Apprentissage Machine / Statistique

Analyse discriminante décisionnelle

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Notations

- p variables quantitatives explicatives X^j,
- une variable qualitative T (m modalités)
- un échantillon Ω de taille n.
- $\{\mathbf{g}_\ell; \ell = 1, \ldots, m\}$ désignent les barycentres des classes
- $\overline{\mathbf{x}}$ le barycentre global

Objectif

- affecter un nouvel individu $\mathbf{x} = [x_1, \ldots, x_p]'$
- dans une classe $\mathcal{T}_\ell$ de T
- Définir des *règles d'affectation*

Règle élémentaire avec m classes

Affecter l'individu x à la modalité de T minimisant :

$$d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell), \ell = 1, \ldots, m.$$

- Métrique de Mahalanobis
- $d^2_{\mathbf{S}_r^{-1}}(\mathbf{x}, \mathbf{g}_\ell) = \|\mathbf{x} - \mathbf{g}_\ell\|^2_{\mathbf{S}_r^{-1}} = (\mathbf{x} - \mathbf{g}_\ell)'\mathbf{S}_r^{-1}(\mathbf{x} - \mathbf{g}_\ell)$
- Ceci revient à maximiser

$$\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{x} - \frac{1}{2}\mathbf{g}_\ell'\mathbf{S}_r^{-1}\mathbf{g}_\ell.$$

- Règle linéaire en $\mathbf{x}$.

Règle élémentaire avec 2 classes

- Un seul axe discriminant Δ passant par $\mathbf{g}_1$ et $\mathbf{g}_2$.
- Règle de Fisher : $\mathbf{x}$ affecté à T_1 si

$$\mathbf{g}_1' \mathbf{S}_r^{-1} \mathbf{x} - \frac{1}{2} \mathbf{g}_1' \mathbf{S}_r^{-1} \mathbf{g}_1 > \mathbf{g}_2' \mathbf{S}_r^{-1} \mathbf{x} - \frac{1}{2} \mathbf{g}_2' \mathbf{S}_r^{-1} \mathbf{g}_2$$

ou si $(\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{S}_r^{-1} \mathbf{x} > (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{S}_r^{-1} \frac{\mathbf{g}_1 + \mathbf{g}_2}{2}$.

- Règle simple mais inadaptée si les variances sont différentes
- Ne tient pas compte de l'échantillonnage.

Risque bayésien : notations

- $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ muni d'une loi de probabilités $\pi_1, \ldots, \pi_m$.
- qui sont les probabilités *a priori* des classes ω_ℓ.
- $\mathbf{x} \mid T$ admet une loi de densité

$$f_\ell(\mathbf{x}) = P[\mathbf{x} \mid \mathcal{T}_\ell].$$

- Application $\delta : \Omega \longmapsto \{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ dépendant de la
  - connaissance ou non de coûts de mauvais classement,
  - connaissance ou non des lois *a priori* sur les classes,
  - nature aléatoire ou non de l'échantillon.

Risque bayésien : définition

- Associé à δ ou coût moyen :

$$R_\delta = \sum_{k=1}^{m} \pi_k \sum_{\ell=1}^{m} c_{\ell \mid k} \int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$$

Avec

- $c_{\ell \mid k}$: coût du classement dans $\mathcal{T}_\ell$ d'un individu de $\mathcal{T}_k$.
- $\int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$:
- Probabilité d'affecter $\mathbf{x}$ à $\mathcal{T}_\ell$ alors qu'il est dans $\mathcal{T}_k$.

Coûts inconnus supposés égaux

- Règle de Bayes : affecter $\mathbf{x}$ à la classe la plus probable
- Celle qui maximise la probabilité conditionnelle *a posteriori* : $P[\mathcal{T}_\ell \mid \mathbf{x}]$.

$$P[\mathcal{T}_\ell \mid \mathbf{x}] = \frac{P[\mathcal{T}_\ell \text{ et } \mathbf{x}]}{P[\mathbf{x}]} = \frac{P[\mathcal{T}_\ell].P[\mathbf{x} \mid \mathcal{T}_\ell]}{P[\mathbf{x}]}$$

- La règle de décision s'écrit :

$$\delta(\mathbf{x}) = \arg \max_{\ell=1,\ldots,m} \pi_\ell f_\ell(\mathbf{x}).$$

Les probabilités *a priori* π_ℓ sont

- connues comme proportions de groupes
- estimées sur un échantillon aléatoire
- inconnues et considérées égales

Si les probabilités *a priori* sont égales

- On maximise $f_\ell(\mathbf{x})$
- C'est la vraisemblance de $\mathbf{x}$ au sein de T_ℓ
- Si $m = 2$, $\mathbf{x}$ est affectée à T_1 si :

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \quad \text{(rapport de vraisemblance)}$$

- Problème : estimer les densités conditionnelles $f_\ell(\mathbf{x})$

Cas gaussien, variances inégales

- Hypothèse : $\mathbf{x} \mid T \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$
- Densité de $\mathbf{x}$ au sein de $\mathcal{T}_\ell$:

$$f_\ell(\mathbf{x}) = \frac{1}{\sqrt{2\pi}(\det(\boldsymbol{\Sigma}_\ell))^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)\right].$$

