Apprentissage supervisé
Approches basées sur un modèle.

Marie Chavent

Université de Bordeaux

4 octobre 2019

Introduction

- On a vu dans le chapitre précédent que la règle de classification de Bayes s'écrit :

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\dots,K\}}{\arg\min} \sum_{k=1}^{K} C_{k\ell}\, \mathbb{P}(Y = k | X = x) \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \mathbb{P}(Y = \ell | X = x) \quad \text{(coût 0-1)} \end{aligned}$$

- Les approches basées sur un modèle consistent à apprendre la loi de Y sachant X pour en déduire ensuite la règle de classification g.
- Exemples : analyse discriminante linéaire, bayésien naïf, régression logistique, etc.

- L'approche directe consiste à apprendre directement la loi de Y sachant X. Par exemple en régression logistique :

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

où β est estimé à partir des données d'apprentissage.

- L'approche indirecte utilise la formule de Bayes

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

Il suffit alors d'apprendre la loi de X sachant Y et la loi de Y.
Par exemple en analyse discriminante $f(x|Y = k)$ est gaussienne et les paramètres sont estimés à partir des données d'apprentissage.

Plan

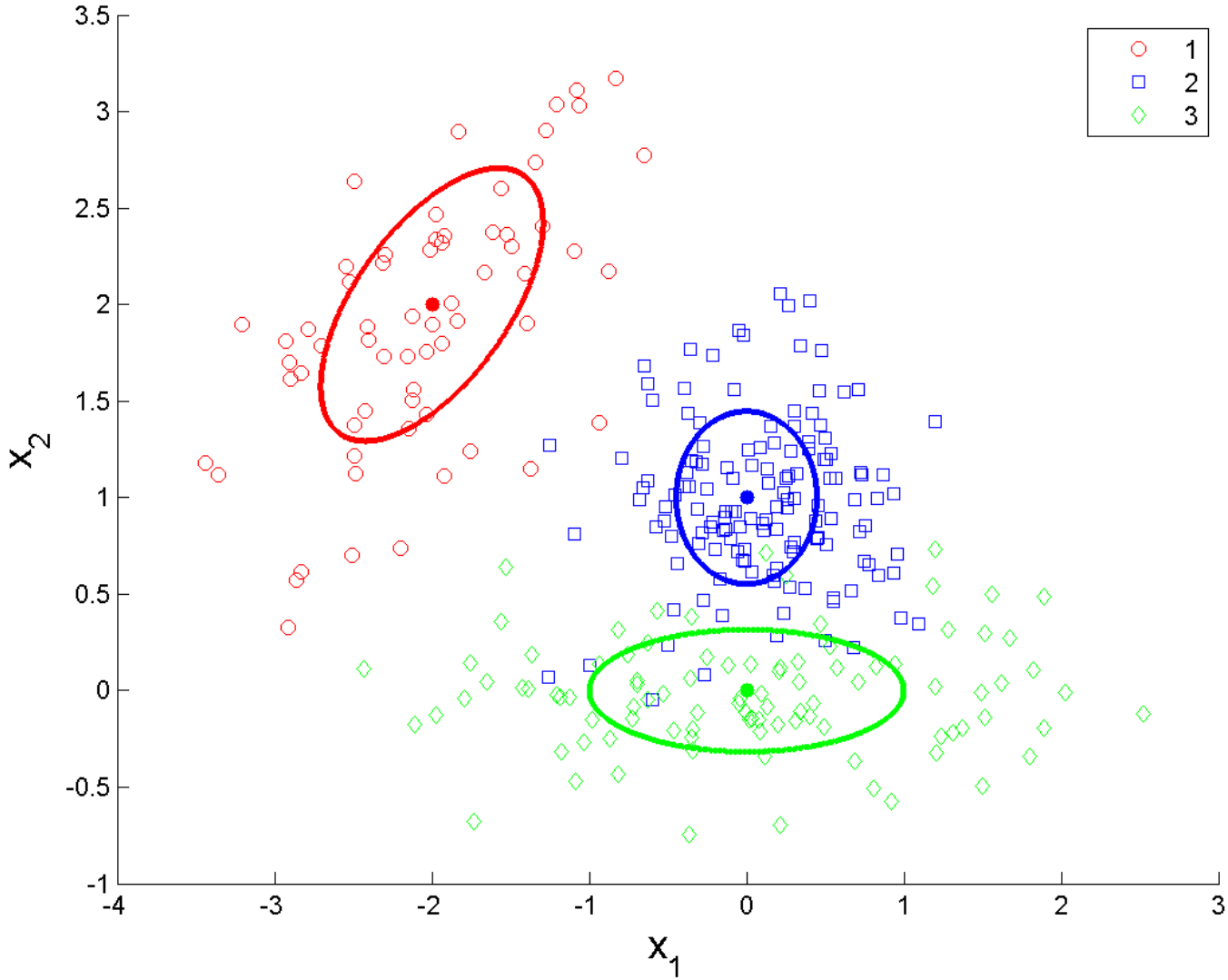
1. Analyse discriminante linéaire et quadratique.
2. Bayésien naif.
3. Régression logistique.

