Correction du Partiel de Classification et Reconnaissance des Formes

du Lundi 9 décembre 2013, 8h00-10h00. December 28, 2013

Exercice 1 : Questions portant sur le cours

1) **Analyse linéaire discriminante** (3pts) : On considère un problème de classification à deux classes ω_1 et ω_2 avec la base d'apprentissage

$$\omega_1 \quad : \quad \mathbf{x}_1 = (1,1)^T, \mathbf{x}_2 = (1+\varepsilon, 1+\varepsilon)^T, \mathbf{x}_3 = (1+\varepsilon, 1-\varepsilon)^T, \mathbf{x}_4 = (1-\varepsilon, 1+\varepsilon)^T, \mathbf{x}_5 = (1-\varepsilon, 1-\varepsilon)^T$$

$$\omega_2 \quad : \quad \left\{ \begin{array}{l} \mathbf{x}_6 = (-1,-1)^T, \mathbf{x}_7 = (-1+\varepsilon, -1+\varepsilon)^T, \mathbf{x}_8 = (-1+\varepsilon, -1-\varepsilon)^T, \mathbf{x}_9 = (-1-\varepsilon, -1+\varepsilon)^T \\ \mathbf{x}_{10} = (-1-\varepsilon, -1-\varepsilon)^T \end{array} \right.$$

où $\varepsilon > 0$ est un réel positif de petite valeur. On note $\mathbf{S}_1$ et $\mathbf{S}_2$ les matrices de covariance intra-classe des données et $\mathbf{B}$ la matrice de covariance inter-classes..

- Déterminer les matrices $\mathbf{S}_1$,$\mathbf{S}_2$ et $\mathbf{B}$.
- Que peut-on dire des valeurs propres de la matrice $\mathbf{S}^{-1}\mathbf{B}$ avec $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$?
- Sans faire de calcul, indiquer l'axe le plus discriminant issu de la maximisation du critère de Fisher.

Réponse : d'après le cours, on a

$$\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

où $\mathbf{m}_1$ et $\mathbf{m}_2$ sont les moyennes des deux classes. On obtient

$$\mathbf{m}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{m}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \mathbf{m}_1 - \mathbf{m}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

et donc

$$\mathbf{B} = 4 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

De plus

$$\begin{aligned} \mathbf{S}_1 &= \sum_{i=1}^{5} (\mathbf{x}_1 - \mathbf{m}_1)(\mathbf{x}_1 - \mathbf{m}_1)^T \\ &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \varepsilon^2 & \varepsilon^2 \\ \varepsilon^2 & \varepsilon^2 \end{pmatrix} + \begin{pmatrix} \varepsilon^2 & -\varepsilon^2 \\ -\varepsilon^2 & \varepsilon^2 \end{pmatrix} + \begin{pmatrix} \varepsilon^2 & -\varepsilon^2 \\ -\varepsilon^2 & \varepsilon^2 \end{pmatrix} + \begin{pmatrix} \varepsilon^2 & \varepsilon^2 \\ \varepsilon^2 & \varepsilon^2 \end{pmatrix} \\ &= \begin{pmatrix} 4\varepsilon^2 & 0 \\ 0 & 4\varepsilon^2 \end{pmatrix} = 4\varepsilon^2 \mathbf{I}_2 \end{aligned}$$

De même

$$\mathbf{S}_2 = 4\varepsilon^2 \mathbf{I}_2$$

La matrice $\mathbf{S}^{-1}\mathbf{B}$ est facile à déterminer

$$\mathbf{S}^{-1}\mathbf{B} = \frac{1}{8\varepsilon^2}\mathbf{B} = \frac{1}{2\varepsilon^2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Cette matrice admet une valeur propre nulle (car on sait qu'il y a $K-1=1$ axe discriminant) et une valeur propre > 0 (des calculs donnent $\lambda = 0$ et $\lambda = 2$). En faisant un dessin, on remarque que l'axe discriminant est d'équation $y = x$.

