1) Règle de décision Bayésienne

La règle de décision Bayésienne associée au problème (1) consiste à accepter l'hypothèse ω_0 si

$$P(\omega_0|\mathbf{X}) > P(\omega_c|\mathbf{X}) \Longleftrightarrow f(\mathbf{X}|\omega_0)\,P(\omega_0) > f(\mathbf{X}|\omega_c)\,P(\omega_c)$$

c'est-à-dire

$$\frac{p_0}{(2\pi)^{n/2}\sqrt{|\mathbf{\Sigma}_0|}}\exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{M}_0)^T\mathbf{\Sigma}_0^{-1}(\mathbf{X}-\mathbf{M}_0)\right] > \frac{p_c}{(2\pi)^{n/2}\sqrt{|\mathbf{\Sigma}_c|}}\exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{M}_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M}_c)\right]$$

En prenant le logarithme de cette inégalité, on obtient

$$\ln\left[\frac{p_0}{p_c}\right] - \frac{1}{2}\ln\left[\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_c|}\right] - \frac{1}{2}(\mathbf{X}-\mathbf{M}_0)^T\mathbf{\Sigma}_0^{-1}(\mathbf{X}-\mathbf{M}_0) + \frac{1}{2}(\mathbf{X}-\mathbf{M}_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M}_c) > 0$$

soit

$$(\mathbf{X}-\mathbf{M}_c)^T\mathbf{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M}_c) - (\mathbf{X}-\mathbf{M}_0)^T\mathbf{\Sigma}_0^{-1}(\mathbf{X}-\mathbf{M}_0) > \ln\left[\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_c|}\right] - 2\ln\left[\frac{p_0}{p_c}\right]$$

On voit donc que

$$b = \ln\left[\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_c|}\right] - 2\ln\left[\frac{p_0}{p_c}\right]$$

- Dans le cas où les deux classes sont équiprobables et où les matrices de covariances sont les mêmes sous les deux hypothèses, i.e., $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_c = \mathbf{\Sigma}$, la règle de Bayes se réduit à accepter l'hypothèse ω_0 si

  $$(\mathbf{X}-\mathbf{M}_0)^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M}_0) < (\mathbf{X}-\mathbf{M}_c)^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M}_c) \Longleftrightarrow d(\mathbf{X},\mathbf{M}_0) < d(\mathbf{X},\mathbf{M}_c)$$

  où $d(\mathbf{X},\mathbf{Y}) = (\mathbf{X}-\mathbf{Y})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{Y})$ est la distance de Mahalanobis entre $\mathbf{X}$ et $\mathbf{Y}$.

- Dans le cas $n = 2$ et où les matrices de covariances sont égales à l'identité, la règle de décision s'écrit (voir cours)

  $$d^*(x) = \omega_0 \text{ si } \left[\mathbf{X} - \frac{1}{2}(\mathbf{M}_0+\mathbf{M}_c)\right]^T(\mathbf{M}_0-\mathbf{M}_c) \geq 0$$

  c'est-à-dire

  $$d^*(x) = \omega_0 \text{ si } x_1 + x_2 - 1 \geq 0$$

  Les zones d'acceptation des classes ω_0 et ω_c sont donc des demi-plans séparés par la droite d'équation $x_1 + x_2 - 1 = 0$.

  Si on appelle R_0 et R_c les zones d'acceptation des hypothèses ω_0 et ω_c, la probabilité d'erreur du classifieur est définie par

  $$P_e = P(\omega_0)\int_{R_c} f(\mathbf{X}|\omega_0)\,d\mathbf{X} + P(\omega_c)\int_{R_0} f(\mathbf{X}|\omega_c)\,d\mathbf{X}$$

