Examen Classification - 18 décembre 2007

1) Règle de décision Bayésienne

- Dans le cas de classes équiprobables, la règle de décision Bayésienne s'écrit

$$d^*(z) = \omega_1 \iff P(\omega_1|z) = \frac{f(z|\omega_1)P(\omega_1)}{f(z)} \geq P(\omega_2|z)$$

$$\underline{d^*(z) = \omega_1 \iff f(z|\omega_1) \geq f(z|\omega_2)}$$

Les densités de z conditionnellement à ω_1 et ω_2 s'écrivent

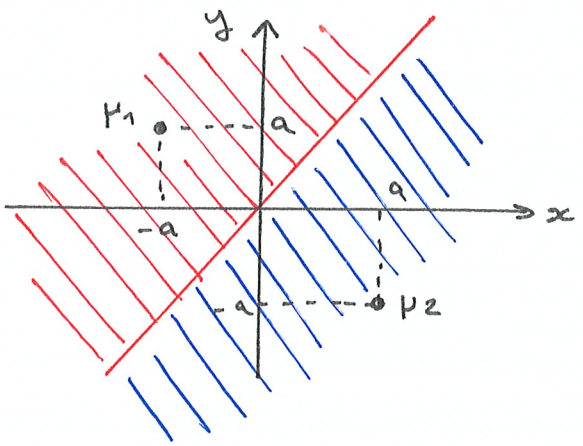
$$f(z|\omega_1) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}\|z-\mu_1\|^2\right]$$

$$f(z|\omega_2) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}\|z-\mu_2\|^2\right]$$

d'où

$$\underline{d^*(z) = \omega_1 \iff \|z-\mu_1\| \leq \|z-\mu_2\|}$$

qui est la règle classique de la distance aux barycentres. Dans le cas où $\mu_1 = \begin{pmatrix} -a \\ a \end{pmatrix}$ et $\mu_2 = \begin{pmatrix} a \\ -a \end{pmatrix}$ avec $a > 0$, on obtient les régions de décision suivantes :



$d^*(z) = \omega_1$

$d^*(z) = \omega_2$

- La probabilité d'erreur associée à la règle d^* ci dessus est définie par

$$P_e = P_1 + P_2 = \int_{R_2} f(z|\omega_1) P(\omega_1)\, dz + \int_{R_1} f(z|\omega_2) P(\omega_2)\, dz$$

où R_1 et R_2 sont les régions du plan associées aux décisions $d^*(z) = \omega_1$ et $d^*(z) = \omega_2$.
Les probabilité P_1 et P_2 se calculent comme suit

$$P_1 = \frac{1}{2}\int_{\mathbb{R}^2} \left[\int_{-\infty}^{x} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(x+a)^2 - \frac{1}{2\sigma^2}(y-a)^2\right] dy \right] dx$$

$$= \frac{1}{2}\int_{\mathbb{R}^2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x+a)^2\right] \left[\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-a)^2\right] dy \right] dx$$

d'où $P_1 = \int_{\mathbb{R}} \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x+a)^2\right] \left[\int_{-\infty}^{\frac{x-a}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right] dx$

$\left[\int_{-\infty}^{\frac{x-a}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \right] = \Phi\left(\frac{x-a}{\sigma}\right)$

↑ Chgt de variables $u = \frac{y-a}{\sigma}$

En faisant le changement de variables $u = \frac{x+a}{\sigma}$, on obtient alors

$$\boxed{P_1 = \frac{1}{2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \Phi\left(u - \frac{2a}{\sigma}\right) du}$$

Par symétrie, le calcul de P_2 conduit au même résultat, d'où

$$\boxed{P_e = \int_{\mathbb{R}} f(u) \Phi\left(u - \frac{2a}{\sigma}\right) du \triangleq P_e(a,\sigma)}$$

Quand $a \to +\infty$ ou $\sigma \to 0$ on a $-\frac{2a}{\sigma} \to -\infty$ d'où $\boxed{P_e \to 0}$

- Lorsque les coûts c_{ij} sont définis par $c_{11} = c_{22} = 0$ et $c_{12} = 2c_{21} = 2$ la règle de décision Bayésienne s'écrit

$$d^*(x) = \omega_1 \iff \frac{f(x|\omega_1)}{f(x|\omega_2)} \geq \frac{c_{12} - c_{22}}{c_{21} - c_{11}} \frac{P(\omega_2)}{P(\omega_1)} = 2$$