- Affectation de $\mathbf{x}$ par maximisation de $\pi_\ell . f_\ell(\mathbf{x})$:

$$\max_\ell \left[\ln(\pi_\ell) - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma}_\ell)) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ell)\right].$$

Cas gaussien, variances inégales

- Les matrices $\boldsymbol{\Sigma}_\ell$ dépendent de ℓ.
- Le critère d'affectation est *quadratique* en $\mathbf{x}$.
- Les π_ℓ sont connues ou égales.
- les $\boldsymbol{\mu}_\ell$ et les $\boldsymbol{\Sigma}_\ell$ sont estimées :

$$\widehat{\boldsymbol{\mu}_\ell} = \mathbf{g}_\ell \quad \text{et} \quad \mathbf{S}^*_{Rl} = \frac{1}{n_\ell - 1} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'.$$

Cas gaussien, variances égales

- Le critère devient : $\ln(\pi_\ell) - \frac{1}{2}\boldsymbol{\mu}_\ell' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_\ell + \boldsymbol{\mu}_\ell' \boldsymbol{\Sigma}^{-1} \mathbf{x}$
- *linéaire* en $\mathbf{x}$.
- $\boldsymbol{\Sigma}$ est estimée par : $\mathbf{S}_R^* = \frac{1}{n-m} \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)'$
- Si les probabilités π_ℓ sont égales :

$$\overline{\mathbf{x}_\ell}' \mathbf{S}_R^{*-1} \mathbf{x} - \frac{1}{2} \overline{\mathbf{x}_\ell}' \mathbf{S}_R^{*-1} \overline{\mathbf{x}_\ell}$$

- C'est le critère élémentaire issu de l'AFD.

Cas non paramétrique

- Pas d'hypothèse (normalité) sur la loi
- Hypothèse de régularité sur la fonction de densité f
- Estimation fonctionnelle de la densité $f(x)$ par $\widehat{f}(x)$.
- Échantillon de grande taille surtout si p est grand
- *The curse of dimensionality* ou fléau de la dimension
- Pour l'analyse discriminante : estimation des $f_\ell(\mathbf{x})$

Méthode du noyau

- $x_1, \ldots, x_n$ n observations d'une v.a.r. X de densité f inconnue.
- $K(y)$ (*noyau*) : densité de probabilité unidimensionnelle ;
- h (largeur de fenêtre) un réel positif.

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

- K est choisi gaussien, uniforme ou triangulaire.

Application à l'analyse discriminante

- Estimation non paramétrique de chaque $f_\ell(\mathbf{x})$
- Noyau K^* multidimensionnel
- K^*densité d'une loi multivariée ou
- ou produit de lois univariées $K^*(\mathbf{x}) = \prod_{j=1}^{p} K(x^j)$

$$\widehat{f}_\ell(\mathbf{x}) = \frac{1}{n_\ell h^p} \sum_{i \in \Omega_\ell} K^* \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right).$$

k-pp : k plus proches voisins

1. Choix d'un entier k : $1 \geq k \geq n$
2. Calculer les distances $d_{\mathbf{S}_R^{-1}}(\mathbf{x}, \mathbf{x}_i)$, $\quad i = 1, \ldots, n$
3. $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}$, les k observations les plus proches ;
4. Nombres d'occurences $k_1, \ldots, k_m$ que ces k observations dans chacune des classes,
5. Estimer les densités par $\widehat{f_\ell}(\mathbf{x}) = \frac{k_\ell}{kV_k(\mathbf{x})}$; où $V_k(\mathbf{x})$ est le volume de l'ellipsoïde $\{\mathbf{z} | (\mathbf{z}-\mathbf{x})' \mathbf{S}_R^{-1} (\mathbf{z}-\mathbf{x}) = d_{\mathbf{S}_R^{-1}}(\mathbf{x}, \mathbf{x}_{(k)})\}$.

Remarques

- Version simplifiée : $V_k(\mathbf{x}) = 1$
- Si $k = 1$, $\mathbf{x}$ est affecté à la classe du plus proche élément
- Si $k = 1$, erreur d'estimation nulle !
- Choix important de la distance entre observations
- Réglage des paramètre : h(largeur de fenêtre) ou k
- par validation croisée ou échantillon de validation
- Estimation de densité déconseillée par Vapnik