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 $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$:

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

- Paramètres inconnus $\{\mu_k, \Sigma_k\}$ et $\pi_k = \mathbb{P}(Y = k)$, pour $k = 1, \ldots, K$.


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
x2
x1

- Paramètres inconnus à estimer :

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

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

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

- Estimateurs du maximum de vraissemblance (à retrouver) :

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

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

- Règle de classification de Bayes (coût 0-1) :

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \mathbb{P}(Y=\ell|X=x) \text{ (approche directe)} \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ f(x|Y=\ell)\mathbb{P}(Y=\ell) \text{ (approche indirecte)} \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \log\left[f(x|Y=\ell)\right] + \log\left[\mathbb{P}(Y=\ell)\right] \end{aligned}$$

- Avec l'hypothèse paramétrique gaussienne on obtient (à montrer) :

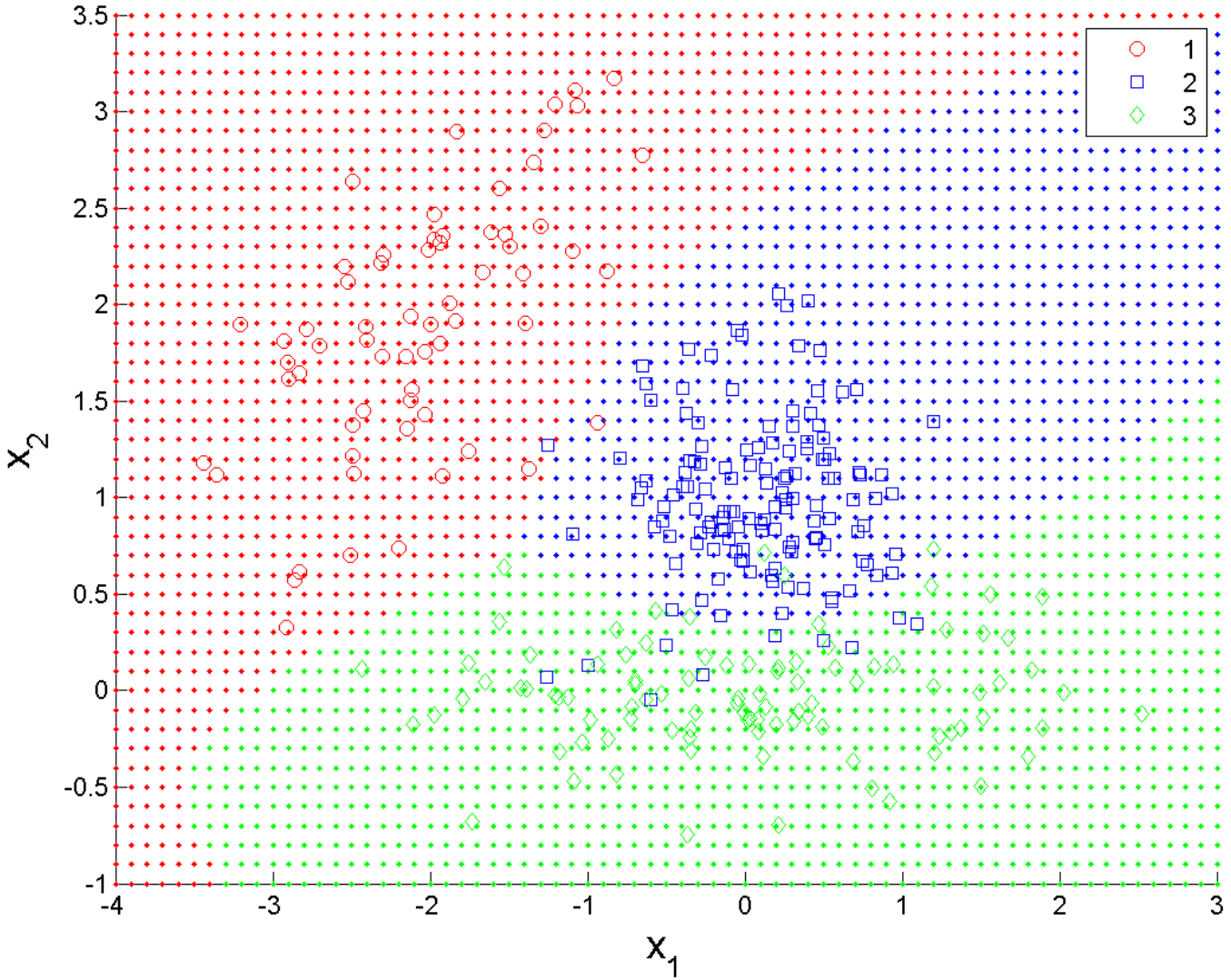
$$g(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ Q_\ell(x) \tag{1}$$