  En utilisant la symmétrie du problème, il est clair que

  $$\int_{R_c} f(\mathbf{X}|\omega_0)\,d\mathbf{X} = \int_{R_0} f(\mathbf{X}|\omega_c)\,d\mathbf{X}$$

d'où

$$\begin{aligned}
P_e &= \int_{R_c} f\left(\mathbf{X}|\,\omega_0\right) d\mathbf{X} \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{1-x_1} \frac{1}{2\pi} \exp\left(-\frac{(x_1-1)^2 + (x_2-1)^2}{2} \right) dx_2 \right] dx_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1-1)^2}{2} \right) \left[\int_{-\infty}^{1-x_1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2-1)^2}{2} \right) dx_2 \right] dx_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1-1)^2}{2} \right) \left[\int_{-\infty}^{-x_1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2} \right) du \right] dx_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1-1)^2}{2} \right) F(-x_1) dx_1 \\
&= \int_{-\infty}^{\infty} p(v) F(-v-1) dv
\end{aligned}$$

où $p(v)$ est la densité de la loi normale $\mathcal{N}(0,1)$ et

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2} \right) du$$

est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. Si on était parti de l'autre intégrale, on aurait obtenu

$$\begin{aligned}
P_e &= \int_{R_0} f\left(\mathbf{X}|\,\omega_c\right) d\mathbf{X} \\
&= \int_{-\infty}^{\infty} \left[\int_{1-x_1}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2} \right) dx_2 \right] dx_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2} \right) \left[\int_{1-x_1}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2} \right) dx_2 \right] dx_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2} \right) \left[1 - F(1-x_1) \right] dx_1 \\
&= \int_{-\infty}^{\infty} p(v) \left[1 - F(1-v) \right] dv \\
&= \int_{-\infty}^{\infty} p(v) F(v-1) dv
\end{aligned}$$

En faisant le changement de variables $u = -v$, on retrouve le même résultat que ci-dessus.

2) **Apprentissage :** dans les applications pratiques, on peut estimer $\mathbf{M}_0, \mathbf{M}_c, \mathbf{\Sigma}_0$ et $\mathbf{\Sigma}_c$ à l'aide de leurs estimateurs du maximum de vraisemblance

$$\begin{aligned}
\widehat{\mathbf{M}}_0 &= \frac{1}{q_0} \sum_{i=1}^{q_0} \mathbf{X}_0^{(i)},\ \widehat{\mathbf{M}}_c = \frac{1}{q_c} \sum_{i=1}^{q_c} \mathbf{X}_c^{(i)} \\
\widehat{\mathbf{\Sigma}}_0 &= \frac{1}{q_0} \sum_{i=1}^{q_0} \left(\mathbf{X}_0^{(i)} - \widehat{\mathbf{M}}_0 \right) \left(\mathbf{X}_0^{(i)} - \widehat{\mathbf{M}}_0 \right)^T,\ \widehat{\mathbf{\Sigma}}_c = \frac{1}{q_c} \sum_{i=1}^{q_c} \left(\mathbf{X}_c^{(i)} - \widehat{\mathbf{M}}_c \right) \left(\mathbf{X}_c^{(i)} - \widehat{\mathbf{M}}_c \right)^T
\end{aligned}$$

3) **Analyse en composantes principales :** la matrice de covariance

$$\mathbf{\Sigma}_l = \frac{1}{q}\sum_{i=1}^{q}\left(\mathbf{X}_i - \mathbf{M}\right)\left(\mathbf{X}_i - \mathbf{M}\right)^T$$

peut s'écrire (en utilisant les notations du cours)

$$\begin{aligned}\mathbf{\Sigma}_l &= \frac{T}{q} = \frac{S}{q} + \frac{B}{q} \\ &= \frac{q_0}{q}\Sigma_0 + \frac{q_c}{q}\Sigma_c + \frac{q_0}{q}\left(\mathbf{M}_0 - \mathbf{M}\right)\left(\mathbf{M}_0 - \mathbf{M}\right)^T + \frac{q_c}{q}\left(\mathbf{M}_c - \mathbf{M}\right)\left(\mathbf{M}_c - \mathbf{M}\right)^T\end{aligned}$$

avec

$$\mathbf{\Sigma}_0 = \frac{1}{q_0}\sum_{\mathbf{X}_i\in\omega_0}\left(\mathbf{X}_i - \mathbf{M}_0\right)\left(\mathbf{X}_i - \mathbf{M}_0\right)^T \text{ et } \mathbf{\Sigma}_c = \frac{1}{q_c}\sum_{\mathbf{X}_i\in\omega_c}\left(\mathbf{X}_i - \mathbf{M}_c\right)\left(\mathbf{X}_i - \mathbf{M}_c\right)^T.$$

On en déduit

$$\mathbf{\Sigma}_m = \frac{q_0}{q}\left(\mathbf{M}_0 - \mathbf{M}\right)\left(\mathbf{M}_0 - \mathbf{M}\right)^T + \frac{q_c}{q}\left(\mathbf{M}_c - \mathbf{M}\right)\left(\mathbf{M}_c - \mathbf{M}\right)^T$$