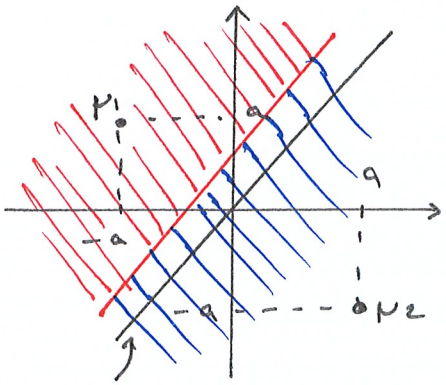
$$\boxed{d^*(x) = \omega_1 \iff f(x|\omega_1) \geq 2 f(x|\omega_2)}$$

c_{12} est le coût de prendre la décision a_1 alors que x est élément de ω_2 - Comme ce coût est plus élevé que c_{21}, on va prendre la décision a_1 moins souvent que dans le cas précédent où $d^*(x) = \omega_1 \iff f(x|\omega_1) \geq f(x|\omega_2)$

On a alors après quelques calculs élémentaires

$$\boxed{d^*(x) = \omega_1 \iff \|x - \mu_1\|^2 \leq \|x - \mu_2\|^2 - 2\sigma^2 \ln 2}$$

Par rapport à la question précédente, on a déplacé la droite de séparation entre les deux régions, comme illustré sur la figure ci-dessous



$d^*(z) = \omega_1$

$d^*(z) = \omega_2$

frontière pour la question précédente

2) Apprentissage Bayésien

On désire estimer μ_1 et σ_1^2 à l'aide des vecteurs $v_1, \dots, v_{n_1}$ et de manière similaire estimer μ_2 et σ_2^2 à l'aide des vecteurs $w_1, \dots, w_{n_2}$.
La vraisemblance associée aux données $v_1, \dots, v_{n_1}$ s'écrit en supposant l'indépendance entre ces différents vecteurs:

$$f(v ; \mu_1, \sigma_1^2) = f(v_1, \dots, v_{n_1} ; \mu_1, \sigma_1^2)$$

$$= \prod_{i=1}^{n_1} \frac{1}{(2\pi\sigma_1^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma_1^2}\|z_i - \mu_1\|^2\right]$$

$$= \frac{1}{(2\pi\sigma_1^2)^{n_1}} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}\|z_i - \mu_1\|^2\right]$$

on en déduit la log-vraisemblance

$$\ln f(v ; \mu_1 ; \sigma_1^2) = -n_1 \ln(2\pi) - n_1 \ln(\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}\|z_i - \mu_1\|^2$$

que l'on peut dériver par rapport à μ_1 et σ_1^2 pour obtenir

$$\frac{\partial \ln f}{\partial \mu_1} = 0 \Rightarrow \boxed{\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} z_i}$$

$$\frac{\partial \ln f}{\partial \sigma_1^2} = 0 \Rightarrow -\frac{n_1}{\sigma_1^2} + \frac{1}{2\sigma_1^4}\sum_{i=1}^{n_1}\|z_i - \mu_1\|^2 = 0 \Rightarrow \boxed{\hat{\sigma}_1^2 = \frac{1}{2n_1}\sum_{i=1}^{n_1}\|z_i - \mu_1\|^2}$$

La règle de classification qui découle de cette procédure d'apprentissage s'écrit

$$d^*(z) = \omega_1 \Leftrightarrow \hat{f}(z|\omega_1) \geq \hat{f}(z|\omega_2)$$

$$\Leftrightarrow \frac{1}{\hat{\sigma}_1^2}\exp\left[-\frac{1}{2\hat{\sigma}_1^2}\|z - \hat{\mu}_1\|^2\right] \geq \frac{1}{\hat{\sigma}_2^2}\exp\left[-\frac{1}{2\hat{\sigma}_2^2}\|z - \hat{\mu}_2\|^2\right]$$

$$\Leftrightarrow \ln\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right) - \frac{1}{2\hat{\sigma}_1^2}\|z - \hat{\mu}_1\|^2 \geq -\frac{1}{2\hat{\sigma}_2^2}\|z - \hat{\mu}_2\|^2$$

c'est-à-dire

$$d^*(z) = w_1 \iff \|z-\hat{\mu}_1\|^2 \leq \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}\|z-\hat{\mu}_2\|^2 + 2\hat{\sigma}_1^2 \ln\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)$$