2) **Classifieur Bayésien** (2pts) : On considère un problème de classification à deux classes équiprobables définies comme suit

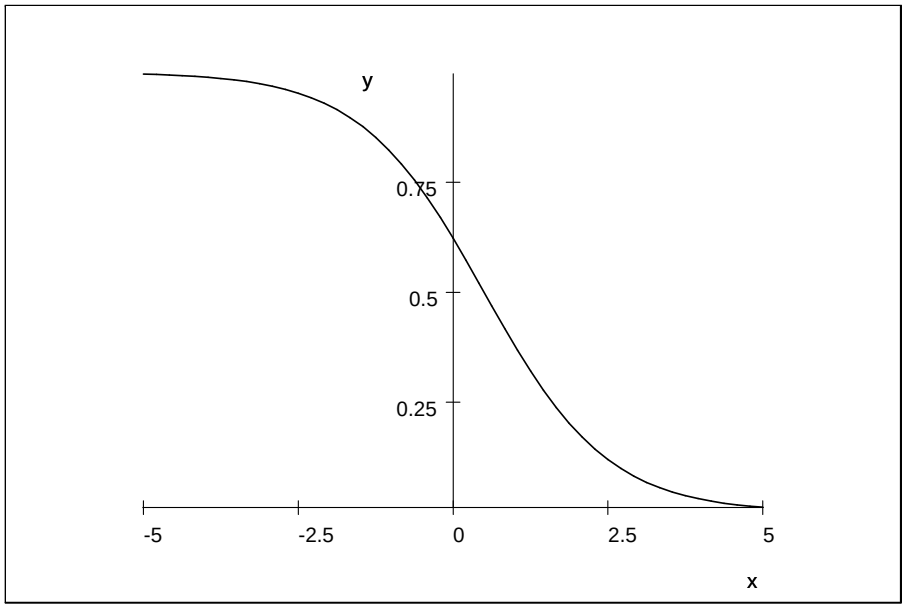
$$f(x\,|\omega_1) = \mathcal{N}(0,\sigma^2) \text{ et } f(x\,|\omega_2) = \mathcal{N}(1,\sigma^2).$$

Déterminer et représenter graphiquement $P(\omega_1\,|x)$ et $P(\omega_2\,|x)$. En déduire la règle de décision du classifieur Bayésien pour ce problème.

Réponse : on sait que

$$\begin{aligned} P(\omega_1\,|x) &= \frac{f(x\,|\omega_1)\,P(\omega_1)}{f(x)} \\ &= \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)+\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right)} \\ &= \frac{1}{1+\exp\left[-\frac{(x-1)^2}{2\sigma^2}+\frac{x^2}{2\sigma^2}\right]} \\ &= \frac{1}{1+\exp\left[\frac{2x-1}{2\sigma^2}\right]} \end{aligned}$$

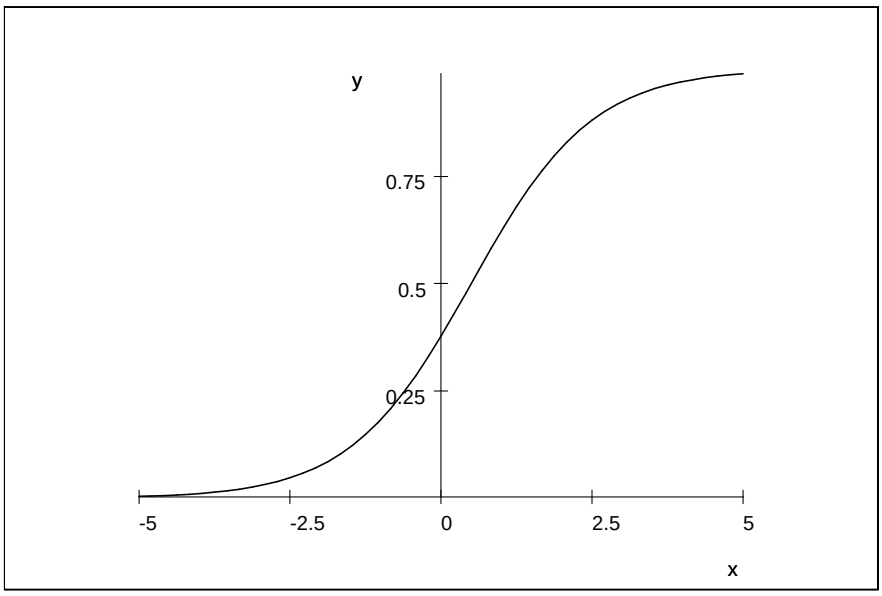
qui est représenté ci-dessous pour $\sigma^2 = 1$