avec

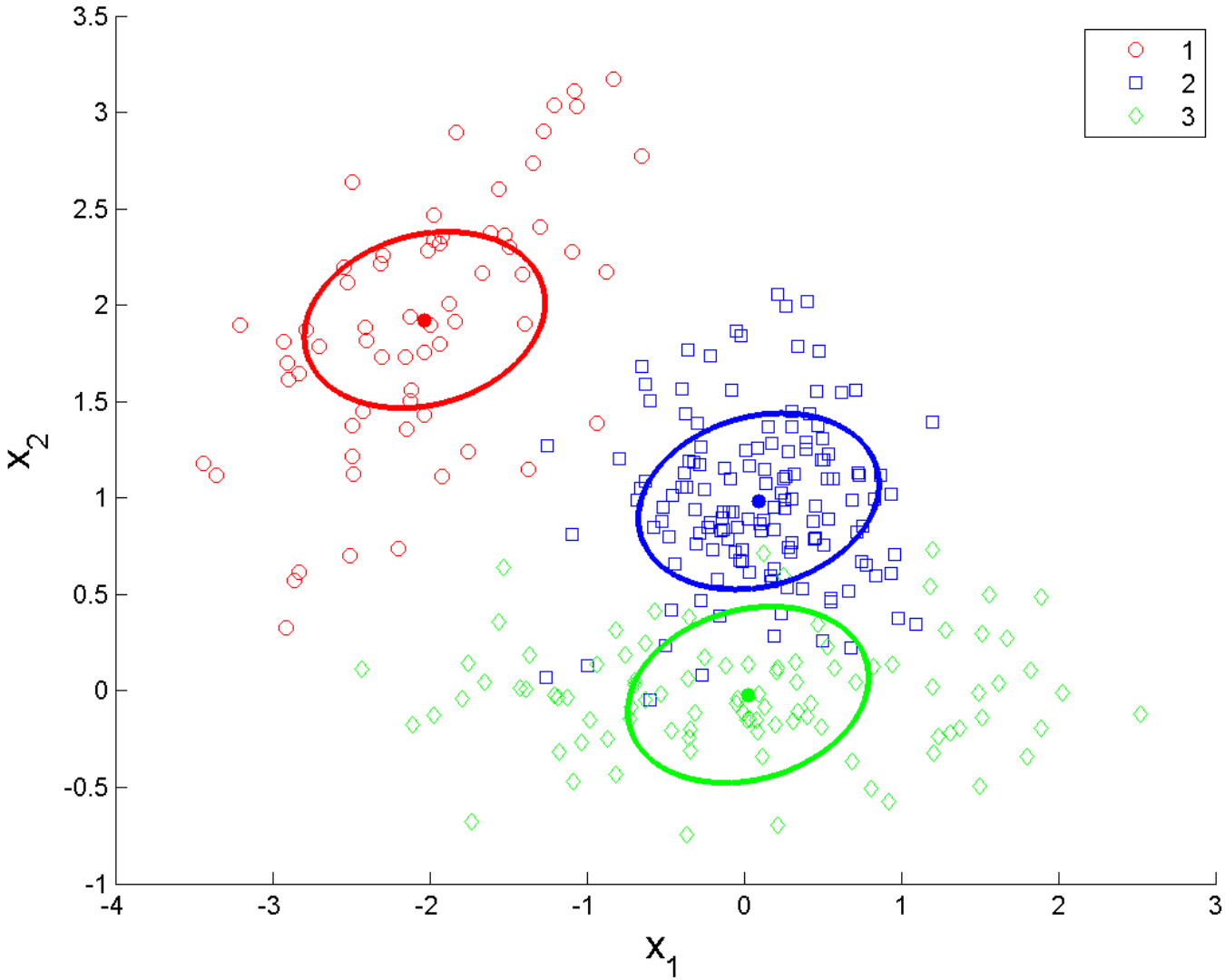
$$Q_\ell(x) = -\frac{1}{2}\log|\Sigma_\ell| - \frac{1}{2}(x-\mu_\ell)^T\Sigma_\ell^{-1}(x-\mu_\ell) + \log(\pi_\ell) \tag{2}$$

- Q_ℓ est appellée fonction discriminante quadratique.
- -2Q_ℓ est appellée dans SAS la distance de Mahalanobis généralisée entre x et μ_ℓ.

- La frontière de décision entre deux classes k et ℓ est décrite par une équation quadratique en x $\{x : Q_k(x) = Q_\ell(x)\}$



On suppose maintenant que $\Sigma_k = \Sigma$ pour tout $k = 1, \ldots, K$.



- Avec l'hypothèse d'égalité des matrices de covariance on obtient (à montrer) :