Cancer : taux d'erreur

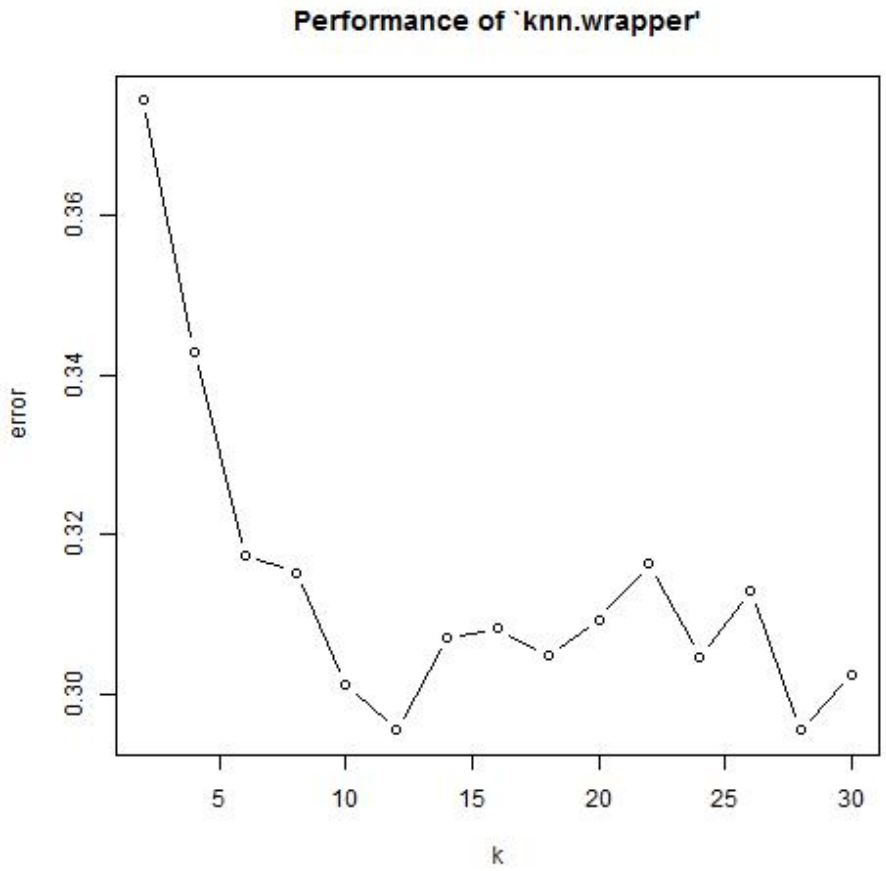
Méthode	apprentissage	validations croisée	test
linéaire	1,8	3,8	3,6
kNN	2,5	2,7	2,9

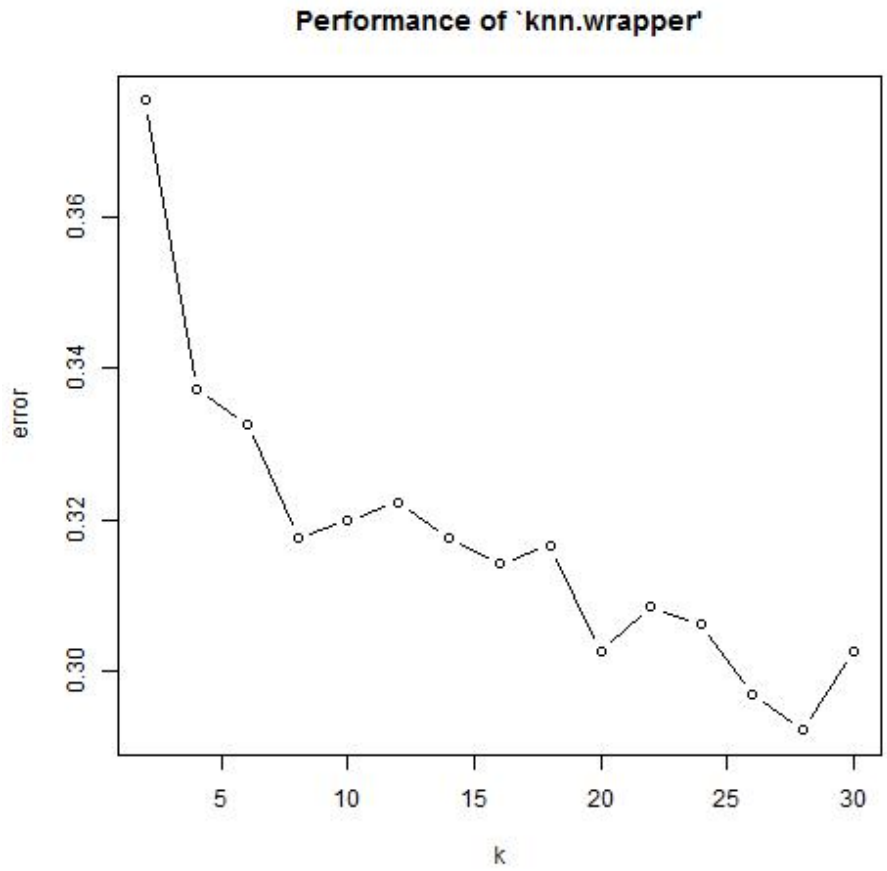
Concentration d'ozone : taux d'erreur

Méthode	apprentissage	validations croisée	test
linéaire	11,9	12,5	12,0
quadratique	12,7	14,8	12,5

Carte visa : taux d'erreur

Méthode	apprentissage	validations croisée	test
linéaire	16,5	18,3	18
quadratique	17,8	22,0	30
kNN	23,5	29,8	29





Carte visa : deux exécutions de la validation croisée.