Estimation et choix de modèles en classification semi-supervisée

V. Vandewalle

Soutenance de thèse
Encadré par C. Biernacki, G. Celeux

Villeneuve d'Ascq, le 9 décembre 2009

Contexte

Estimation

Juger de la pertinence d'un modèle

Choix de modèle

Modèles multinomiaux parcimonieux

Conclusion et perspectives

Forme des données et problème posés

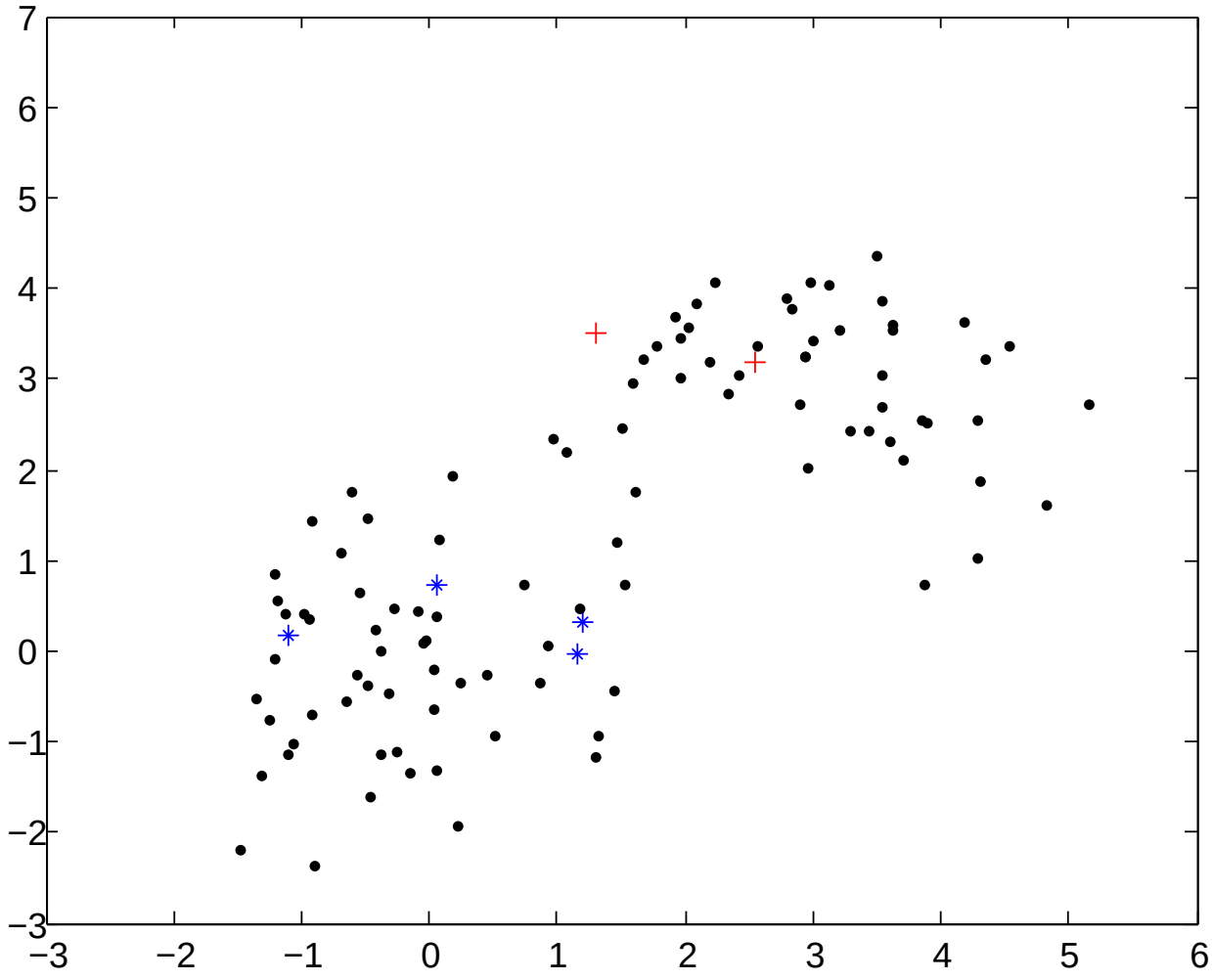
Données partiellement classées

- n_ℓ observations classées :
  $(\mathbf{x}_\ell, \mathbf{z}_\ell) = \{(\boldsymbol{x}_1, \boldsymbol{z}_1), (\boldsymbol{x}_2, \boldsymbol{z}_2), \ldots, (\boldsymbol{x}_{n_\ell}, \boldsymbol{z}_{n_\ell})\}$
- $\boldsymbol{x}_i$ le vecteur des covariables de l'individu i
- $\boldsymbol{z}_i$ la classe d'appartenance de l'individu i
- n_u données non classées : $\mathbf{x}_u = \{\boldsymbol{x}_{n_\ell+1}, \boldsymbol{x}_{n_\ell+2}, \ldots, \boldsymbol{x}_{n_\ell+n_u}\}$.
  Typiquement $n_u >> n_\ell$

Problème de classification

- Classification *supervisée* :
  - Méthodes prédictives
  - Méthodes génératives
- Classification *non supervisée* :
  - Méthodes géométriques
  - Modèles probabilistes : Modèles de mélange (méthodes génératives)

Illustration



Cadre semi-supervisé

Différentes questions et différentes méthodes

- Il faut choisir un point de vue : *supervisé* ou *non supervisé*.
- Il faut choisir une méthode.

Objectif privilégié : Analyse discriminante

Amélioration de la précision de la règle de classement apprise à partir des données non classées.

Analyse discriminante

Diverses méthodes dans le cadre supervisé

- Modèles génératifs :
  - Analyse discriminante linéaire de Fisher
  - Analyse discriminante quadratique
- Modèles prédictifs :
  - Régression logistique
  - Séparateurs vaste marge
  - k plus proches voisins
  - Arbres de régression et de classification

Illustration méthodes génératives

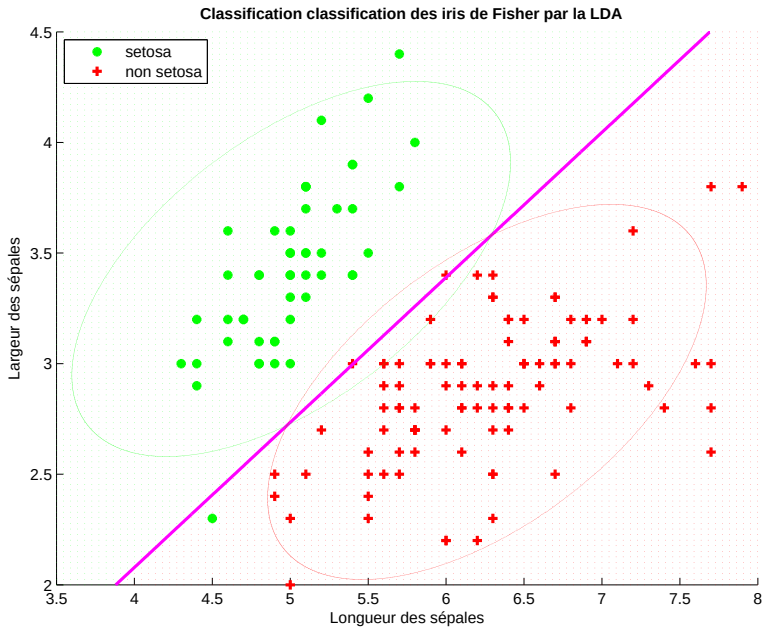


Fig.: Illustration de la LDA sur les iris de Fisher.