De même, on obtient

$$
\begin{aligned}
P(\omega_2\,|x\,) &= \frac{f\,(x\,|\omega_2\,)\,P(\omega_2)}{f\,(x)} \\
&= \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right)}{\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)+\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right)} \\
&= \frac{1}{1+\exp\left[\frac{(x-1)^2}{2\sigma^2}-\frac{x^2}{2\sigma^2}\right]} \\
&= \frac{1}{1+\exp\left[\frac{1-2x}{2\sigma^2}\right]}
\end{aligned}
$$

qui est représenté ci-dessous pour $\sigma^2 = 1$



On remarque que

$$P(\omega_1\,|x\,) > P(\omega_2\,|x\,) \Longleftrightarrow x < \frac{1}{2}.$$

La règle de décision Bayésienne est donc

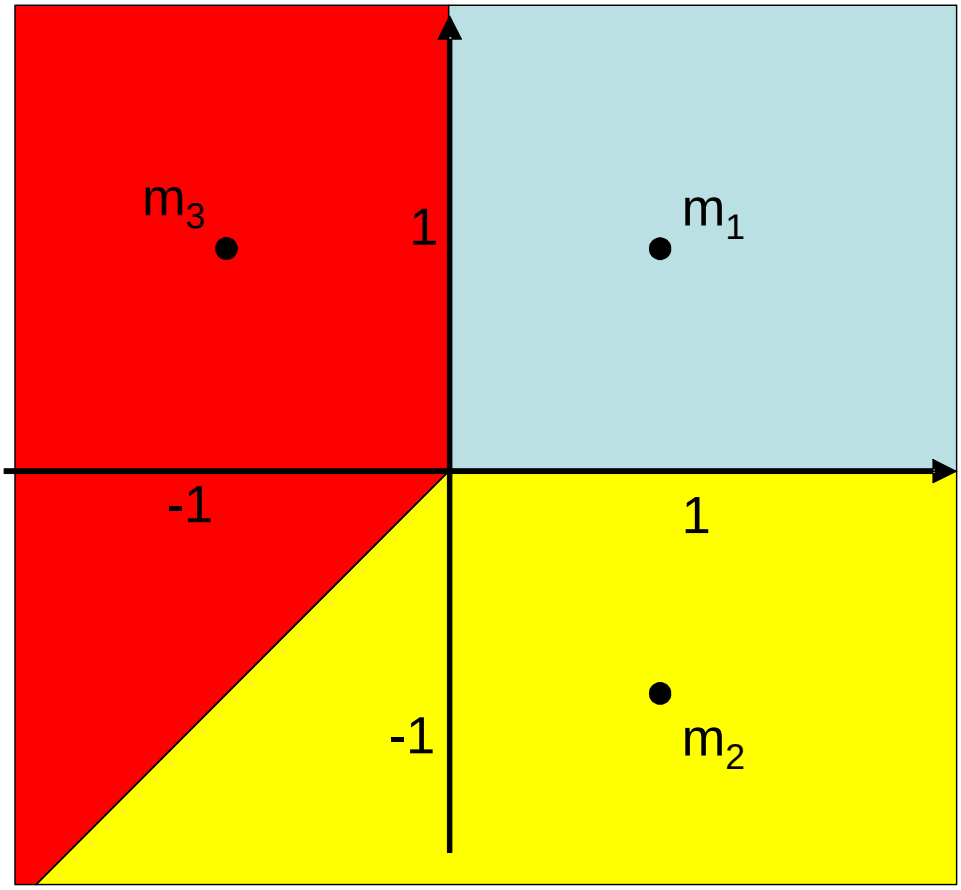
$$\text{Affecter } x \text{ à la classe } \omega_1 \text{ si } x < \frac{1}{2}.$$