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

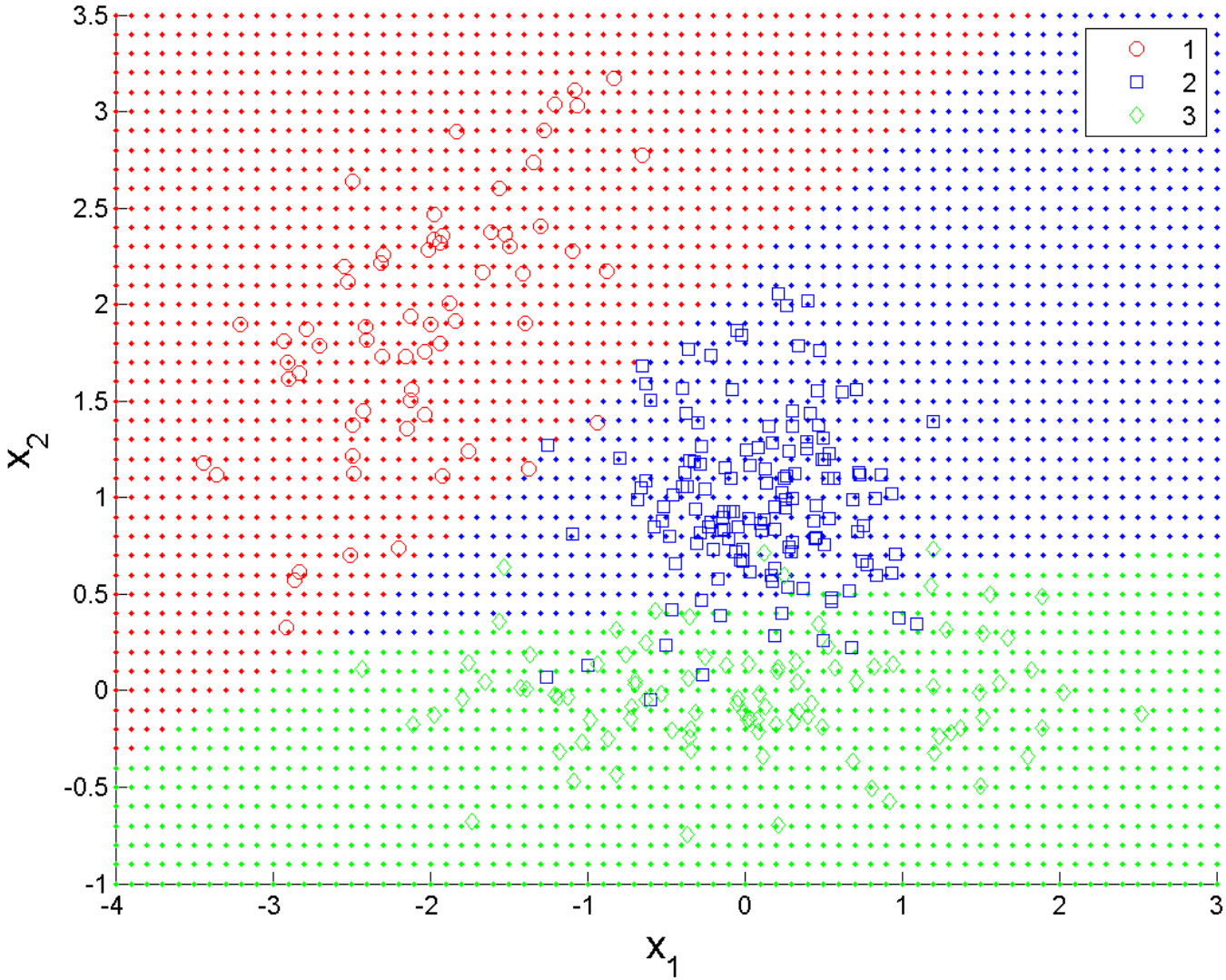
avec

$$L_\ell(x) = x^T \Sigma^{-1} \mu_\ell - \frac{1}{2} \mu_\ell^T \Sigma^{-1} \mu_\ell + \log(\pi_\ell) \tag{3}$$

- L_ℓ est alors appellée fonction discriminante linéaire.
- L'estimateur 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 frontière de décision entre deux classes k et ℓ est décrite par une équation linéaire en x $\{x : L_k(x) = L_\ell(x)\}$



Cas particulier de la classification binaire où $K = 2$.

- Le score de Fisher est définit par :

$$\begin{aligned}\Delta(x) &= L_1(x) - L_2(x) \\ &= x^T \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \log(\frac{\pi_1}{\pi_2}).\end{aligned}$$

  Ce score est une fonction linéaire en x.

- On affecte x à la classe 1 si $\Delta(x) \geq 0$, sinon on affecte x à la classe 2.