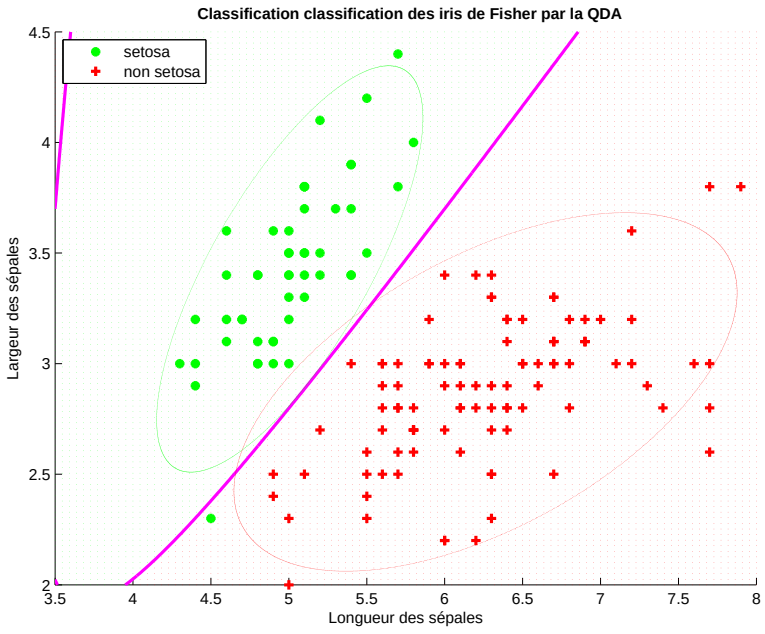


Fig.: Illustration de la QDA sur les iris de Fisher.

Illustration des méthodes prédictives

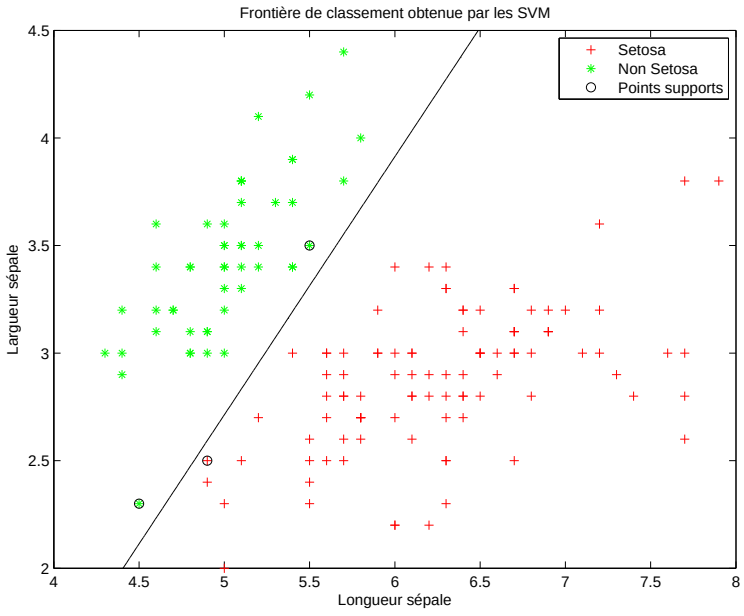


Fig.: Régle de classement apprise par les SVM.

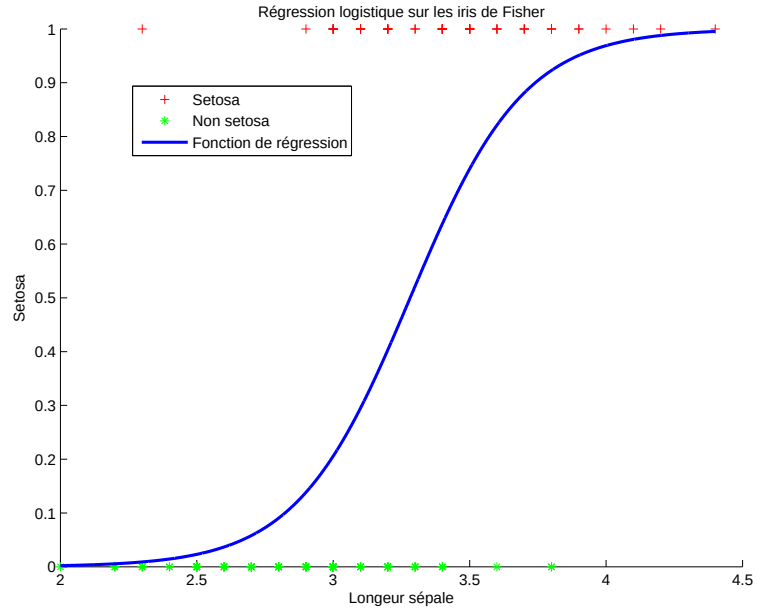


Fig.: Régression logistique sur la variable longueur des sépales des iris de Fisher pour la distinction *setosa*/non-*setosa*.

Différence entre modèles génératifs et prédictifs

Modèles génératifs

- Modélisation de $p(x, z)$
- Donc modélisation de $p(x) = \sum_z p(x, z)$
- Prise en compte « naturelle » des données non classées

Modèles prédictifs

- Modélisation de $p(z|x)$, voire de $\mathbf{1}\{p(z|x) > 1/2\}$
- Hypothèses sur $p(x)$ évitées
- Prise en compte difficile des données non classées

Les hypothèses du semi-supervisé (1/2)

Situation souhaitable (Chapelle *et al. eds.*, 2006) :

- **Hypothèse de régularité :** Si x_1 et x_2 proches dans des zones de forte densité alors z_1 et z_2 proches.
- **Hypothèse de cluster :** Si x_1 et x_2 dans le même *cluster* alors il est probable qu'ils soient dans la même classe.
- **Hypothèse de séparation par zones de faible densité :** La frontière de classification se trouve dans des zones de faible densité.
- **Hypothèse de dimensionnalité :** Les données en grande dimension appartiennent à des espaces de petites dimensions.

Applications aux modèles prédictifs

- Propagation des étiquettes dans un graphe (Zhou *et al.*, 2004)
- Régularisation par l'entropie (Grandvalet & Bengio, 2006)
- SVM transductifs (Vapnik, 1998)

Les hypothèses du semi-supervisé (2/2)

Applications aux modèles génératifs

- Hypothèse de *cluster* faite d'office
- Hypothèse de séparation des classes nécessaire pour un faible taux d'erreur
- Hypothèse de dimensionalité faite par certains modèles génératifs

Conclusion

Les hypothèses où le semi-supervisé est susceptible de bien fonctionner sont les hypothèses où l'utilisation d'un modèle génératif fait sens.