- Dans le cas de données multidimensionnelles, on peut estimer une densité à l'aide de la méthode des fenêtres de Parzen ou la méthode des k plus proches voisins. Considérons par exemple la méthode des fenêtres de Parzen. Un estimateur de densité s'écrit pour $x \in \mathbb{R}^p$

$$\hat{f}(x) = \frac{1}{nh^p}\sum_{i=1}^{n} \phi\left(\frac{z-z_i}{h}\right)$$

où z_i est un vecteur de la base d'apprentissage de densité f, ϕ est un noyau choisi a priori (par exemple noyau Gaussien) et h est un paramètre (appelé parfois bande passante du noyau) qu'il faut régler de façon à avoir une estimation peu bruitée avec une résolution suffisante.

On parle d'estimation de densité non paramétrique car ces méthodes ne supposent pas comme que la loi est définie par un vecteur paramètre θ que l'on cherche à estimer comme pour la méthode d'apprentissage basée sur l'estimateur du maximum de vraisemblance.

3) Perceptron Multi-couches

- Cherchons tout d'abord le vecteur $w_1 = (w_{11}, w_{21}, a)^T$ qui minimise

$$E[e_1^2(n)] = E[(d_1(n) - y_1(n))^2]$$

Puisque $f(t) = t$, on a $y_1(n) = w_{11}x_1(n) + w_{21}x_2(n) - a$ et donc

$$\begin{cases} \dfrac{\partial E[e_1^2(n)]}{\partial w_{11}} = 0 \\ \dfrac{\partial E[e_1^2(n)]}{\partial w_{21}} = 0 \\ \dfrac{\partial E[e_1^2(n)]}{\partial a} = 0 \end{cases} \Rightarrow \begin{cases} E\left[e_1(n)\dfrac{\partial y_1(n)}{\partial w_{11}}\right] = 0 \\ E\left[e_1(n)\dfrac{\partial y_1(n)}{\partial w_{21}}\right] = 0 \\ E\left[e_1(n)\dfrac{\partial y_1(n)}{\partial a}\right] = 0 \end{cases}$$

c'est-à-dire

$$\begin{cases} E[e_1(n)x_1(n)]=0 \\ E[e_1(n)x_2(n)]=0 \\ E[e_1(n)]=0 \end{cases} \Leftrightarrow \begin{cases} E[d_1(n)x_1(n)] = w_{11}E[x_1^2(n)] + w_{21}E[x_1(n)x_2(n)] - a\,E[x_1(n)] \\ E[d_1(n)x_2(n)] = w_{11}E[x_1(n)x_2(n)] + w_{21}E[x_2^2(n)] - a\,E[x_2(n)] \\ E[d_1(n)] = w_{11}E[x_1(n)] + w_{21}E[x_2(n)] - a \end{cases}$$

On a donc un système de 3 équations à 3 inconnues qui nous permet de trouver w_{11}, w_{21} et a. Ceci suppose bien sur qu'on connaisse toutes les espérances intervenant dans ce système

On fait de même pour estimer w_2 en minimisant $E[e_2^2(n)]$ et pour estimer w_3 en minimisant $E[e_3^2(n)]$.

- On sait que la dérivée de la fonction sigmoïde vérifie

$$f'(x) = \alpha f(x)[1-f(x)]$$

La mise à jour des poids w_{11}, w_{21} et a à l'aide de la règle du gradient s'écrit sous la forme

$$w_{11}(n+1) = w_{11}(n) - \mu \left.\frac{\partial e_1^2(n)}{\partial w_{11}}\right|_{w_{11}=w_{11}(n)}$$

$$w_{21}(n+1) = w_{21}(n) - \mu \left.\frac{\partial e_1^2(n)}{\partial w_{21}}\right|_{w_{21}=w_{21}(n)}$$

$$a(n+1) = a(n) - \mu \left.\frac{\partial e_1^2(n)}{\partial a}\right|_{a=a(n)}$$