- La probabilité à posteriori d'appartenir à la classe 1 est une fonction logistique du score de Fisher :

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

On suppose maintenant que $\Sigma_k = \Sigma$ et que $\mathbb{P}(Y = k) = 1/K$ pour tout $k = 1, \ldots, K$.

- Avec cette hypothèse supplémentaire des probabilités à priori égales on obtient (à montrer) :

$$g(x) = \underset{\ell \in \{1,\ldots,K\}}{\arg\min} \; D_\ell(x)$$

avec

$$D_\ell(x) = (x - \mu_\ell)^T \Sigma^{-1} (x - \mu_\ell) \tag{4}$$

- $D_\ell(x)$ est le carré de la distance de Mahalanobis (métrique Σ^{-1}) entre x et le centre μ_ℓ de la classe ℓ.
- On affecte alors x à la classe la plus proche.
- On parle de règle géométrique de classement.

En résumé :

- QDA (Quadratic Discriminant Analysis) :

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

où Q_ℓ définie en (2)

- LDA (Linear Discriminant Analysis) :

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

où L_ℓ définie en (3)

- $Q_k(x)$ ou encore $L_k(x)$ mesurent un score d'appartenance aux classes,
- Les probabilités à posteriori des classes se calculent avec la formule :

$$\mathbb{P}(Y = k|X = x) = \frac{\exp\left(Q_k(x)\right)}{\sum_{\ell=1}^{K} \exp\left(Q_\ell(x)\right)} \text{ en QDA}$$

$$= \frac{\exp\left(L_k(x)\right)}{\sum_{\ell=1}^{K} \exp\left(L_\ell(x)\right)} \text{ en LDA}$$

Bayésien naïf

Les variables d'entrées $X = (X^1, \ldots, X^p)$ sont quantitatives ou qualitatives et $Y \in \{1, \ldots, K\}$.

- Hypothèse d'indépendance des variables $X^1, \ldots, X^p$ conditionnellement à Y :

$$f(x|Y = k) = \prod_{j=1}^{p} f_j(x_j|Y = k),$$

où $f_j(x_j|Y = k)$ est la notation utilisée ici pour désigner de manière unifiée :

  - la densité conditionnelle de X^j sachant $Y = k$ si X^j continue,
  - la probabilité conditionnelle de X^j sachant $Y = k$ si X^j discrète.

- L'approche indirect donne :

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\ldots,K\}}{\arg\max}\ \pi_\ell f(x|Y = \ell) \\ &= \underset{\ell \in \{1,\ldots,K\}}{\arg\max}\ \pi_\ell \prod_{j=1}^{p} f_j(x_j|Y = \ell). \end{aligned}$$

- Les K paramètres π_ℓ et les $p \times K$ lois conditionnelles $f_j(x_j|Y = k)$ sont à estimer sur les données d'apprentissage i.e. sur un échantillon de couples de variables aléatoires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y).

- Si la variable X^j est qualitative à valeurs dans $\mathcal{M}_j$, on estime les probabilités conditionnelles $f_j(x_j|Y=k)$ par les fréquences dans la classe k des modalités $x_j \in \mathcal{M}_j$:

$$\hat{f}_j(x_j|Y=k) = \frac{\sum_{i:y_i=k} \mathbb{1}_{X_i^j=x_j}}{n_k}.$$

- Si la variable X^j est quantitative à valeur dans $\mathbb{R}$:

  - On peut supposer une forme paramétrique pour $f_j(x_j|Y=k)$ et estimer les paramètres par maximum de vraissemblance. Par exemple

$$\hat{f}_j(x_j|Y=k) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{kj}^2}} \exp\left[-\frac{1}{2\hat{\sigma}_{kj}^2}(x-\hat{\mu}_{kj})^2\right]$$

    où $\hat{\mu}_{kj}$ est la moyenne empirique et $\hat{\sigma}_{kj}^2$ est la variance empirique corrigée de la variable X^j dans la classe k.