Hypothèses d'échantillonnage

- g classes
- $(\mathbf{X}, \mathbf{Z})$ couple de v.a. à valeurs dans $\mathcal{X} \times \mathcal{Z}$
- $\mathcal{X}$ l'espace des covariables
- $\mathcal{Z} = \{0, 1\}^g$ l'espace des étiquettes

- $(\mathbf{x}_\ell, \mathbf{z}_\ell)$ provient de n_ℓ réalisations i.i.d. de $(\mathbf{X}, \mathbf{Z})$
- $\mathbf{x}_u$ provient de n_u réalisations i.i.d. de $\mathbf{X}$
- n_ℓ est la réalisation $N_\ell \sim \mathcal{B}(n, \beta)$
- chaque donnée est étiquetée de manière indépendante avec une probabilité $\beta \in]0, 1[$

Conséquence

Les données manquantes (ici les étiquettes des données non classées) sont manquantes totalement au hasard.

Modèle génératif

- $(\mathbf{X}, \mathbf{Z})$ admet une densité de probabilité p sur $\mathcal{X} \times \mathcal{Z}$.
- p appartient à une famille paramétrique paramétrée par $\theta \in \Theta$.
  $\exists \theta^* \in \Theta$ tel que $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)\ \forall (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}$.
- $p(\boldsymbol{x}, \boldsymbol{z}; \theta) = \prod_{k=1}^{g} (\pi_k p(\boldsymbol{x}; \theta_k))^{z_k}$ avec :
  - π_k la probabilité d'appartenir à la classe k
  - θ_k les paramètres spécifiques à la distribution de la classe k
  - $\theta = (\pi_1, \ldots, \pi_{g-1}, \theta_1, \ldots, \theta_g)$
- Génération des données :
  - $\mathbf{Z} \sim \mathcal{M}(1, \pi_1^*, \ldots, \pi_g^*)$
  - $\mathbf{X} | Z_k = 1$ a pour densité $p(\cdot; \theta_k^*)$
- $p(\boldsymbol{x}; \theta) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \theta_k)$: *distribution mélange*
- La règle de classement optimale pour un individu est
  $\hat{k} = \arg\max_k p(z_k = 1 | \boldsymbol{x}; \theta^*)$ où $p(z_k = 1 | \boldsymbol{x}; \theta^*) \propto \pi_k^* p(\boldsymbol{x}; \theta_k^*)$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} \mathbb{E}[Z_{ik} | \mathbf{x}_i; \theta^{(r)}] \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} p(Z_{ik} = 1 | \boldsymbol{x}_i; \theta^{(r)}) \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} t_{ik}^{(r+1)} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} t_{ik}^{(r+1)} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

Estimation par maximum de vraisemblance (1/2)

Log-vraisemblance :

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données classées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données non classées}}$$

Algorithme EM (Dempster *et al.* 1977)

Initialisation de l'algorithme $\theta^{(0)}$

Jusqu'à convergence, boucler :

- **Étape E :** Calcul de l'espérance de la vraisemblance complétée

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} t_{ik}^{(r+1)} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))$$

- **Étape M :** $\theta^{(r+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(r)})$

Estimation par maximum de vraisemblance (2/2)

Propriétés de EM

- $\mathcal{L}(\theta^{(r+1)}; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) \geq \mathcal{L}(\theta^{(r)}; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)$
- Convergence vers un point stationnaire de la vraisemblance

Avantages

- Bonne initialisation disponible : initialisation à partir des données classées.
- Réduction de la variance asymptotique sous l'hypothèse du bon modèle.

Inconvénients

- Numériquement plus long que l'estimation supervisée.
- Possibilités de dégradation si le modèle est éloigné de la réalité.

Exemple de modèle utilisé si $\mathcal{X} = \mathbb{R}^d$ (1/2)

Distribution Gaussienne multivariée conditionnellement à la classe.

- $\mathbf{X}|Z_k = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$
- μ_k la moyenne de la classe k
- Σ_k la matrice de covariance de la classe k

Modèles parcimonieux pour Σ_k

EDDA (Bensmail et Celeux 1996)

- $\Sigma_k = \lambda_k D_k A_k D_k'$
- λ_k volume, D_k orientation, A_k forme
- 14 modèles parcimonieux

HDDA (Bouveyron *et al.* 2006)

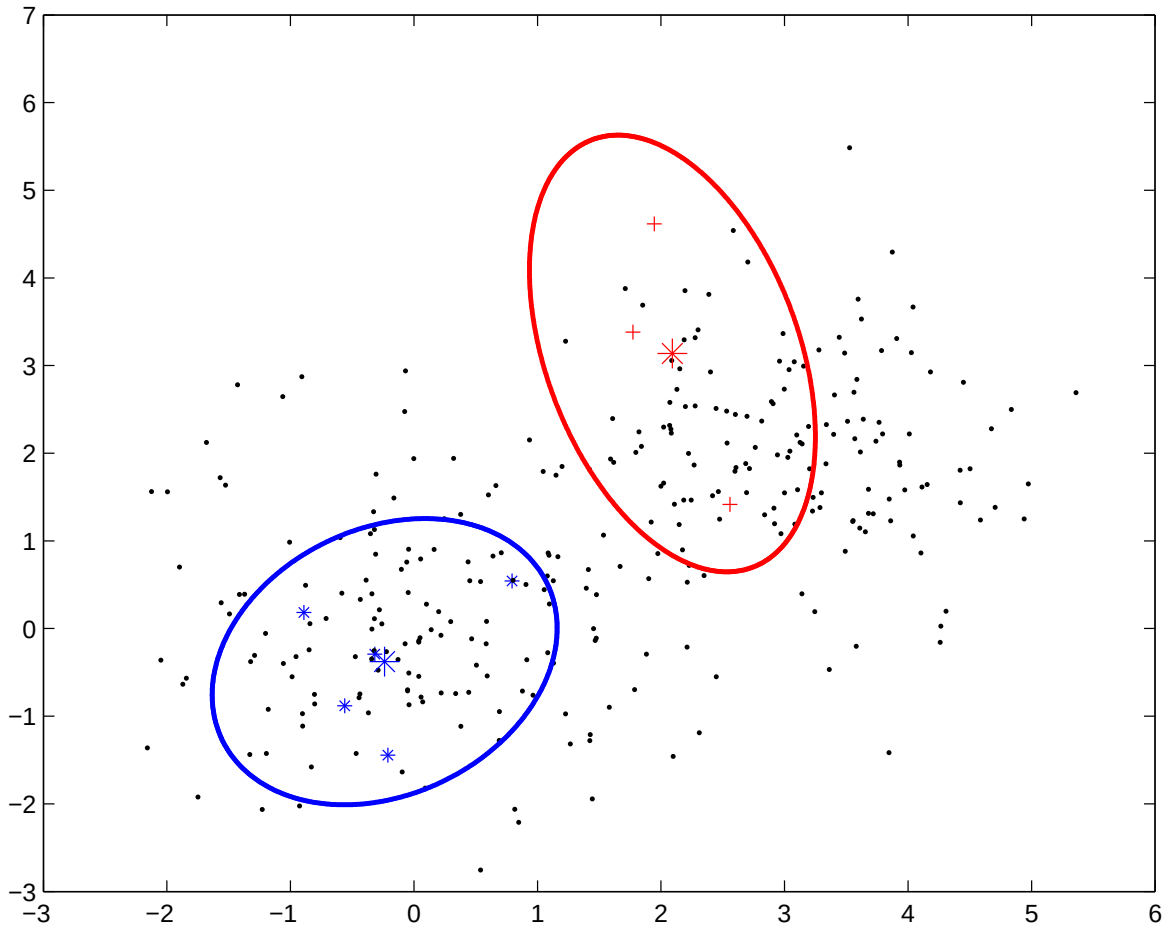
- $\Sigma_k = D_k \Delta_k D_k'$
- Δ_k matrices diagonales des vp
- Les p plus petites vp identiques