3) **Classifieur Bayésien** (1pt) : On considère un problème de classification à trois classes équiprobables définies comme suit

$$f\,(\mathbf{x}\,|\omega_1\,) = \mathcal{N}(\mathbf{m}_1, \sigma^2\mathbf{I}), f\,(\mathbf{x}\,|\omega_2\,) = \mathcal{N}(\mathbf{m}_2, \sigma^2\mathbf{I}_2) \text{ et } f\,(\mathbf{x}\,|\omega_3\,) = \mathcal{N}(\mathbf{m}_3, \sigma^2\mathbf{I}_2)$$

où $\mathbf{m}_1 = (1,1)^T, \mathbf{m}_2 = (1,-1)^T$ et $\mathbf{m}_3 = (-1,1)^T$ et où $\mathbf{I}_2$ est la matrice identité de taille 2×2. Sans faire de calcul, représenter les régions du plan associées aux décisions du classifieur Bayésien $d^*(\mathbf{x}) = \omega_1$ (on affecte le vecteur $\mathbf{x}$ à la classe ω_1), $d^*(\mathbf{x}) = \omega_2$ (on affecte le vecteur $\mathbf{x}$ à la classe ω_2) et $d^*(\mathbf{x}) = \omega_3$ (on affecte le vecteur $\mathbf{x}$ à la classe ω_3).

Réponse : la règle de la distance aux barycentres conduit au résultat suivant



4) **Algorithme K-means** (2pts) : On considère un problème de classification avec la base d'apprentissage

$$\mathbf{x}_1 = (1,1)^T, \mathbf{x}_2 = (1,0)^T, \mathbf{x}_3 = (1,-1)^T, \mathbf{x}_4 = (-1,-1)^T, \mathbf{x}_5 = (-1,0)^T, \mathbf{x}_6 = (-1,1)^T.$$

On suppose que ce problème admet deux classes ω_1 et ω_2 et que ces classes admettent comme représentants $\mathbf{p}_1 = (-0.5,0)^T$ et $\mathbf{p}_2 = (0.5,0)^T$. Décrire les itérations de l'algorithme K-means initialisé avec les représentants $\mathbf{p}_1$ et $\mathbf{p}_2$ et donner la répartition finale des six points de la base d'apprentissage dans les deux classes ω_1 et ω_2. Faire de même avec $\mathbf{p}_1 = (0,-0.5)^T$ et $\mathbf{p}_2 = (0,1)^T$. Que peut-on en conclure ?

Réponse : l'algorithme K-means commence par affecter les points les plus proches de $\mathbf{p}_1$ à la classe ω_1 et les les points les plus proches de $\mathbf{p}_2$ à la classe ω_2. Après avoir fait un petit dessin, on obtient

$$C_1 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\} \text{ et } C_2 = \{\mathbf{x}_1, \mathbf{x}_6\}.$$

Ensuite, on calcule les centres de gravités des deux classes

$$\mathbf{g}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \mathbf{x}_5 \text{ et } \mathbf{g}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{x}_2.$$

et on applique la règle de la distance aux barycentres. Puisque les classes C_1 et C_2 restent inchangées, l'algorithme s'arrête.

Si on effectue le même travail avec $\mathbf{p}_1 = (0, -0.5)^T$ et $\mathbf{p}_2 = (0, 1)^T$, on obtient

$$C_1 = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\} \text{ et } C_2 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$$

puis

$$\mathbf{g}_1 = \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} = \mathbf{p}_1 \text{ et } \mathbf{g}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \mathbf{p}_2$$

et l'algorithme s'arrête. L'algorithme K-means est donc sensible à l'initialisation qui doit être effectuée correctement.

5) **Perceptron** (3pts) : on considère un problème de classification à deux classes ω_1 et ω_2 avec la base d'apprentissage constituée des $n = 7$ points suivants

$$\begin{aligned} \omega_1 &: x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 5 \\ \omega_2 &: x_5 = 8, x_6 = 9, x_7 = 11 \end{aligned}$$

- Quelle est la règle de décision issue de la méthode des machines à vecteurs supports ?
- On construit un perceptron à une couche possédant un seul neurone avec une fonction d'activation linéaire. Si l'entrée de ce neurone est x, alors la sortie de ce neurone est $y = ax + b$. De plus pour toute entrée x_i de la classe ω_1, la sortie désirée est $d_i = -1$ tandis que pour toute entrée x_i de la classe ω_2, la sortie désirée est $d_i = 1$. Quelle est la fonction de coût $E(a, b)$ à considérer pour le filtre de Wiener ? Montrer que lorsqu'on remplace les espérances mathématiques par des moyennes empiriques, cette fonction de coût est minimale pour

$$\begin{aligned} b &= \frac{1}{n}\sum_{i=1}^{n} d_i - \frac{a}{n}\sum_{i=1}^{n} x_i \\ a &= \frac{\frac{1}{n}\sum_{i=1}^{n} d_i x_i - \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \end{aligned}$$

Après application numérique, on trouve $a \simeq 0.153$ et $b \simeq -1.014$. Quelle est la règle de décision associée ?