  - On peut aussi estimer $f_j(x_j|Y=k)$ de façon non paramétrique à l'aide d'un histogramme ou d'un estimateur de densité à noyau.

- L'hypothèse d'indépendance des variables $X^1, \ldots, X^p$ conditionnellement à Y 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.

Exercice : Un fournisseur d'élécticité veut prédire la demande de puissance éléctrique de ses clients à partir de trois variables binaires : Chauffage électrique ou non, Maison/Appartement, Sèche linge ou non. La variable de sortie a $K = 4$ modalités correspondant à $3, 6, 9$ et 12 kWh.
Comment construiriez-vous un classifieur Bayésien naïf ?

Régression logistique

Les variables d'entrées sont quantitatives ou qualitatives et $Y \in \{0, 1\}$.

- Les variables qualitatives sont recodées par les indicatrices des modalités et $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$ avec X^j quantitative ou binaire.
- En régression logistique, on s'intéresse à la loi de Y sachant X qui est une loi de Bernoulli de paramètre p avec :

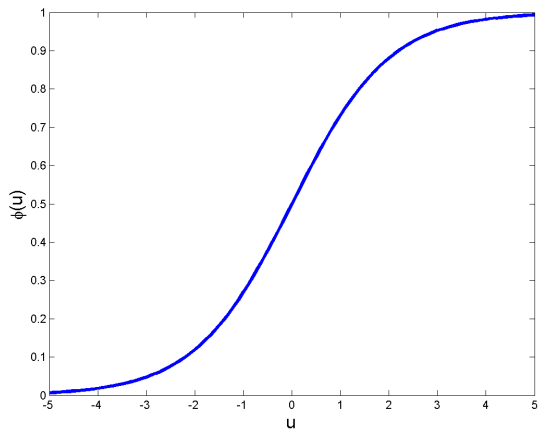
$$\mathbb{P}(Y = 1|X = x) = p$$
$$\mathbb{P}(Y = 0|X = x) = 1 - p$$

- 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 p 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 (conditionnelle) 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

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

- L'estimateur du maximum de vraissemblancede β 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 :

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

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

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

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

La régression logistique peut s'étendre au cas de classes multiples. On parle alors de régression logistique multinomiale.

- On a maintenant $Y \in \{1, \dots, K\}$ et on note $X = (1, X^1, \dots, X^p)$.
- Le modèle prend la forme

$$\log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = K|X = x)} = x^T \beta_1$$
$$\log \frac{\mathbb{P}(Y = 2|X = x)}{\mathbb{P}(Y = K|X = x)} = x^T \beta_2$$
$$\vdots$$
$$\log \frac{\mathbb{P}(Y = K-1|X = x)}{\mathbb{P}(Y = K|X = x)} = x^T \beta_{K-1}$$

  avec $\beta_1, \dots, \beta_{K-1}$ des vecteurs de $\mathbb{R}^{p+1}$.

- Les $K - 1$ vecteurs β_k sont estimés par maximum de vraisemblance sur les données d'apprentissage.

- Les probabilités à posteriori sont alors :

$$\mathbb{P}(Y = K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(x^T \beta_\ell)}$$