Exemple de modèle utilisé si $\mathcal{X} = \mathbb{R}^d$ (2/2)

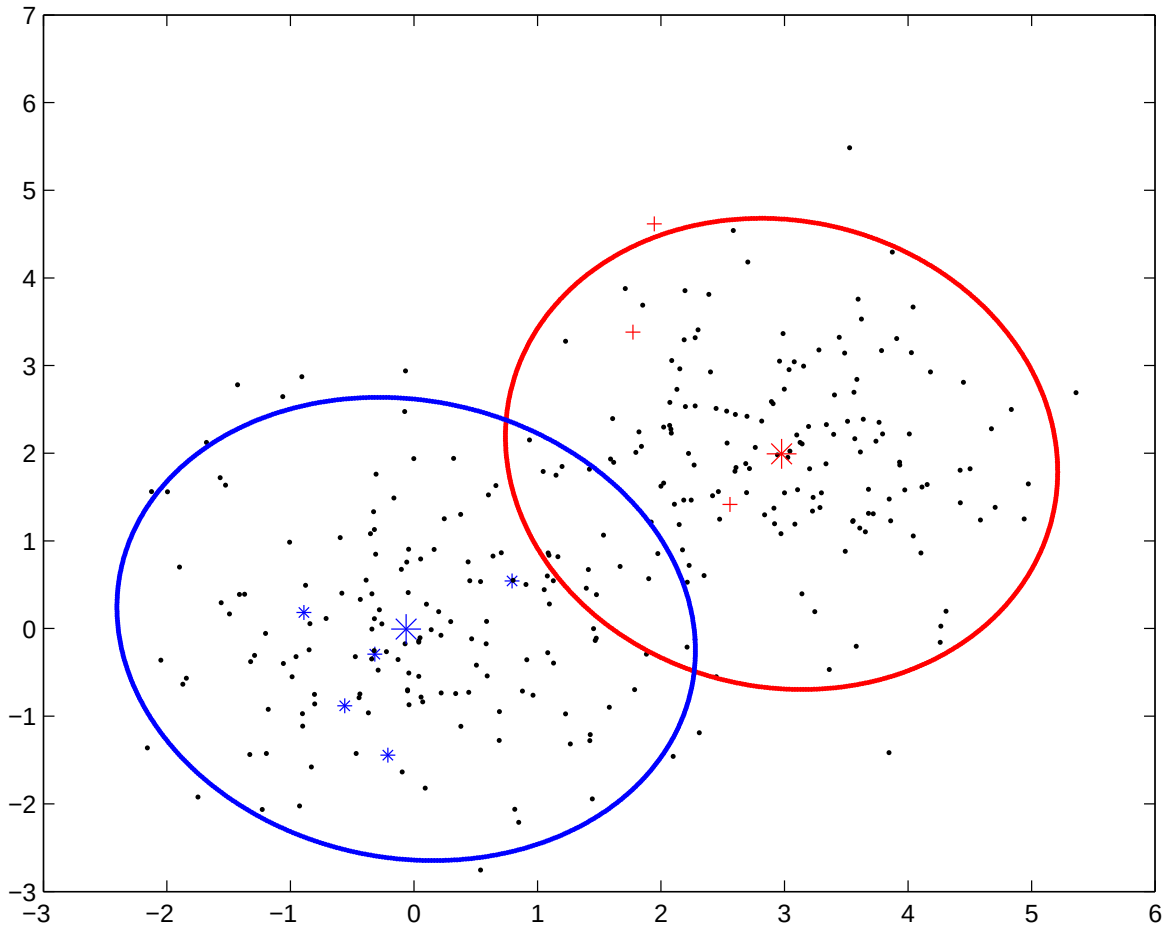
Estimation des paramètres : Étape M

$$
\begin{aligned}
n_k^{(r+1)} &= \sum_{i=1}^{n_\ell} z_{ik} + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} \\
\pi_k^{(r+1)} &= \frac{n_k^{(r+1)}}{n} \\
\mu_k^{(r+1)} &= \frac{\sum_{i=1}^{n_\ell} z_{ik} \boldsymbol{x}_i + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} \boldsymbol{x}_i}{n_k^{(r+1)}} \\
\Sigma_k^{(r+1)} &= \frac{\sum_{i=1}^{n_\ell} z_{ik} (\boldsymbol{x}_i - \mu_k^{(r+1)})(\boldsymbol{x}_i - \mu_k^{(r+1)})'}{n_k^{(r+1)}} \\
&\quad + \frac{\sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} (\boldsymbol{x}_i - \mu_k^{(r+1)})(\boldsymbol{x}_i - \mu_k^{(r+1)})'}{n_k^{(r+1)}}
\end{aligned}
$$

Illustration



Illustration



Modèles utilisés dans le cas discret (1/2)

Données

- $\boldsymbol{x} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d)$
- $\boldsymbol{x}^j = (x^{j1}, \ldots, x^{jm_j})$
- $x^{jh} = 1$ si l'individu présente la modalité h de la variable j et 0 sinon.

Modèle d'indépendance conditionnelle (Everitt 1984)

- $\alpha_k^{jh} = P(X^{jh} = 1 | Z_k = 1)$
- $\alpha_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$
- $\alpha_k = (\alpha_k^1, \ldots, \alpha_k^d)$
- $\theta = (\pi_1, \ldots, \pi_{g-1}, \alpha_1, \ldots, \alpha_g)$
- $p(\boldsymbol{x}|\alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}}$

Modèles utilisés dans le cas discret (2/2)

Données

- d mots $(w_1, \ldots, w_d)$
- Texte de longueur ℓ
- $\boldsymbol{x} = (x^1, \ldots, x^d)$
- x^j le nombre d'occurrences du mot w_j dans le texte

Modèle multinomial (Nigam *et al.* 1999)