Réponse : après avoir fait un petit dessin, on observe que les deux points de ω_1 et ω_2 les plus proches sont $x_4 = 5$ et $x_5 = 8$. Ces points sont les vecteurs supports et la règle de décision est

$$\text{Affecter } x \text{ à la classe } \omega_1 \text{ si } x < \frac{5+8}{2} = 6.5.$$

La fonction de coût associée à la détermination du filtre de Wiener est

$$E(a, b) = E\left[(d_i - ax_i - b)^2\right].$$

Lorsqu'on remplace l'espérance mathématique par une moyenne empirique, on obtient

$$\begin{aligned} E(a, b) &= \sum_{i=1}^{n} (d_i - ax_i - b)^2 \\ &= \sum_{i=1}^{4} (1 + ax_i + b)^2 + \sum_{i=5}^{7} (1 - ax_i - b)^2 \end{aligned}$$

Pour trouver le minimum de cette fonction, il suffit d'annuler les dérivées partielles de $E(a, b)$, soit

$$\frac{\partial E(a, b)}{\partial a} = \frac{\partial E(a, b)}{\partial b} = 0.$$

Des calculs élémentaires conduisent à

$$\begin{cases} \sum_{i=1}^{n} (d_i - ax_i - b) = 0 \\ \sum_{i=1}^{n} x_i (d_i - ax_i - b) = 0 \end{cases}$$

La première équation permet d'obtenir

$$b = \frac{1}{n} \sum_{i=1}^{n} d_i - \frac{a}{n} \sum_{i=1}^{n} x_i.$$

En remplaçant cette expression de b dans la seconde équation, on obtient

$$\sum_{i=1}^{n} x_i d_i - a \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} d_i - \frac{a}{n} \sum_{i=1}^{n} x_i \right) \sum_{i=1}^{n} x_i = 0$$

d'où

$$a = \frac{\frac{1}{n} \sum_{i=1}^{n} d_i x_i - \left(\frac{1}{n} \sum_{i=1}^{n} d_i \right) \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)}{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2}.$$

La règle de décision associée est

$$\text{Affecter } x \text{ à la classe } \omega_1 \text{ si } x < -\frac{b}{a} = 6.65.$$

Exercice 2 : **Questions portant sur l'article**

1) (2pts) Dans l'introduction de l'article les auteurs parlent de "standard linear and quadratic discriminant analysis". Expliquer le principe de ces règles de classification (vues en cours).

Réponse : ces deux règles correspondent au classifieur Bayésien dans le cas Gaussien. La règle linéaire correspond au cas de matrices de covariances identiques (i.e., $\Sigma_i = \Sigma$, $\forall i$) et s'écrit

$$\text{Affecter } x \text{ à la classe } \omega_i \text{ si } g_i(x) \leq g_j(x), \forall j$$

avec

$$g_i(x) = m_i^T \Sigma^{-1} m_i - 2x^T \Sigma^{-1} m_i + \ln |\Sigma| - 2 \ln P(\omega_i)$$

qui est une fonction affine de x. Le cas quadratique correspond au cas de matrices de covariance différentes et s'écrit

$$\text{Affecter } x \text{ à la classe } \omega_i \text{ si } g_i(x) \leq g_j(x), \forall j$$

avec

$$g_i(x) = (x - m_i)^T \Sigma_i^{-1} (x - m_i) + \ln |\Sigma_i| - 2 \ln P(\omega_i).$$

2) (1pt) Rappeler l'expression du classifieur Bayésien dans le cas de densités normales avec une fonction de coût $c_{ij} = 1 - \delta_{ij}$. N'y-t-il pas une erreur dans l'équation (3) de l'article.