4) **Fonctions radiales de base**

- Lorsque $\sigma^2 = 1$, $\mathbf{c}_1 = (0,0)$ et $\mathbf{c}_2 = (1,1)$, on a

$$\begin{pmatrix} f_1(\mathbf{X}) \\ f_2(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \exp\left[-\frac{x_1^2+x_2^2}{2}\right] \\ \exp\left[-\frac{(x_1-1)^2+(x_2-1)^2}{2}\right] \end{pmatrix}$$

  Donc

$$\begin{aligned}\begin{pmatrix} f_1[(0,0)] \\ f_2[(0,0)] \end{pmatrix} &= \begin{pmatrix} 1 \\ \exp[-1] \end{pmatrix}, \begin{pmatrix} f_1[(1,1)] \\ f_2[(1,1)] \end{pmatrix} = \begin{pmatrix} \exp[-1] \\ 1 \end{pmatrix} \\ \begin{pmatrix} f_1[(0,1)] \\ f_2[(0,1)] \end{pmatrix} &= \begin{pmatrix} f_1[(1,0)] \\ f_2[(1,0)] \end{pmatrix} = \begin{pmatrix} \exp[-1/2] \\ \exp[-1/2] \end{pmatrix}\end{aligned}$$

  En faisant un dessin, on peut remarquer que les deux images de la classe ω_0 et l'image des deux points de la classe ω_1 sont linéairement séparables. En remarquant que la pente de la droite séparatrice est -1, son équation s'écrit

$$f_1 + f_2 + w_0 = 0.$$

  Il suffit alors de choisir w_0 tel que $d\left[(0,0)^T\right] > 0$ et $d\left[(0,1)^T\right] < 0$, soit

$$1 + \frac{1}{e} + w_0 > 0 \text{ et } \frac{2}{\sqrt{e}} + w_0 < 0$$

  soit

$$-\left(1 + \frac{1}{e}\right) < w_0 < -\frac{2}{\sqrt{e}}.$$

- Dans le cas où on a deux centres $\mathbf{c}_1$ et $\mathbf{c}_2$, on a

$$d(\mathbf{X}) = w_0 + w_1 \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X} - \mathbf{c}_1\right\|^2\right) + w_2 \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{X} - \mathbf{c}_2\right\|^2\right)$$

– Le critère des moindres carrés

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]^2$$

est quadratique par rapport à $\mathbf{w}$ donc il admet un minimum global défini par

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0$$

En détaillant les trois équations, on obtient

$$\begin{aligned}
\sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)] &= 0 \\
\sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]\, f_1(\mathbf{X}_i) &= 0 \\
\sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]\, f_2(\mathbf{X}_i) &= 0
\end{aligned}$$

d'où

$$\begin{pmatrix} q & \sum_{i=1}^{q} f_1(\mathbf{X}_i) & \sum_{i=1}^{q} f_2(\mathbf{X}_i) \\ q & \sum_{i=1}^{q} f_1^2(\mathbf{X}_i) & \sum_{i=1}^{q} f_1(\mathbf{X}_i) f_2(\mathbf{X}_i) \\ q & \sum_{i=1}^{q} f_1(\mathbf{X}_i) f_2(\mathbf{X}_i) & \sum_{i=1}^{q} f_2^2(\mathbf{X}_i) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{q} e_i \\ \sum_{i=1}^{q} e_i f_1(\mathbf{X}_i) \\ \sum_{i=1}^{q} e_i f_2(\mathbf{X}_i) \end{pmatrix}$$

qui est un système linéaire qui permet de déterminer $\mathbf{w}$ dans la mesure où la matrice ci-dessus est inversible.

