Chapitre 6
Classification supervisée

Approches basées sur un modèle

Plan

Introduction

Deux approches possibles pour constuire une règle de classification g.

- Approche basée sur un modèle.
  - Apprentissage de $Loi(Y|X)$ puis déduction de g
  - Exemples : analyse discriminante linéaire, bayésien naïf, régression logistique, etc.
- Approche de type prototype.
  - Apprentissage direct de la règle classification g
  - Exemples : k-plus proches voisins, arbres de classification, forêts aléatoires, etc.

Règle de classification de Bayes :

$$g(x) = \arg\max_{\ell \in \{1,\ldots,K\}} \mathbb{P}(Y = \ell | X = x)$$

Dans les approches basées sur un modèle, on distingue :

- l'approche directe comme en régression logistique :

$$\mathbb{P}[Y = 1|X = x] = \frac{exp(X^T\beta)}{1 + exp(X^T\beta)}$$

  Estimation du paramètre β à partir des données d'apprentissage.

- l'approche indirecte comme en analyse discriminante linéaire ou en bayésien naïf. Cette approche utilise la formule de Bayes :

$$\mathbb{P}(Y = k|X = x) = \frac{f(x|Y = k)\mathbb{P}(Y = k)}{\sum_{j=1}^{K} f(x|Y = j)\mathbb{P}(Y = j)}$$

  L'approche indirect nécessite donc l'estimation de $f_k(x) = f(x|Y = k)$ et de $\pi_k = \mathbb{P}(Y = k)$.

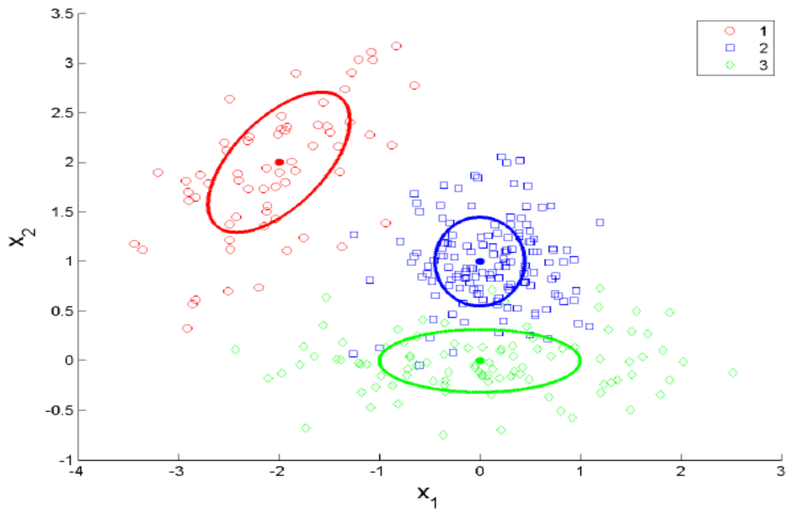
  - $f_k(x)$ prend une forme paramétrique (e.g. gaussienne, etc.) de paramètre θ_k :
  - Estimation des paramètres $\{\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K\}$ à partir des données d'apprentissage

1. Analyse discriminante linéaire et quadratique

- $X \in \mathbb{R}^p$ et $Y \in \{1, \ldots, K\}$
- Ensemble d'apprentissage (X_i, Y_i), $i = 1, \ldots, n$
- Hypothèse paramétrique gaussienne $X \sim \mathcal{N}(\mu_k, \Sigma_k)$ dans chaque groupe k i.e.

