'affectation de $\boldsymbol{x}_{n+1}$ à une classe résulte du modèle inféré. Or le modèle dépend du contexte : on prendra pour fonction $f_g(\bullet; \boldsymbol{\theta}_g)$ la densité d'une loi normale par exemple lorsque les observations sont continues, la fonction de probabilité d'une loi multinomiale lorsque les observations sont qualitatives, etc. Il existe des logiciels permettant d'inférer un modèle dans la plupart des cas classiques : données continues, discrètes, nominales, mixtes, etc. Table 1 cites quelques uns de ces logiciels.

A. Lourme, Faculté d'économie, gestion & AES, Université de Bordeaux http://alexandrelourme.free.fr

[a]Maximum A Posteriori

type de données	*contexte*	*lois conditionnelles*[b]	*package*	*références*	*fonctions utiles*
quantitatives continues	données standard	normales	`Rmixmod`	[7, 8]	`mixmodLearn`, `mixmodPredict`
	données bruitées	Student	`teigen`	[2, 1]	`teigen`
	données de grande dimension	normales	`HDclassif`	[3, 4]	`hdda`
			`pgmm`	[9]	`pgmmEM`
discrètes	données de comptage	Poisson	`MixtComp`	[5]	`Learn`, `Predict`
qualitatives	données nominales	multinomiales	`Rmixmod`	[7, 8]	`mixmodLearn`, `mixmodPredict`
mixtes	données quantitatives, qualitatives, ...	mixtes	`MixtComp`	[5]	`Learn`, `Predict`

Table 1: *Les logiciels dédiés à l'inférence d'un modèle d'analyse discriminante selon le contexte (données continues/qualitatives/mixtes/etc.)*

Régression logistique

La régression logistique s'inscrit dans une approche prédictive : elle propose un modèle semi paramétrique sur $\boldsymbol{z}$ conditionnellement à $\boldsymbol{x}$.

Le modèle. On se place dans le cas de données continues ($\mathcal{X} = \mathbb{R}^d$) et on suppose la partition binaire[c] ($G = 2$). La régression logistique modélise seulement la probabilité conditionnelle $t_1(\boldsymbol{x})$ d'appartenir à la classe 1 : $t_1(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}'\boldsymbol{x})/(1 + \exp(\alpha + \boldsymbol{\beta}'\boldsymbol{x}))$ où $\alpha \in \mathbb{R}$ et $\boldsymbol{\beta} \in \mathbb{R}^d$. Ainsi $\log\{t_1(\boldsymbol{x})/(1 - t_1(\boldsymbol{x}))\} = \alpha + \boldsymbol{\beta}'\boldsymbol{x}$ et la frontière séparant les deux classes est un hyperplan.

La méthode.

∘ estimation du paramètre $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ par maximum de vraisemblance conditionnelle :

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} z_{i,1}(\alpha + \boldsymbol{\beta}'\boldsymbol{x}_i) - \log(1 + \exp(\alpha + \boldsymbol{\beta}'\boldsymbol{x}_i)) \right\}. \quad (1)$$

Remarque. Les équations de vraisemblance permettant de déterminer (1) ne sont pas linéaires ; un algorithme de Newton-Raphson permet cependant de déterminer $\hat{\boldsymbol{\theta}}$. En pratique, sous R, l'estimation de $\boldsymbol{\theta}$ peut être réalisée grâce à la fonction `glm`.

∘ prédiction :

- soit la fonction de score est : $\mathfrak{s}_1(\boldsymbol{x}) = \hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}$ et $\boldsymbol{x}$ est affecté à la classe 1 si $\mathfrak{s}_1(\boldsymbol{x}) > 0$, à la classe 2 dans le cas contraire
- soit la fonction de score est : $\mathfrak{s}_1(\boldsymbol{x}) = t_1(\boldsymbol{x}; \hat{\alpha}, \hat{\boldsymbol{\beta}}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x})/(1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}))$ et $\boldsymbol{x}$ est affecté à la classe 1 si $\mathfrak{s}_1(\boldsymbol{x}) > 0,5$, à la classe 2 dans le cas contraire.

References

[1] Jeffrey L Andrews and Paul D Mcnicholas. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, 22(5):1021–1029, 2012.

[2] Jeffrey L Andrews, Paul D McNicholas, and Maintainer Jeffrey L Andrews. Package 'teigen'. 2015.

[3] Laurent Bergé, Charles Bouveyron, and Stéphane Girard. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

[4] Charles BOUVEYRON. Hdclassif: an r package for model-based classification of high-dimensional data. *Talk, Université Paris*, 1, 2012.

[5] MODAL Team INRIA Lille Nord Europe. MixtComp MOdels for Data Analysis and Learning. `https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp`, 2015.

[6] Gérard Govaert. *Data analysis*. John Wiley & Sons, 2013.

[7] Rémi Lebret, Serge Iovleff, Florent Langrognet, Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. *Journal of Statistical Software*, pages In–press, 2015.

[8] Remi Lebret, Serge Iovleff, Florent Langrognet, Maintainer Remi Lebret, and LinkingTo Rcpp. Package 'rmixmod'. 2012.

[9] PD McNicholas, KR Jampani, AF McDaid, TB Murphy, and L Banks. pgmm: Parsimonious gaussian mixture models. *R package version*, 1:24–47, 2011.

[c] il est possible de généraliser la régression logistique à plus de deux classes (voir [6], p.194)