Réponse : Avec une fonction de coût $c_{ij} = 1 - \delta_{ij}$, la règle de décision Bayésienne s'écrit

$$\text{Affecter } x \text{ à la classe } \omega_k \text{ si } P(\omega_k | x) \geq P(\omega_i | x), \forall i.$$

En utilisant le fait que

$$P\left(\omega_k \left| x \right.\right) = \frac{f\left(x \left| \omega_k \right.\right) P(\omega_k)}{f\left(x\right)}$$

et la forme de la densité Gaussienne, on obtient la règle de décision suivante

$$\text{Affecter } x \text{ à la classe } \omega_k \text{ si } g_k(x) \leq g_i(x), \forall i$$

avec

$$g_k(x) = (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) + \ln|\Sigma_k| - 2\ln \pi_k$$

Dans l'équation (3) de l'article, il y a donc une erreur typographique puisque $\underset{k=1,...,K}{\text{argmax}}$ devrait être $\underset{k=1,...,K}{\text{argmin}}$.

3) (2pts) Notons $\{\mathbf{x}_{ik}, i = 1, ..., N_k\}$ les vecteurs de la base d'apprentissage associés à la classe ω_k. Démontrer que l'estimateur du maximum de vraisemblance du vecteur moyenne $\boldsymbol{\mu}_k$ dans le cas Gaussien est

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{x}_{ik}.$$

Comment devrait-on procéder pour retrouver l'expression de $\widehat{\Sigma}_k$ donnée dans l'article (on demande juste d'expliquer la méthode) ?

Réponse : la vraisemblance des vecteurs $\{\mathbf{x}_{ik}, i = 1, ..., N_k\}$ s'écrit

$$f\left(\mathbf{x}_{ik}, ..., \mathbf{x}_{N_k k} \left| \boldsymbol{\mu}_k \right.\right) = \prod_{i=1}^{N_k} \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x_{ik}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(x_{ik}-\boldsymbol{\mu}_k)\right].$$

L'estimateur du maximum de vraisemblance de $\boldsymbol{\mu}_k$ est obtenu en annulant la dérivée de la log-vraisemblance par rapport à $\boldsymbol{\mu}_k$. On obtient

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln f\left(\mathbf{x}_{ik}, ..., \mathbf{x}_{N_k k} \left| \boldsymbol{\mu}_k \right.\right) = 0 \Longleftrightarrow \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{i=1}^{N_k} (x_{ik}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(x_{ik}-\boldsymbol{\mu}_k) = 0.$$

On sait que

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} \boldsymbol{\mu}_k^T \Sigma_k^{-1} x_{ik} &= \frac{\partial}{\partial \boldsymbol{\mu}_k} x_{ik}^T \Sigma_k^{-1} \boldsymbol{\mu}_k = \Sigma_k^{-1} x_{ik} \\ \frac{\partial}{\partial \boldsymbol{\mu}_k} \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k &= 2\Sigma_k^{-1} \boldsymbol{\mu}_k \end{aligned}$$

donc

$$-2\Sigma_k^{-1} \sum_{i=1}^{N_k} x_{ik} + 2N_k \Sigma_k^{-1} \boldsymbol{\mu}_k = 0 \Longleftrightarrow \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{x}_{ik}.$$

Pour déterminer l'estimateur du maximum de vraisemblance du vecteur $\boldsymbol{\mu}_k$ et de la matrice de covariance Σ_k, il faudrait chercher le vecteur $\boldsymbol{\mu}_k$ et la matrice Σ_k qui annulent les dérivées de la log-vraisemblance par rapport à $\boldsymbol{\mu}_k$ et Σ_k On obtient alors la même expression de $\boldsymbol{\mu}_k$ obtenue ci-dessus. La dérivée de la log-vraisemblance par rapport à Σ_k est moins triviale que celle par rapport à $\boldsymbol{\mu}_k$(mais pas très dure à calculer).

4) (1pt) Expliquer l'influence du paramètre h_{kj} de l'équation (5) sur l'estimateur $\widehat{f}_k(\mathbf{x})$.