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k))$$

- Paramètres inconnus $\theta_k = \{\mu_k, \Sigma_k\}$ et π_k


3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-4
-3
-2
-1
0
1
2
3
x_2
x_1
1
2
3

- Paramètres inconnus estimés par maximum de vraisemblance :

$$\theta = (\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K).$$

- Log-vraissemblance de l'échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$

$$\begin{aligned}
\ell(\theta) &= \log \prod_{i=1}^{n} f(x_i, y_i) \\
&= \sum_{i=1}^{n} \log(\pi_{y_i} f_{y_i}(x_i)) \\
&= \sum_{k=1}^{K} n_k \log(\pi_k) + \sum_{k=1}^{K} \sum_{i:y_i=k} \log(f_k(x_i))
\end{aligned}$$

- Estimateurs

$$\widehat{\pi}_k = \frac{n_k}{n}, \ \widehat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^T$$

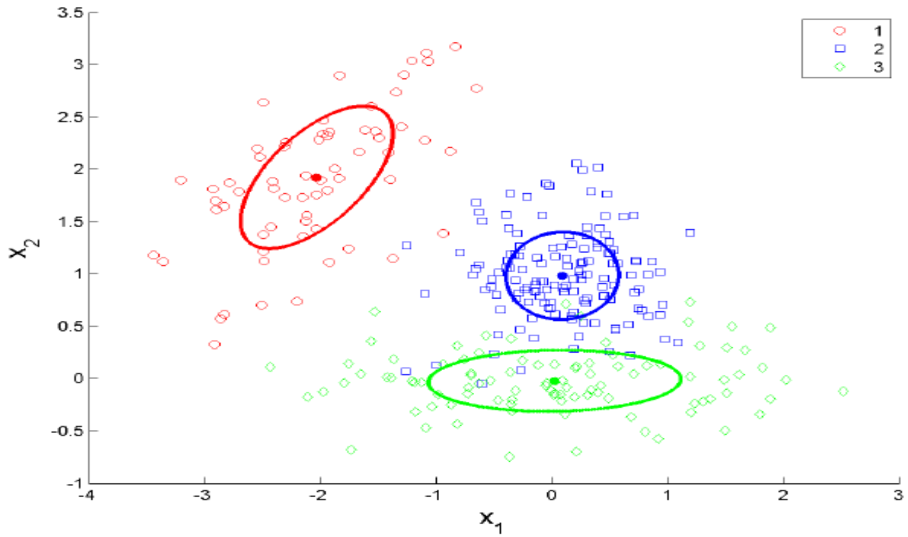
- Règle de classification

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \mathbb{P}(Y=\ell|X=x) \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \log\left(\mathbb{P}(Y=\ell|X=x)\right) \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \delta_\ell(x) \end{aligned}$$

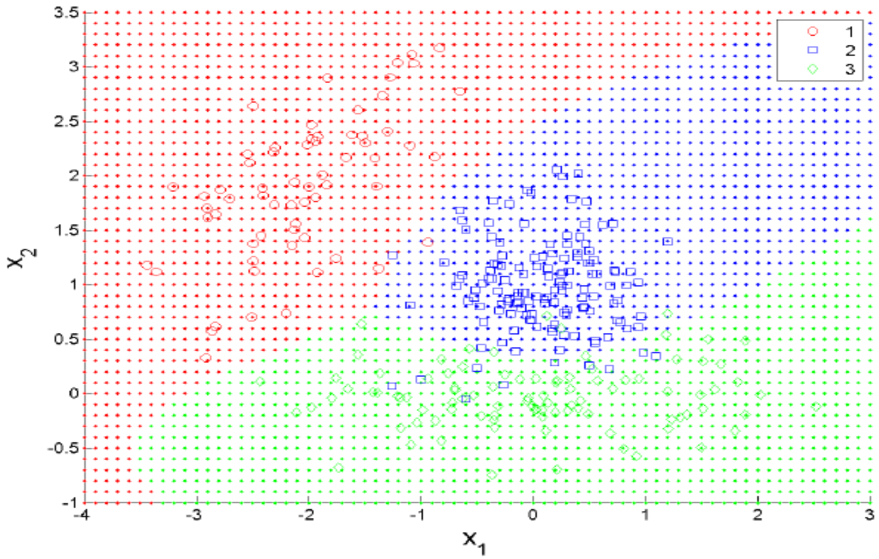
où

$$\delta_\ell(x) = -\frac{1}{2}\log|\widehat{\Sigma}_\ell| - \frac{1}{2}(x-\widehat{\mu}_\ell)^T\widehat{\Sigma}_\ell^{-1}(x-\widehat{\mu}_\ell) + \log(\widehat{\pi}_\ell)$$