c'est-à-dire

$$\begin{cases} w_{11}(n+1) = w_{11}(n) - 2\mu e_1(n)\dfrac{\partial e_1(n)}{\partial w_{11}} = w_{11}(n) + 2\mu e_1(n)\dfrac{\partial y_1(n)}{\partial w_{11}} \\ w_{21}(n+1) = w_{21}(n) - 2\mu e_1(n)\dfrac{\partial e_1(n)}{\partial w_{21}} = w_{21}(n) + 2\mu e_1(n)\dfrac{\partial y_1(n)}{\partial w_{21}} \\ a(n+1) = a(n) - 2\mu e_1(n)\dfrac{\partial e_1(n)}{\partial a} = a(n) + 2\mu e_1(n)\dfrac{\partial y_1(n)}{\partial a} \end{cases}$$

Les dérivées de $y_1(n) = f[w_{11}x_1(n) + w_{21}x_2(n) - a]$ par rapport à w_{11} w_{12} et a se calculent facilement - Par exemple

$$\frac{\partial y_1(n)}{\partial w_{11}} = \alpha y_1(n)[1-y_1(n)]\,x_1(n)$$

On en déduit les règles de mise à jour de w_{11}, w_{21} et a :

$$w_{11}(n+1) = w_{11}(n) + \delta\, e_1(n)\, y_1(n)\,[1 - y_1(n)]\, x_1(n)$$
$$w_{21}(n+1) = w_{21}(n) + \delta\, e_1(n)\, y_1(n)\,[1 - y_1(n)]\, x_2(n)$$
$$a(n+1) = a(n) + \delta\, e_1(n)\, y_1(n)\,[1 - y_1(n)]\,(-1)$$

4) Règle DCMS

- Le coût de prendre la décision "classe c" prend les valeurs

s_{1c} avec proba p_{1c}

s_{2c} avec proba p_{2c}

⋮

s_{Cc} avec proba p_{Cc}

La moyenne de ce coût s'écrit donc $E[\phi_c] = \sum_{i=1}^{C} s_{ic}\, p_{ic}$

Puisque le coût associée à une décision correcte, noté s_{cc}, est nul, on a

$$E[\phi_c] = \sum_{\substack{i=1 \\ i \neq c}}^{C} s_{ic}\, p_{ic} \qquad (*)$$

- Le coût de ne pas prendre la décision "classe c" alors que les données proviennent de la classe c, prend les valeurs

s_{c1} avec la proba p_{c1}

s_{c2} avec la proba p_{c2}

⋮

s_{cC} avec la proba p_{cC}

Sa valeur moyenne est donc

$$E[r_c] = \sum_{\substack{i=1 \\ i \neq c}}^{C} s_{ci}\, p_{ci} \qquad (**)$$

- La question la ☺ difficile de l'examen car il fallait avoir compris l'article ! On utilise un tiers des données pour déterminer les vecteurs w_1, w_2 et w_3 en appliquant la règle du gradient. Une fois que ces vecteurs ont été déterminés, on estime les probabilités p_{ij} à l'aide de la relation (1) et du second tiers des données de la base d'apprentissage. Ensuite, à l'aide des relations (*) et (**) ci dessus, on peut déterminer $E[\phi_c]$ et $E[r_c]$.

On en déduit alors la matrice de coûts

$$\begin{bmatrix} 0 & E[V_c] \\ E[\phi_c] & 0 \end{bmatrix}$$

à laquelle on peut appliquer une règle de Bayes standard.

- Distance de Mahalandbis

Pour une chaque classe, on estime les vecteurs moyennes m_i et les matrice de covariance Σ_i à l'aide de la base d'apprentissage. On notera $\hat{m}_i$ et $\hat{\Sigma}_i$ les estimations obtenues. Le classifieur construit à partir de la règle de la distance aux barycentres s'écrit

$$d(x_j) = w_j \Leftrightarrow d(z, \hat{\mu}_j) \leq d(z, \hat{\mu}_k) \quad \forall k$$

avec $d(z, \mu_j) = (z - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (z - \hat{\mu}_j)$.

Classifieur KNN

Pour classifier z, on cherche ses K plus proches voisins parmi la base d'apprentissage et on affecte z à la classe majoritairement représentée par ces voisins.