Réponse : le paramètre h_{kj} joue le même rôle que le paramètre "bandwidth" intervenant dans les méthodes à noyaux. Quand ce paramètre est "grand", $\widehat{f}_k(\mathbf{x})$ est une somme de fonctions à variations lentes, donc c'est un estimateur de $f(x)$ avec peu de résolution. Par contre, lorsque ce paramètre est "petit", $\widehat{f}_k(\mathbf{x})$ est une somme de fonctions à variations rapides et donc est une estimation bruitée de $f(x)$. Il faut donc faire un compromis et choisir ce paramètre avec soin.

5) (2pts) Démontrer le résultat de l'équation (6).

Réponse : la matrice $\mathbf{H}_k$ rend les variables $Y_1, ..., Y_p$ indépendantes, donc la densité de $\mathbf{Y} = (Y_1, ..., Y_p)^T$ est

$$f(\mathbf{y}) = \prod_{i=1}^{p} f(y_i).$$

Puisque $\mathbf{Y} = \mathbf{H}_k\mathbf{X}$, en utilisant le théorème du changement de variables, on a

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \Longrightarrow \mathbf{Y} \sim \mathcal{N}\left(\mathbf{H}_k\boldsymbol{\mu}_k, \mathbf{H}_k\Sigma_k\mathbf{H}_k^T\right) = \mathcal{N}(\mathbf{H}_k\boldsymbol{\mu}_k, \mathbf{D}_k).$$

Puisque la matrice $\mathbf{D}_k$ est diagonale, i.e.,

$$\mathbf{D}_k = \text{diag}\left(\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_p^2}\right)$$

la densité de $\mathbf{Y}$ s'écrit

$$\begin{aligned} f(\mathbf{y}) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left[-\frac{(y_i - m_i)^2}{2\sigma_{ik}^2}\right] \\ &= \prod_{i=1}^{p} \frac{1}{\sigma_{ik}} \phi\left(\frac{y_i - m_i}{\sigma_{ik}}\right) \end{aligned}$$

où m_i est la moyenne de Y_i, c'est-à-dire la $i$ème composante de $\mathbf{H}_k\boldsymbol{\mu}_k$ notée dans l'article $(\mathbf{H}_k\boldsymbol{\mu}_k)_i$. Il suffit alors de faire le changement de variables $\mathbf{X} = \mathbf{H}_k^T\mathbf{Y}$ pour avoir la densité de $\mathbf{X}$. On obtient, en utilisant le fait que le déterminant d'une matrice orthogonale est $|\mathbf{H}_k| = 1$

$$f(\mathbf{x}) = \prod_{i=1}^{p} \frac{1}{\sigma_{ik}} \phi\left(\frac{(\mathbf{H}_k x)_i - (\mathbf{H}_k\boldsymbol{\mu}_k)_i}{\sigma_{ik}}\right).$$

6) (2pts) Expliquer en quoi les théorèmes 3.1 et 3.2 permettent d'estimer la matrice $\mathbf{M}$ à partir de données $\{\mathbf{x}_{ik}, i = 1, ..., N_k\}$ (expliquer le principe de l'estimation de $\mathbf{M}$).

Réponse : le théorème 3.1 indique que la matrice $\mathbf{M}$ peut être obtenue en cherchant le minimum de la fonction de contraste $J_G(\mathbf{Y_M})$. Par exemple, si on impose aux variances de $\mathbf{Y_M}$ d'être égales à 1 et si les lois marginales de $\mathbf{Y_M}$ ont un kurtosis non nul, d'après le théorème 3.1, on peut chercher la matrice en minimizant la fonction

$$J_G(\mathbf{Y_M}) = \sum_{i=1}^{p} \left[E\left(Y_i^4\right) - 3\right]^2$$

car le moment d'ordre 4 d'une variable aléatoire Gaussienne centrée réduite est $E(Z^4) = 3$. En pratique, on ne connaît pas $E(Y_i^4)$ et donc on doit approcher cette moyenne à l'aide des vecteurs d'apprentissage

$$E\left(Y_i^4\right) \simeq \frac{1}{N_k} \sum_{j=1}^{N_k} \mathbf{y}_{ijk}^4$$

où $\mathbf{y}_{ijk} = [\mathbf{M}\mathbf{x}_{jk}]_i$ est la $i$ème composante du vecteur $\mathbf{y}_{jk}$.