– L'algorithme de plus profonde descente est défini par

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \left.\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}(n)}$$

soit

$$\begin{aligned}
w_0(n+1) &= w_0(n) + \mu \sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)] \\
w_1(n+1) &= w_1(n) + \mu \sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]\, f_1(\mathbf{X}_i) \\
w_2(n+1) &= w_2(n) + \mu \sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]\, f_2(\mathbf{X}_i)
\end{aligned}$$

– L'algorithme LMS consiste à utiliser une règle de mise à jour basée sur l'erreur instantanée, soit

$$\begin{aligned}
w_0(n+1) &= w_0(n) + \mu\,[e_n - d(\mathbf{X}_n)] \\
w_1(n+1) &= w_1(n) + \mu\,[e_n - d(\mathbf{X}_n)]\, f_1(\mathbf{X}_n) \\
w_2(n+1) &= w_2(n) + \mu\,[e_n - d(\mathbf{X}_n)]\, f_2(\mathbf{X}_n)
\end{aligned}$$

où $\mathbf{X}_n$ est le vecteur de la base d'apprentissage (d'étiquette e_n) présenté à l'entrée du réseau pour la mise à jour de $\mathbf{w}(n)$.

- Lorsque le paramètre σ est inconnu, on peut essayer de l'estimer conjointement avec le vecteur $\mathbf{w}$. L'algorithme LMS consiste alors à ajouter aux relations précédentes une étape de mise à jour de σ basée sur la règle

$$\boldsymbol{\sigma}(n+1) = \boldsymbol{\sigma}(n) - \mu \left. \frac{\partial J(\mathbf{w}, \boldsymbol{\sigma})}{\partial \sigma} \right|_{\sigma=\sigma(n)}$$

On obtient alors

$$\begin{aligned} \boldsymbol{\sigma}(n+1) &= \boldsymbol{\sigma}(n) + \mu \left[e_n - d(\mathbf{X}_n)\right] \left. \frac{\partial d(\mathbf{X}_n)}{\partial \sigma} \right|_{\sigma=\sigma(n)} \\ &= \boldsymbol{\sigma}(n) + \frac{\mu}{\sigma^3(n)} \left[e_n - d(\mathbf{X}_n)\right] \left[\sum_{k=1}^{2} w_k(n) \left\|\mathbf{X}_n - \mathbf{c}_k\right\|^2 \exp\left[-\frac{\left\|\mathbf{X}_n - \mathbf{c}_k\right\|^2}{2\sigma^2(n)}\right]\right] \end{aligned}$$

5) **Questions portant sur l'article**

- Expliquer comment on peut obtenir la règle (4)

Réponse : les matrices de covariance Σ_0 et Σ_c peuvent être diagonalisées et s'écrivent alors

$$\Sigma_0 = \phi_0 \Lambda_0 \phi_0^T \text{ et } \Sigma_c = \phi_c \Lambda_c \phi_c^T$$

où Λ_0 et Λ_c sont des matrices diagonales contenant les valeurs propres λ_k^0 et λ_k^c pour $k = 1, ..., n$. On en déduit

$$\begin{aligned} (\mathbf{X} - \mathbf{M}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{X} - \mathbf{M}_c) &= (\mathbf{X} - \mathbf{M}_c)^T \phi_c \Lambda_c^{-1} \phi_c^T (\mathbf{X} - \mathbf{M}_c) \\ &= g^T \Lambda_c^{-1} g \end{aligned}$$

avec $g = \phi_c^T (\mathbf{X} - \mathbf{M}_c)$. Si on note $g = (g_1, ..., g_n)^T$ et $\Lambda_c = \text{diag}(\lambda_1^c, ..., \lambda_n^c)$, on en déduit

$$(\mathbf{X} - \mathbf{M}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{X} - \mathbf{M}_c) = \sum_{k=1}^{n} \frac{g_k^2}{\lambda_k^c}$$

De même

$$(\mathbf{X} - \mathbf{M}_0)^T \boldsymbol{\Sigma}_0^{-1} (\mathbf{X} - \mathbf{M}_0) = \sum_{k=1}^{n} \frac{h_k^2}{\lambda_k^0}$$

avec $h = \phi_0^T (\mathbf{X} - \mathbf{M}_0)$ et $\Lambda_c = \text{diag}(\lambda_1^0, ..., \lambda_n^0)$.

- Quelle est la motivation pour utiliser la matrice de covariance asymétrique donnée dans l'équation (6) de l'article ?

Réponse : la forme asymétrique de la matrice de covariance donnée dans (6) a pour objectif de compenser les erreurs d'estimation des matrices de covariance Σ_0 et Σ_c. Plus les erreurs d'estimation de la matrice Σ_c sont importantes, plus le coefficient α_c doit être important (tout en s'assurant que $\alpha_0 + \alpha_c = 1$).