- δ_ℓ est appellée fonction discriminante quadratique.
- -2δ_ℓ est appellée dans SAS la distance de Mahalanobis généralisée entre x et $\widehat{\mu}_\ell$.
- La frontière de décision entre deux classes k et ℓ est décrite par une équation quadratique en x $\{x : \delta_k(x) = \delta_\ell(x)\}$



1
2
3
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-4
-3
-2
-1
0
1
2
3
x_2
x_1


1
2
3
x_1
x_2

- On suppose maintenant que $\Sigma_k = \Sigma$ pour tout k
- L'estimée du maximum de vraisemblance de Σ est la matrice de covariance intra-groupe définie par :

$$\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} n_k \widehat{\Sigma}_k$$

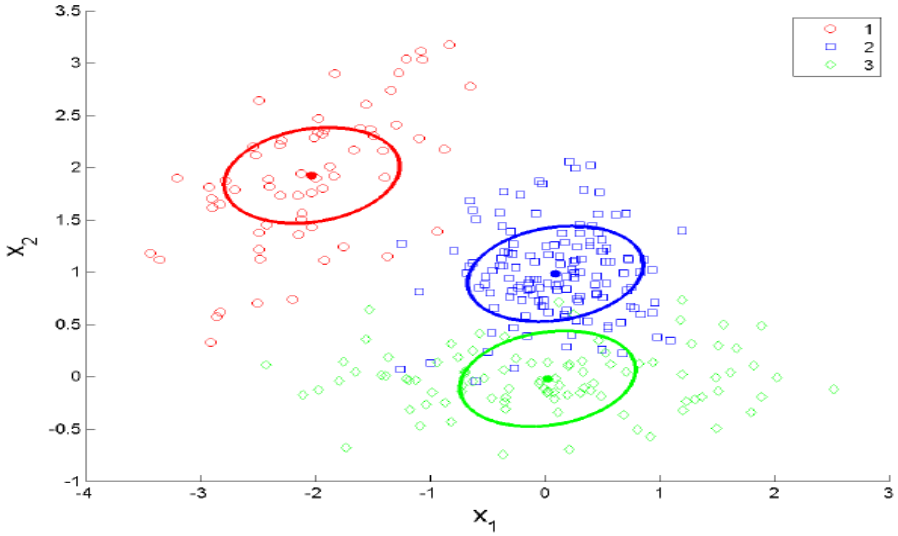
- La règle de classification devient

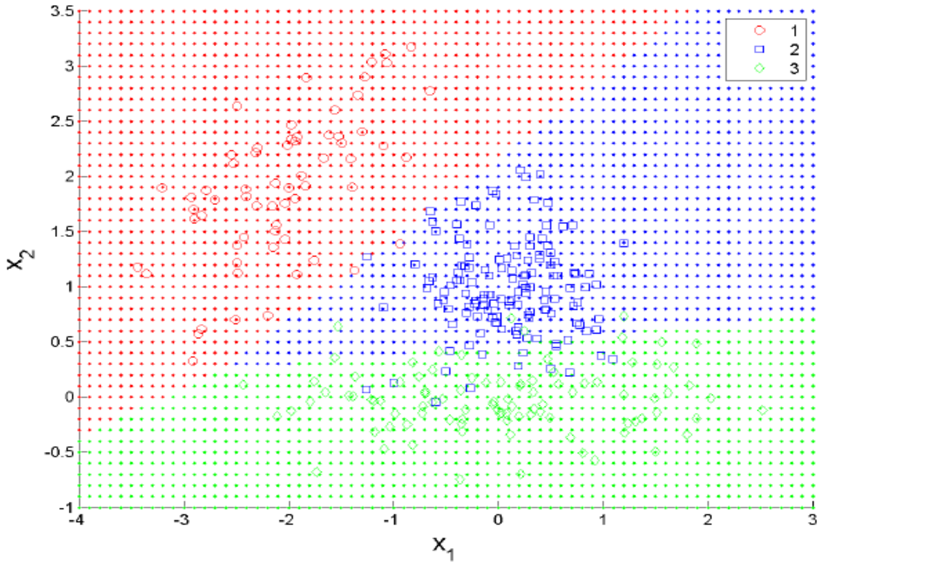
$$g(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \delta_\ell(x)$$