7) (1pt) Quelle est la principale motivation pour utiliser la méthode de classification proposée dans la section 4 ?

Réponse : Lorsqu'on applique l'analyse en composantes indépendantes, on transforme le vecteur $\mathbf{X}$ qui a des composantes dépendantes en un vecteur $\mathbf{Y}$ qui a des composantes indépendantes. L'apprentissage de la loi de $\mathbf{Y}$ se réduit à l'apprentissage de chacune de ses composantes qui est un problème univarié alors que l'apprentissage de la loi de $\mathbf{X}$ nécessite d'estimer une loi multivariée. Le fait d'estimer p lois univariées au lieu d'une loi multivariée rend l'apprentissage plus simple et plus performant.

8) Question Bonus (2pts) Pour mettre en oeuvre le classifieur Bayésien associé au premier exemple de la section 5.1, on doit connaitre la loi de $\mathbf{x} = (x_1, ..., x_{21})^T$ conditionnellement à chaque classe. Pour simplifier, on se limite à la loi d'une des composantes de $\mathbf{x}$, à savoir x_i. Montrer que pour la première classe, cette loi s'écrit

$$f(x_i|\, C_1) = \frac{F(v_1) - F(v_0)}{|h_1(i) - h_2(i)|}$$

où v_1 et v_0 sont deux quantités à déterminer dépendant de $x_i, h_1(i), h_2(i)$ et de la variance de ε_i notée σ_ε^2 et où F est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.

Réponse : Lorsque u est fixé, puisque ε_i suit une loi normale $\mathcal{N}(0, \sigma_\varepsilon^2)$, x_i suit également une loi normale de moyenne $uh_1(i) + (1-u)h_2(i)$ et de variance σ_ε^2, soit

$$f(x_i|\, u, C_1) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(x_i - uh_1(i) - (1-u)h_2(i))^2}{2\sigma_\varepsilon^2}\right].$$

En utilisant le fait que u suit une loi uniforme sur l'intervalle $[0,1]$, on obtient

$$\begin{aligned}
f(x_i|\, C_1) &= \int_0^1 f(x_i|\, u, C_1) du \\
&= \int_0^1 \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(x_i - uh_1(i) - (1-u)h_2(i))^2}{2\sigma_\varepsilon^2}\right] du \\
&= \int_0^1 \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(u[h_2(i) - h_1(i)] + x_i - h_2(i))^2}{2\sigma_\varepsilon^2}\right] du \\
&= \int_0^1 \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{\left(u - \frac{-x_i + h_2(i)}{h_2(i) - h_1(i)}\right)^2}{2\sigma_\varepsilon^2 / [h_2(i) - h_1(i)]^2}\right] du.
\end{aligned}$$

On fait ensuite le changement de variables

$$v = \frac{u - \frac{-x_i + h_2(i)}{[h_2(i) - h_1(i)]}}{\sqrt{\sigma_\varepsilon^2 / [h_2(i) - h_1(i)]^2}} = \frac{|h_1(i) - h_2(i)|}{\sigma_\varepsilon}\left[u - \frac{-x_i + h_2(i)}{h_2(i) - h_1(i)}\right]$$

ce qui permet d'aboutir au résultat suivant

$$f(x_i|\, C_1) = \int_{v_0}^{v_1} \frac{1}{\sqrt{2\pi}\,|h_1(i) - h_2(i)|} \exp\left(-\frac{v^2}{2}\right) dv$$

avec

$$\begin{aligned} v_0 &= \frac{|h_1(i) - h_2(i)|}{\sigma_\varepsilon} \left(\frac{x_i - h_2(i)}{h_2(i) - h_1(i)} \right) \\ v_1 &= \frac{|h_1(i) - h_2(i)|}{\sigma_\varepsilon} \left(1 - \frac{-x_i + h_2(i)}{h_2(i) - h_1(i)} \right). \end{aligned}$$

Si on utilise la fonction de répartition de la loi normale $\mathcal{N}(0,1)$, la densité s'exprime sous la forme

$$f(x_i | C_1) = \frac{F(v_1) - F(v_0)}{|h_1(i) - h_2(i)|}.$$