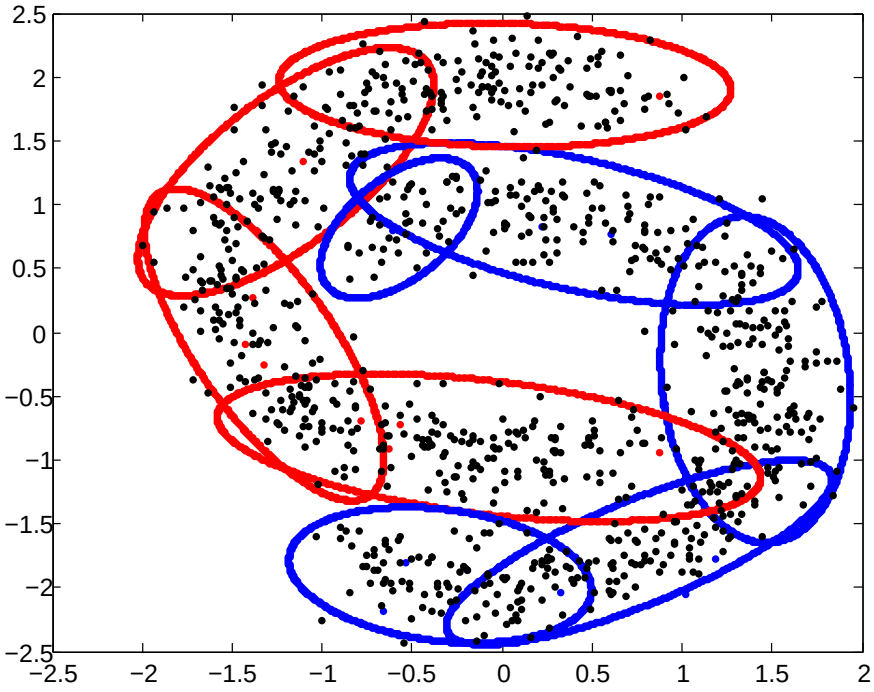
- $\mathbf{X}|Z_k = 1 \sim \mathcal{M}(\ell, \alpha_k^1, \ldots, \alpha_k^d)$
- α_k^j fréquence du mot w_j dans les textes de la classe k

Modèles utilisés

Modèles à plusieurs composants par classe

La densité conditionnellement à la classe est un mélange (Miller et Uyar 1997) :

$$p(\boldsymbol{x}; \theta_k) = \sum_{h=1}^{H_k} \tau_{hk} p(\boldsymbol{x}; \theta_{hk})$$



Remarques

- Assez naturel une fois qu'on travaille avec des mélanges
- Deux niveaux de données manquantes

Bilan sur les modèles

Réutilisation facile

Simple utilisation de l'algorithme EM

Plusieurs modélisations possibles :

- Paramétrisation de Σ_k
- Choix du nombre de composants par classe
- Choix des variables à retenir

Question

- Le modèle proposé est-il pertinent ?
- Comment choisir le modèle qui produit les meilleures performances compte tenu de l'objectif d'analyse discriminante ?

Heuristique

Principe

Si les données proviennent du modèle postulé, les estimations supervisée ($\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$) et non supervisée ($\hat{\theta}_{\mathbf{x}_u}$) devraient être proches.

Soit :

$$\theta_s^* = \arg\max_{\theta\in\Theta} \mathbb{E}_{\mathbf{X},\mathbf{Z}}[\log p(\mathbf{X},\mathbf{Z};\theta)].$$

et

$$\theta_{ns}^* = \arg\max_{\theta\in\Theta} \mathbb{E}_{\mathbf{X}}[\log p(\mathbf{X};\theta)].$$

On a $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} \xrightarrow{P} \theta_s^*$ et $\hat{\theta}_{\mathbf{x}_u} \xrightarrow{P} \theta_{ns}^*$

- Si le modèle est correct $\theta_s^* = \theta_{ns}^*$.
- Si le modèle n'est pas correct $\theta_s^* \neq \theta_{ns}^*$ en général.

Formalisation comme un test d'hypothèse

On teste $H_0 : \{\theta_s^* = \theta_{ns}^*\}$ contre $H_1 : \{\theta_s^* \neq \theta_{ns}^*\}$.
Test utilisé : rapport des vraisemblances maximales (LRT).
La statistique de test s'écrit :

$$LRT = 2[\mathcal{L}(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}; \mathbf{x}_\ell, \mathbf{z}_\ell) + \mathcal{L}(\hat{\theta}_{\mathbf{x}_u}; \mathbf{x}_u) - \mathcal{L}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)].$$

ν : le nombre de paramètres.
Sous des conditions de régularité classiques,

$$LRT \approx \chi^2_\nu$$

Formalisation comme un choix de modèle

M_1 : associe les données étiquetées et non étiquetées pour estimer les paramètres

$$\begin{aligned} M_1 \quad = \quad & \{\forall (\boldsymbol{x}, z, s) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{S} \\ & \exists \theta \in \Theta / p(\boldsymbol{x}, z, s) = p(\boldsymbol{x}, z) p(s) = p(\boldsymbol{x}, z; \theta) p(s; \beta)\}, \end{aligned}$$

M_2 : dissocie les données étiquetées et non étiquetées pour estimer les paramètres

$$\begin{aligned} M_2 \quad = \quad & \{\forall (\boldsymbol{x}, z) \in \mathcal{X} \times \mathcal{Z} \\ & \exists (\theta, \theta') \in \Theta^2 / p(\boldsymbol{x}, z, s = 1) = p(\boldsymbol{x}, z; \theta) \beta \\ & \text{ et } p(\boldsymbol{x}, z, s = 0) = p(\boldsymbol{x}, z; \theta')(1 - \beta)\}. \end{aligned}$$

$$BIC(M_1) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \hat{\theta}_{\mathbf{x}, \mathbf{z}_\ell}) - \frac{\nu}{2} \log n.$$

$$BIC(M_2) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x}_\ell, \mathbf{z}_\ell}) + \log p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x}_u}) - \frac{\nu}{2} \log n_\ell n_u.$$

Illustration sur un exemple (1/2)

Mélange de deux gaussiennes : $\mu_1 = \begin{pmatrix} 0 \\ 3/2 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 3/2 \\ 0 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 4/5 \\ 4/5 & 1 \end{pmatrix}$, $\pi_1 = 3/5$. Utilisation du modèle $[\pi_k \lambda C]$ à tort :

$n_\ell \backslash n_u$	20	40	80	100	200	500	1000	5000
20	87	91	95	96	98	97	99	99
40	97	100	95	99	99	98	100	98
80	92	93	95	99	95	99	99	100
100	93	96	95	98	96	94	96	97
200	93	84	88	84	93	86	91	83
500	81	77	74	64	52	30	18	10
1000	82	74	57	62	21	0	0	0
5000	84	77	50	41	6	0	0	0

Tab.: Nombre de fois où le modèle M_1 est choisi pour n_ℓ données étiquetées et n_u données non étiquetées sur 100 simulations.

Illustration sur un exemple (2/2)

$n_\ell \backslash n_u$	20	40	80	100	200	500	1000	5000
20	76	79	69	60	58	60	48	49
40	86	79	64	74	64	56	59	60
80	90	87	71	77	68	54	56	40
100	83	78	74	62	69	64	53	58
200	83	77	71	69	57	56	62	49
500	90	68	39	37	30	27	33	37
1000	89	63	32	23	8	10	9	26
5000	88	50	11	8	0	0	0	0

Tab.: Nombre de fois sur 100 simulations, où la distribution jointe estimée à partir du modèle M_1 est plus proche de la distribution jointe d'échantillonnage (au sens de Kullback) que de la distribution jointe estimée à partir du modèle M_2.