où

$$\delta_\ell(x) = x^T \widehat{\Sigma}^{-1} \widehat{\mu}_k - \frac{1}{2} \widehat{\mu}_k^T \widehat{\Sigma}^{-1} \widehat{\mu}_k + \log(\widehat{\pi}_k)$$

- δ_ℓ est alors appellée fonction discriminante linéaire.
- La frontière de décision entre deux classes k et ℓ est décrite par une équation linéaire en x $\{x : \delta_k(x) = \delta_\ell(x)\}$


3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-4
-3
-2
-1
0
1
2
3
x_1
x_2
1
2
3


1
2
3
x_1
x_2
-4
-3
-2
-1
0
1
2
3
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5

Les fonctions discriminantes δ_k permettent :

- de calculer un un score d'appartenance d'une entrée x à la classe k,
- de calculer les probabilités à posteriori avec :

$$\mathbb{P}(Y = k|X = x) = \frac{\exp \delta_k(x)}{\sum_{\ell=1}^{K} \exp \delta_\ell(x)}$$

Lorsque $\Sigma_k = \Sigma$ et $\widehat{\pi}_k = 1/K$ pour tout k :

- on fait de l'analyse discriminante linéaire avec probabilités à priori égales,
- les fonctions discriminantes calculent les distances de Mahalanobis (métrique $\widehat{\Sigma}^{-1}$) entre x et les centres de gravité $\widehat{\mu}_k$,
- on affecte x à la classe la plus proche,
- on parle de règle géométrique de classement de Mahalanobis-Fisher.

En analyse discriminante linéaire et $K = 2$ classes :

- le score de Fisher est une fonction linéaire qui s'écrit :

$$\begin{aligned}\Delta(x) &= \delta_1(x) - \delta_2(x)\\ &= x^T\widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2) - \frac{1}{2}(\widehat{\mu}_1 + \widehat{\mu}_2)'\widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2) + \log(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}).\end{aligned}$$

- la probabilité à posteriori d'appartenir à la classe 1 s'écrit comme une fonction logistique du score de Fisher :

$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(\Delta(x))}{1 + \exp(\Delta(x))}$$

- La règle de classification de Fisher consiste à comparer le score de Fisher à 0 pour prédire la classe de x.

Plan

On se place dans le cadre où les variables d"entrées $X = (X_1, \ldots, X_p)$ sont de type quelconque (quantitatif ou qualitatif) et $Y \in \{1, \ldots, K\}$.

- Hypothèse : indépendance des variables X_j dans chaque groupe k

$$f_k(x) = \prod_{j=1}^{p} f_{k,j}(x_j)$$

- L'approche indirect donne :

$$\begin{aligned} g(x) &= \underset{k \in \{1,\ldots,K\}}{\arg\max} \; \pi_k f_k(x) \\ &= \underset{k \in \{1,\ldots,K\}}{\arg\max} \; \pi_k \prod_{j=1}^{p} f_{k,j}(x_j) \end{aligned}$$

- Les paramètres π_k et les p densités en dimension 1 $f_{k,j}(x_j)$ sont estimés sur les données d'apprentissage.

Si la variable X_j est qualitative, on estime la probabilité $f_{k,j}(x) = \mathbb{P}(X_j = x | Y = k)$ par la fréquence empirique de la modalité x dans le groupe k.