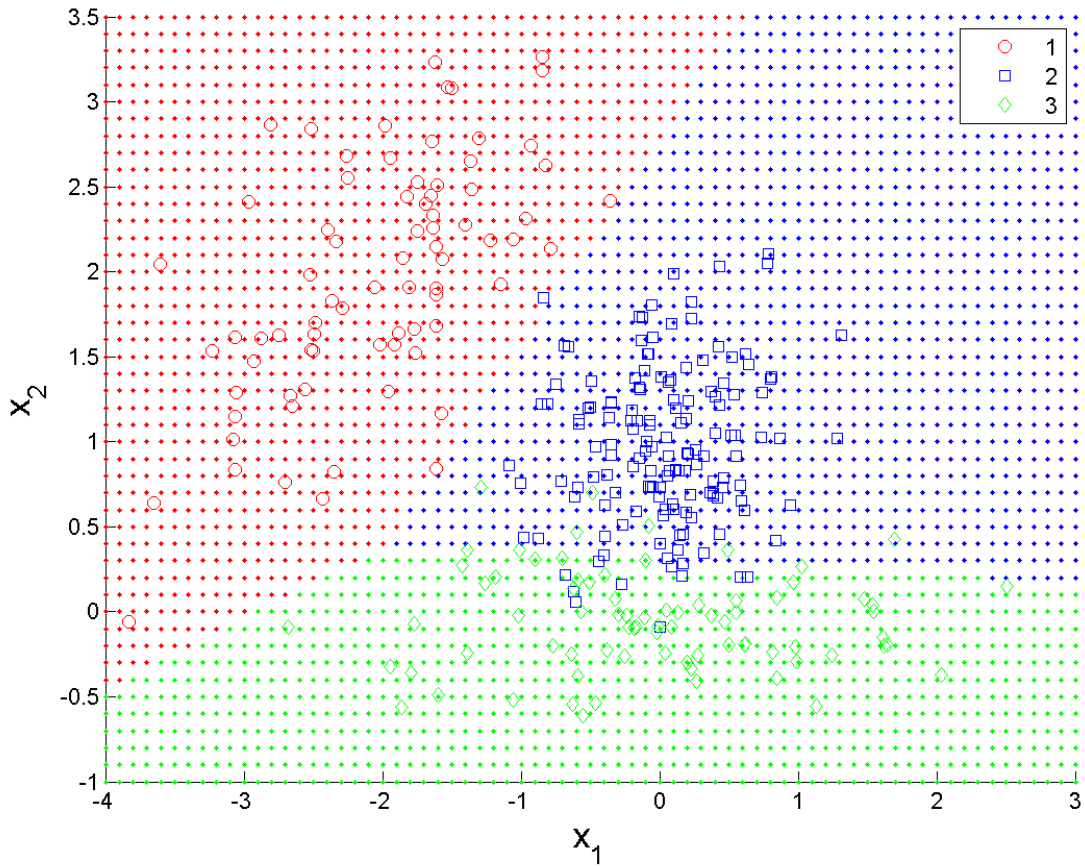
$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(x^T \beta_1)}{1 + \sum_{\ell=1}^{K-1} \exp(x^T \beta_\ell)}$$

$$\vdots$$

$$\mathbb{P}(Y = K - 1|X = x) = \frac{\exp(x^T \beta_{K-1})}{1 + \sum_{\ell=1}^{K-1} \exp(x^T \beta_\ell)}$$

- La règle de classification g affecte alors une nouvelle observation x à la classe la plus probable à posteriori.

Exemple : $p = 2$, $K = 3$ classes



Comparaison avec l'analyse discriminante linéaire.

Les deux modélisations induisent des écritures différentes de la vraissemblance et donc des estimations différentes des paramètres :

- en ADL, on maximise la loi jointe

$$\prod_{i=1}^{n} f_{X,Y}(x_i, y_i) = \underbrace{\prod_{i=1}^{n} f(x_i|Y = y_i)}_{\text{gaussien}} \underbrace{\prod_{i=1}^{n} \mathbb{P}(Y = y_i)}_{\text{Bernoulli}}$$

- en régression logistique on maximise directement la loi conditionnelle

$$\prod_{i=1}^{n} f_{X,Y}(x_i, y_i) = \underbrace{\prod_{i=1}^{n} \mathbb{P}(Y = y_i|X = x_i)}_{\text{logistique}} \underbrace{\prod_{i=1}^{n} f_X(x_i)}_{\text{inconnu}}$$

Exemple : $p = 2$, $K = 3$ classes.

Régression logistique (à gauche) versus LDA (à droite).