Choix adaptatif d'un modèle

Si le modèle M_2 est choisi

L'usage du modèle paramétrique proposé n'est pas appropié.

Solution

Proposer des modèles de plus en plus complexes tant que M_2 est choisi. Par exemple augmenter le nombre de composants par classe.

Illustration sur les deux lunes enchevêtrées (1/2)

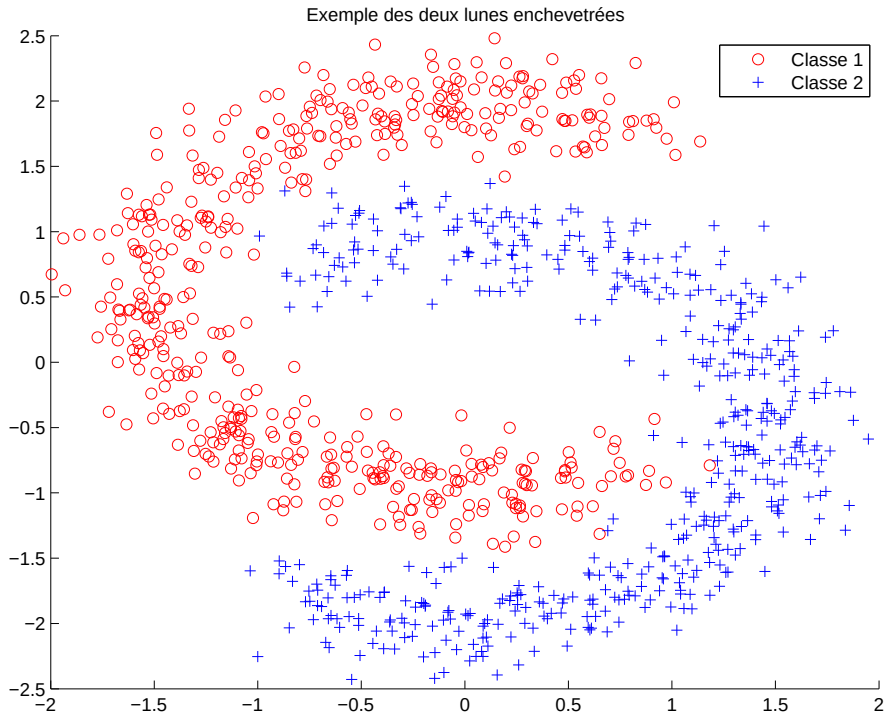


Fig.: Exemple des deux lunes enchevêtrées.

80% des étiquettes sont cachés aléatoirement.

Illustration sur les deux lunes enchevêtrées (2/2)

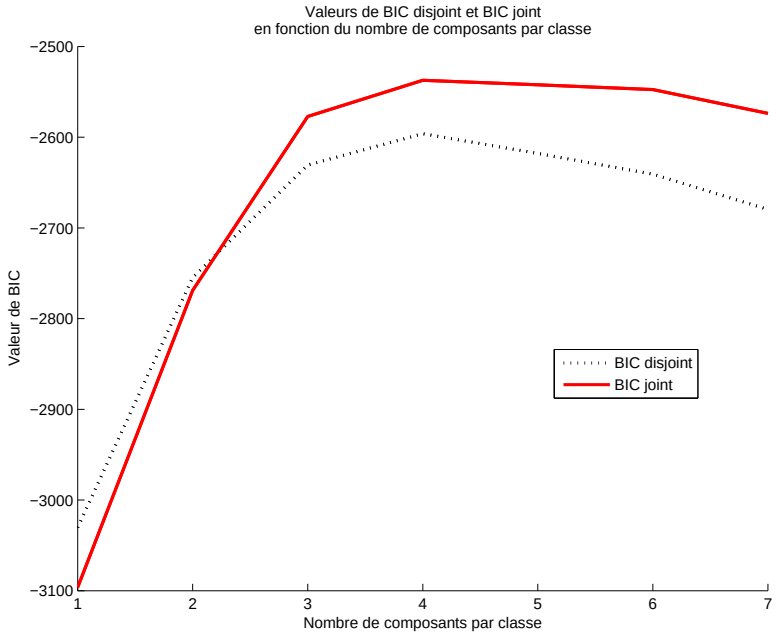


Fig.: Critères BIC disjoint et BIC joint.

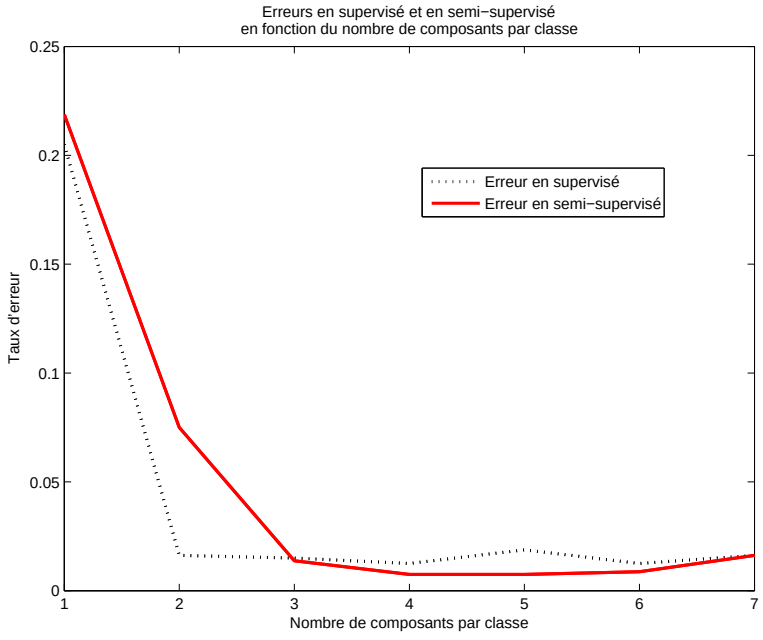


Fig.: Taux d'erreur supervisé et semi-supervisé.

Critères de choix de modèle

Dans le cadre de l'estimation par ML

- Approximation de la déviance moyenne :

$$AIC(m) = 2 \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 2\nu_m$$

- Approximation du log de la vraisemblance intégrée :

$$BIC(m) = \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \frac{\nu_m}{2} \log n$$

Critère universel en analyse discriminante

Validation croisée du taux d'erreur

Remarques sur les critères

AIC et BIC

- Ne prennent pas en compte l'objectif décisionnel
- Minimisent asymptotiquement :
  $\beta DK(p(\mathbf{X}, \mathbf{Z}), p(\mathbf{X}, \mathbf{Z}; m)) + (1 - \beta) DK(p(\mathbf{X}), p(\mathbf{X}; m))$
- Dangereux dans le cadre semi-supervisé puisque β peut être petit

Validation croisée

- Potentiellement coûteuse en temps
- Nécessite le choix de k en k-fold

Prise en compte de l'objectif décisionnel : BEC (1/2)

Idée

- Dans un cadre décisionnel on veut un modèle qui explique bien les étiquettes.
- C'est-à-dire qui maximise $p(\mathbf{z}|\mathbf{x}, m)$.