Si la variable X_j est quantitative existe différentes approches pour estimer la densité $f_{k,j}$:

- on peut supposer une forme paramétrique pour $f_{k,j}(x)$. Par exemple

$$f_{k,j}(x) = \frac{1}{\sqrt{2\pi\sigma_{k,j}^2}} \exp\left[-\frac{1}{2\sigma_{k,j}^2}(x - \mu_{k,j})^2\right]$$

  où les estimateurs du maximum de vraissemblance de $\mu_{k,j}$ et $\sigma_{k,j}^2$ sont la moyenne et la variance empirique de la variable j dans le groupe k.

- $f_{k,j}(x)$ peut aussi être estimé de façon non paramétrique à l'aide d'un histogramme ou d'un estimateur de densité à noyau.

L'hypothèse d'indépendance des variables d'entrée dans les groupes est généralement fausse. Pourtant cette approche est très courante :

- car elle est simple, rapide et fonctionne pour une variable de sortie non binaire, et des variables d'entrées de type quelconque.
- elle permet de traiter des données de grande dimension.

Plan

On se place dans le cadre où les variables d'entrées peuvent être de type quelconque (quantitatif ou qualitatif) et $Y \in \{0, 1\}$.

- Après recodage des données qualitatives avec les indicatrices des modalités, les variables d'entrée sont toutes quantitatives ou binaires et on aura $X = (X_1, \dots, X_p) \in \mathbb{R}^p$.
- En régression logistique, on s'intéresse à la loi de $Y|X$ qui est une loi de Bernoulli de paramètre p avec :

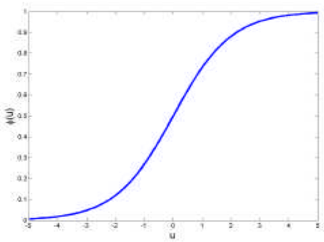
$$\mathbb{P}(Y = 1|X = x) = p$$
$$\mathbb{P}(Y = 0|X = x) = 1 - p$$

- On fait l'hypothèse que la probabilité $p = \mathbb{P}(Y = 1|X = x)$ est une fonction logistique d'un score linéaire

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_1 x_p \in \mathbb{R}$$

et la fonction logistique $f : \mathbb{R} \to [0, 1]$ est définie par :

$$f(u) = \frac{\exp(u)}{1 + \exp(u)}.$$



- On modélise donc la probabilité à posteriori "de succès" par :

$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}$$

- Le score linéaire est alors :

$$\beta_0 + \sum_{j=1}^{p} \beta_j x_j = f^{-1}(p) = \log \frac{p}{1-p}.$$

La fonction f^{-1} est appelée fonction logit avec :

$$logit(p) = log \frac{p}{1-p}.$$

- Paramètres inconnus estimés par maximum de vraisemblance :

$$\beta = (\beta_0, \ldots, \beta_p).$$

- Log-vraissemblance de l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$

$$\begin{aligned}
\ell(\beta) &= \log \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | X_i = x_i,) \\
&= \log \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}
\end{aligned}$$

avec

$$p_i = \mathbb{P}(Y_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j})}.$$

- L'estimateur $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ du maximum de vraissemblance n'a pas de forme explicite. Les logiciels utilisent donc des algorithmes d'optimisation pour estimer les paramètres $\beta_0, \ldots, \beta_p$ sur les données d'apprentissage.
- L'algorithme souvent utilisé est celui de Newton-Raphson qui est une méthode itérative de type gradient basée sur la relation suivante :

$$\beta^{(t)} = \beta^{(t-1)} - \left(\left. \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right|_{\beta^{(t-1)}} \right)^{-1} \left. \frac{\partial \ell(\beta)}{\partial \beta} \right|_{\beta^{(t-1)}}$$

- La règle de classification g affecte alors une nouvelle observation x à la classe 1 si

$$p_i = \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j)}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j)}$$

est supérieur à 0.5. Elle est affectée à la classe 0 sinon.