- Montrer que les problèmes (P_1) et (P_2) admettent les mêmes vecteurs propres x avec des valeurs propres liées par une relation qu'on précisera

$$(P_1) : \widehat{\Sigma}_0 \Phi = \lambda \left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right) \Phi \text{ et } (P_2) : \widehat{\Sigma}_c \Phi = \mu \left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right) \Phi$$

Réponse : supposons que Φ vérifie (P_1), alors on a

$$\widehat{\Sigma}_0 (1-\lambda) \Phi = \lambda \widehat{\Sigma}_c \Phi$$

soit

$$\begin{aligned} (1-\lambda) \left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right) \Phi &= \lambda \widehat{\Sigma}_c \Phi + (1-\lambda) \widehat{\Sigma}_c \Phi \\ &= \widehat{\Sigma}_c \Phi \end{aligned}$$

donc Φ vérifie (P_2) avec $\mu = 1 - \lambda$ et inversement. On notera que puisque $\widehat{\Sigma}_0 \Phi = \lambda \left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right) \Phi$, on a

$$\lambda = \frac{\Phi^T \widehat{\Sigma}_0 \Phi}{\Phi^T \left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right) \Phi}$$

qui est un élément de $[0, 1]$ car les matrices $\widehat{\Sigma}_0$ et $\widehat{\Sigma}_c$ sont positives.

- Après l'équation (12), expliquer "in the APCA subspace".

  Réponse : la technique de prétraitement proposée dans cet article consiste tout d'abord à projeter les données sur les vecteurs propres associés aux m valeurs propres les plus grandes du problème (8). On se place ainsi dans le sous espace engendré par ces m vecteurs propres, d'où le terme "in the APCA subspace". Après avoir effectué cette projection, on a un vecteur X de taille réduite m défini par $\widehat{\Phi}^T X$ (notons que X est un vecteur de taille n et $\widehat{\Phi}$ est une matrice de taille $n \times m$).

- Après (13), les auteurs définissent une matrice $U = \widehat{\Phi}\widetilde{\Phi}$. Quelle est la différence entre $\widehat{\Phi}$ et $\widetilde{\Phi}$?

  Réponse : on va tout d'abord projeter X sur m vecteurs propres de l'APCA rangés dans $\widehat{\Phi}$. Le vecteur obtenu de taille m va être projeté sur les d vecteurs propres associés aux valeurs propres les plus grandes du problème (12). Ces d vecteurs propres sont rangés dans la matrice $\widetilde{\Phi}$ de taille $m \times d$. Après avoir effectué l'opération $U^T X = \widetilde{\Phi}^T \widehat{\Phi}^T X$, on obtient un vecteur de taille d.

- Expliquer les acronymes PCA, APCA, PLCDA et APCDA en précisant le type d'analyse effectué pour chacune de ces méthodes.

  Réponse :

  - PCA signifie "Principal Component Analysis". Il s'agit d'effectuer une analyse en composantes principales qui consiste à projeter le vecteur X sur les vecteurs propres associés aux valeurs propres les plus grande de la matrice de covariance des données X de la base d'apprentissage
  - APCA signifie "Asymmetric Principal Component Analysis" : au lieu de travailler avec la matrice de covariance

    $$\Sigma_l = p_0 \Sigma_0 + p_c \Sigma_c + \Sigma_m$$

    on travaille avec la matrice de covariance

    $$\Sigma_\alpha = \alpha_0 \Sigma_0 + \alpha_c \Sigma_c + \Sigma_m$$

    qui donne moins d'importance à la matrice la moins bien estimée
  - PLCDA signifie "PCA + LDA + CDA" : on fait une ACP symétrique définie par la recherche des valeurs propres et vecteurs propres de (2), puis une analyse discriminante linéaire définie par (10) et enfin une analyse discriminante de covariance dont on parle juste après (10).

- APCDA signifie “Asymmetric Principal Component Discriminant Analysis” : il s’agit de la méthode proposée dans cet article qui consiste à faire une ACP non symétrique définie en (8) suivie d’une analyse discriminante non symétrique définie dans (12)

- Dans la partie 4.3, de quoi est constitué le vecteur de données X que l’on veut classifier et quelle est sa dimension ?

  Réponse : le vecteur X est constitué de tous les pixels de l’image. C’est donc un vecteur de taille $n = 400$.