On a :

$$\log p(\mathbf{z}|\mathbf{x}, m) = \log p(\mathbf{x}, \mathbf{z}|m) - \log p(\mathbf{x}|m).$$

Après deux approximations BIC, on obtient le critère BEC (Bouchard et Celeux 2006) :

$$BEC(m) = \log p(\mathbf{x}, \mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}}) = \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - \log \frac{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})},$$

avec $\hat{\theta}^m_{\mathbf{x}}$ l'EMV obtenu à partir de $\mathbf{x}$.

Prise en compte de l'objectif décisionnel : BEC (2/2)

Propriétés

- Si la distribution d'échantillonnage est incluse dans un seul des modèles en compétition alors celui-ci sera asymptotiquement sélectionné.
- Soient deux modèles m et m' emboîtés. Si la distribution d'échantillonnage est incluse dans m et m' alors : $\lim_{n\to\infty} \mathbb{E}[BEC(m) - BEC(m')] = 0$.
- Problème : apparition d'un plateau

Minimisation de la déviance conditionnelle

On veut minimiser :

$$\Delta_{cond} = 2\mathbb{E}_{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{z}'}[\log p(\mathbf{z}'|\mathbf{x}') - \log p(\mathbf{z}'|\mathbf{x}'; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})],$$

Proposition

Si les données sont issues du modèle m :

$$\Delta_{cond} = 2[\log p(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})] + 2[\nu_m - \mathrm{tr}(JJ_\beta^{-1})] + \mathcal{O}_p(\sqrt{n}),$$

J et J_β sont les informations de Fisher pour les données sans les étiquettes et les données partiellement étiquetées.

En supprimant les termes qui ne dépendent pas de m on a :

$$\mathrm{AIC}^*_{cond}(m) = 2\log p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 2[\nu_m - \mathrm{tr}(JJ_\beta^{-1})].$$

Remplacement de la pénalité par une quantité calculable

Proposition

Soit $\delta = \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})$, si les données proviennent du modèle m, alors $\lim_{n\to\infty} \mathbb{E}[\delta] = \frac{1}{2}[\nu_m - \mathrm{tr}(JJ_\beta^{-1})]$.

Ceci justifie le remplacement de $\nu_m - \mathrm{tr}(JJ_\beta^{-1})$ par $2[\log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})]$, et conduit au critère AIC_{cond} :

$$\mathrm{AIC}_{cond}(m) = 2 \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 4 \log \frac{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})},$$

qu'on peut réécrire :

$$\begin{aligned} \mathrm{AIC}_{cond}(m) &= 2 \log \frac{p(\mathbf{x},\mathbf{z}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})}{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})} - 2 \log \frac{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})} \\ \mathrm{AIC}_{cond}(m) &= 2BEC(m) - 2 \log \frac{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})}. \end{aligned}$$

Comportement pour des modèles emboîtés

Proposition

Si la distribution d'échantillonnage appartient à deux modèles emboîtés m et m' avec $m \subset m'$ et si le nombre de données est assez grand alors

$$\mathbb{E}[\mathrm{AIC}_{cond}(m) - \mathrm{AIC}_{cond}(m')] > 0.$$

Exemple pour le choix de la paramétrisation de Σ_k

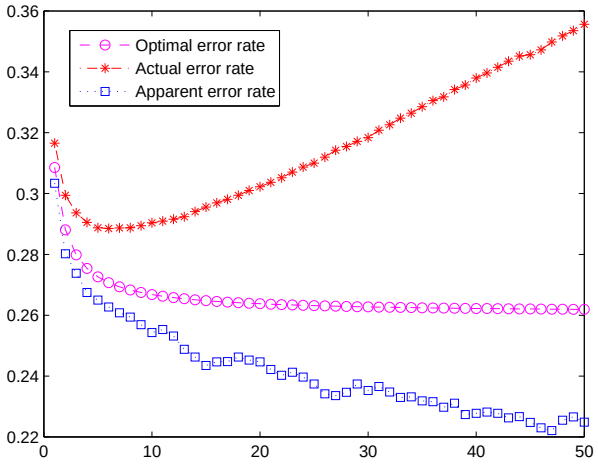
100 échantillons d'apprentissage avec $n_u = 100$, $n_\ell = 20$, et échantillon test de taille 50000. Deux classes $\mathcal{N}((0,0)^t, (1, 0.5; 0.5, 2))$ et $\mathcal{N}((0,2)^t, (1, 0.5; 0.5, 2))$ en proportions identiques.

	BIC	AIC	BEC	AIC_{cond}	CV3	CV10	$\bar{Err}$
λI	9	3	6	6	48	46	29,18
λB	25	9	5	6	9	6	30,18
λC	65	72	38	41	20	30	25,92
$\lambda_k I$	0	0	7	10	7	4	29,29
$\lambda_k B_k$	0	4	8	7	7	3	30,45
$\lambda_k C_k$	1	12	36	30	8	11	27,31
Err^*	26,76	26,86	27,03	26,73	26,83	**26,43**	24,57

Tab.: Nombre de fois où chaque modèle est sélectionné

Exemple en sélection de variables (1/2)

- $P(Z_1 = 1) = P(Z_2 = 2) = 0.5$, $\mathbf{X}|Z_1 = 1 \sim \mathcal{N}(0_{50\times 1}, I_{50})$, $\mathbf{X}|Z_2 = 1 \sim \mathcal{N}(\mu, I_{50})$, $\mu_i = \frac{1}{i}\ \forall i \in \{1, \ldots, 50\}$
- Échantillon test de taille 50000
- Quatre combinaisons de n_ℓ et n_u. S_1 : $n_\ell = 100$, $n_u = 0$; S_2 : $n_\ell = 1000$, $n_u = 0$; SS_1 : $n_\ell = 100$, $n_u = 1000$; SS_2 : $n_\ell = 1000$, $n_u = 10000$
- 100 échantillons d'apprentissage



	BEC	AIC_{cond}	CV3	CV10	NbVar*
S_1	10,5	**3,1**	7,8	7,8	3
S_2	21,7	**11,3**	12,2	14,0	11
SS_1	17,5	**9,2**	10,7	10,0	6
SS_2	33,8	**22,0**	21,1	21,4	23

Exemple en sélection de variables (2/2)

	BEC	AIC_{cond}	CV3	CV10	Err*
S_1	31,53	**30,40**	31,08	31,08	29,68
S_2	27,90	**27,68**	27,77	27,78	27,55
SS_1	30,42	29,75	**29,70**	29,82	28,55
SS_2	27,18	**27,17**	**27,17**	27,21	27,03

Tab.: Taux d'erreur moyen.

- Sélection en moyenne du bon nombre de variables par AIC_{cond}
- Conséquence : erreur moyenne assez faible

Exemple en choix du nombre de composants par classe

- $C_i \sim \mathcal{N}(\mu_i, I_2)$ $\pi_i = 0.25$.
- Classe 1 composée de C_1 et C_2 avec $\mu_1 = (0, 2)^t$ et $\mu_2 = (2, 3)^t$.
- Classe 2 composée de C_3 et C_4 avec $\mu_3 = (0, 0)^t$ et $\mu_3 = (2, 0)^t$.
- $n_\ell = 100$, $n_u = 1000$. Echantillon test de taille 50000.
- Modèle diagonal hétéroscédastique allant de 1 à 5 composants par classe

CC	BIC	AIC	BEC	AIC_{cond}	CV3	CV10	$\bar{Err}$
1	55	0	0	4	27	24	12,41
2	45	91	22	40	31	30	11,78
3	0	8	27	29	19	17	12,03
4	0	1	29	17	9	16	12,21
5	0	0	22	10	14	13	12,36

Tab.: Nombre de composants par classe (CC) sélectionnés par chaque critère.

Exemple sur des données réelles

Choix de la paramétrisation de la matrice de covariance : complet, diagonal, sphérique, homo ou hétéroscédastique.

Dataset	n	d	g	Echantillon test	n_u	n_ℓ
Crab	200	5	4	non	150	50
Iris	150	4	3	non	100	50
Parkinson	195	22	2	non	95	100
Pima	532	7	2	oui	332	200
Wine	178	13	3	non	89	89

Tab.: Dispositif expérimental.

	BIC	AIC	BEC	AIC_{cond}	CV3	CV10
Crab	**6,63**	6,75	6,80	6,77	7,81	7,78
Iris	2,98	2,98	**2,91**	**2,91**	3,25	3,21
Parkinson	26,45	30,68	15,43	**15,16**	18,20	16,38
Pima	25,00	25,00	**19,58**	**19,58**	22,53	**19,58**
Wine	3,24	**1,17**	1,45	1,47	1,73	1,70

Tab.: Taux d'erreur moyen produit par chaque critère.

Remarques

Risques

- $\nu_m - \mathrm{tr}(JJ_\beta^{-1})$ remplacé par une quantité d'espérance requise sous l'hypothèse du bon modèle.
- Si ce n'est pas le cas $2[\log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}})]$ peut devenir très grand par rapport à la pénalité requise.

Solution

Recherche d'une majoration de $\nu_m - \mathrm{tr}(JJ_\beta^{-1})$.

Dimension prédictive

Proposition

$\nu_m - \text{tr}(JJ_\beta^{-1}) \leq \bar{\nu}_m$, avec $\bar{\nu}_m$ qu'on appelle dimension prédictive du modèle m.

$\bar{\nu}_m$ représente le nombre de paramètres algébriquement indépendants impliqués dans l'estimation de la distribution de $\mathbf{Z}|\mathbf{X}$ en maximisant $p(\mathbf{z}|\mathbf{x};\theta)$.

Remarque

- Calcul de $\bar{\nu}_m$ possible pour certains modèles de la famille exponentielle.
- Dans le cas gaussien homoscédastique il est égal au nombre de paramètres de la régression logistique linéaire.

On propose alors le critère AIC_p :

$$\text{AIC}_p(m) = 2 \log p(\mathbf{z}|\mathbf{x}; \hat{\theta}^m_{\mathbf{x},\mathbf{z}}) - 2\bar{\nu}_m$$

Exemple sur données réelles

Choix entre les modèles λC, λB, λI, $\lambda_k C_k$, $\lambda_k B_k$ et $\lambda_k I$.

	d	g	n	n_{test}	AIC	BIC	AIC_p	CV3	CV10
Breast Cancer	30	2	400	169	**4,31**	**4,31**	4,52	4,73	4,76
Wine	13	3	89	89	4,89	2,99	**2,24**	2,94	2,72
Pima	7	2	200	332	23,49	**20,18**	**20,18**	24,10	**20,18**
Crab	5	4	100	100	6,57	5,60	**5,49**	5,84	5,84
Iris	4	3	75	75	2,81	2,93	**2,74**	3,86	3,76
Parkinson	22	2	146	49	**12,59**	12,79	12,89	13,44	13,16
Synt	2	2	250	1000	**10,20**	10,90	10,80	**10,20**	**10,20**
Transfusion	4	2	374	374	28,93	28,93	27,76	24,35	**24,34**
Ionosphere	32	2	175	176	**14,87**	**14,87**	16,14	16,34	16,34

Tab.: Erreur produite par les différents critères de choix de modèle.

Modèles à modalité majoritaire

Paramétrisation proposée par Biernacki *et al.* 2006

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j & \text{si } h = h(k,j), \\ \frac{\varepsilon_k^j}{m_j - 1} & \text{sinon.} \end{cases}$$

Contrainte : $\varepsilon_k^j \leq \frac{m_j - 1}{m_j}$.

Contraintes d'égalité entre classes et/ou entre variables pour le paramètre ε_k^j :

- $[\varepsilon^j]$: on impose aux classes de partager le même paramètre,
- $[\varepsilon_k]$: on impose aux variables de partager le même paramètre,
- $[\varepsilon]$: on impose les deux contraintes précédentes.

Modèles à modalité majoritaire

Variante proposée

$$\alpha_k^{jh} = \begin{cases} 1 - \frac{m_j - 1}{m_j} \varepsilon_k^j & \text{si } h = h(k,j), \\ \frac{\varepsilon_k^j}{m_j} & \text{sinon.} \end{cases}$$

Contrainte : $\varepsilon_k^j \leq 1$.
Contraintes d'égalité entre classes et/ou entre variables pour le paramètre ε_k^j :

- $[\varepsilon^j]$: on impose aux classes de partager le même paramètre,
- $[\varepsilon_k]$: on impose aux variables de partager le même paramètre,
- $[\varepsilon]$: on impose les deux contraintes précédentes.

Comptage du nombre de paramètres

Paramètres discrets

Comment prendre en compte les paramètres dans la pénalité de BIC ?

Solution

Sommer sur tous les états possibles du paramètre discret.

- Fonctionne bien sur des exemples jouets.
- Application difficile à des problèmes concrets avec beaucoup de variables et beaucoup de classes.

Conclusion

Pertinence d'un modèle

- Proposition d'un test pour juger de le pertinence d'un modèle.
- Reformulation du test proposé comme un choix de modèle.

Choix de modèle

- Proposition d'un critère de choix de modèle, AIC_{cond}, dans le cadre semi-supervisé, et du critère AIC_p spécifique au cadre supervisé.
- Bon comportement sur données réelles et simulées.

Modèles multinomiaux parcimonieux

- Proposition de modèles multinomiaux parcimonieux pour lequels les contraintes sont automatiquement satisfaites.
- Proposition d'une variante de BIC pour prendre en compte les paramètres discrets.

Perspectives

Pertinence d'un modèle

- Bâtir un test estimant la partie non supervisée à partir de $(\mathbf{x}_u, \mathbf{x}_\ell)$.
- Utiliser le test proposé pour faire de la sélection de variable.
- Mettre en compétition un modèle génératif et un modèle prédictif.

Choix de modèle

- Mettre en place d'autres stratégies pour estimer $\nu_m - \text{tr}(JJ_\beta^{-1})$.
- Apparition de nouvelles classes dans les données non classées.
- Choix actif des données à étiqueter.

Modèles multinomiaux parcimonieux

- Étude des modèles proposés sur données réelles et simulées.
- Application du critère BIC proposé à des situations réelles en utilisant des méthodes de Monte Carlo.