Correction du Partiel de Classification et Reconnaissance des Formes

du Vendredi 23 janvier 2009, 8h00-10h00 January 26, 2009

1) Règle de décision Bayésienne

- Les hypothèses se résument à

$$\begin{aligned} C_0 &: x \sim \mathcal{N}(\mu_0, \Sigma), \qquad \pi_0 = P(C_0) \\ C_1 &: x \sim \mathcal{N}(\mu_1, \Sigma), \qquad \pi_1 = P(C_1) \end{aligned}$$

  - On en déduit

$$\begin{aligned} P(C_1|x) &= \frac{f(x|C_1)\pi_1}{f(x)}, \\ &= \frac{f(x|C_1)\pi_1}{f(x|C_1)\pi_1 + f(x|C_0)\pi_0}, \\ &= \left(1 + \frac{f(x|C_0)\pi_0}{f(x|C_1)\pi_1}\right)^{-1}. \end{aligned}$$

Mais, puisque les matrices de covariance de x sont les mêmes conditionnellement à chacune des deux classes C_0 et C_1, on a

$$\begin{aligned} \frac{f(x|C_0)}{f(x|C_1)} &= \frac{\exp\left[-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right]}{\exp\left[-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right]}, \\ &= \exp\left[x^T\Sigma^{-1}(\mu_0-\mu_1) + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right)\right], \\ &= \exp\left[-x^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right)\right]. \end{aligned}$$

En remarquant que

$$\frac{\pi_0}{\pi_1} = \exp\left[\ln\left(\frac{\pi_0}{\pi_1}\right)\right],$$

on a finalement

$$P(C_1|x) = \left\{1 + \exp\left[-x^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right) - \ln\left(\frac{\pi_1}{\pi_0}\right)\right]\right\}^{-1}$$

qui peut s'écrire

$$P(C_1|x) = \frac{1}{1+\exp(-x^T\beta - \beta_0)}$$

avec

$$\begin{aligned} \beta &= \Sigma^{-1}(\mu_1-\mu_0), \\ \beta_0 &= \frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right) + \ln\left(\frac{\pi_1}{\pi_0}\right). \end{aligned}$$

Par ailleurs

$$\begin{aligned} P(z=-1|x) &= P(C_0|x), \\ &= 1-P(C_1|x), \\ &= 1-\frac{1}{1+\exp(-x^T\beta-\beta_0)}, \\ &= \frac{\exp(-x^T\beta-\beta_0)}{1+\exp(-x^T\beta-\beta_0)}, \\ &= \frac{1}{1+\exp(x^T\beta+\beta_0)}. \end{aligned}$$

On a donc

$$P(z=\varepsilon|x) = \frac{1}{1+\exp[-\varepsilon(x^T\beta+\beta_0)]}, \qquad \varepsilon \in \{-1,1\}.$$

On notera $P(\varepsilon|x) = P(z=\varepsilon|x)$ dans ce qui suit.

- Si

$$P_1 = \frac{1}{1+\exp(-x^T\beta-\beta_0)}$$

alors

$$1-P_1 = \frac{\exp(-x^T\beta-\beta_0)}{1+\exp(-x^T\beta-\beta_0)}$$

et donc

$$\frac{P_1}{1-P_1} = \frac{1}{\exp(-x^T\beta-\beta_0)} = \exp(x^T\beta+\beta_0).$$

En prenant le logarithme de cerre expression, on retrouve bien l'expression (1).

- La règle de décision Bayésienne associée à ce problème dans le cas où les deux classes C_0 et C_1 sont équiprobables et pour des fonctions de coût $c_{ij} = 1-\delta_{ij}$ est définie par

$$\begin{aligned} d^*(x) &= C_1 \Leftrightarrow P(C_1|x) > P(C_0|x), \\ &\Leftrightarrow \frac{1}{1+\exp(-x^T\beta-\beta_0)} > \frac{1}{1+\exp(x^T\beta+\beta_0)}, \\ &\Leftrightarrow 1+\exp(x^T\beta+\beta_0) > 1+\exp(-x^T\beta-\beta_0), \\ &\Leftrightarrow T(x) = x^T\beta+\beta_0 > 0. \end{aligned}$$

Dans le cas particulier $n=2$ (i.e. $x=(x_1,x_2)^T$) l'équation de la frontière de séparation entre les régions $d^*(x)=C_1$ et $d^*(x)=C_0$ est une droite d'équation

$$T(x)=0 \Leftrightarrow \beta_1x_1+\beta_2x_2+\beta_0=0.$$

- Puisque x est un vecteur Gaussien et que β n'est pas le vecteur nul, on a

$$C_0 : T(x) = x^T\beta+\beta_0 \sim \mathcal{N}(m_0,\sigma^2)$$

avec

$$\begin{aligned} m_0 &= E[T(x)|C_0] = \mu_0^T\beta+\beta_0, \\ \sigma^2 &= \beta^T\Sigma\beta. \end{aligned}$$

De même

$$C_1 : T(x) = x^T\beta+\beta_0 \sim \mathcal{N}(m_1,\sigma^2)$$

avec

$$m_1 = E\left[T(x)|\, C_1\right] = \mu_1^T \beta + \beta_0.$$

On a alors

$$\begin{aligned} m_1 + m_0 &= \mu_1^T \beta + \mu_0^T \beta + 2\beta_0, \\ &= \mu_1^T \Sigma^{-1} (\mu_1 - \mu_0) + \mu_0^T \Sigma^{-1} (\mu_1 - \mu_0) + \left(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1\right) + 2 \ln \left(\frac{\pi_1}{\pi_0}\right). \end{aligned}$$

En développant cette expression et en faisant $\pi_0 = \pi_1$ (classes équiprobables), on obtient

$$m_1 + m_0 = 0.$$

La probabilité d'erreur du classifieur Bayésien s'écrit

$$P_e = \int_{R_0} f\left(t|\, C_1\right) \pi_1 dt + \int_{R_1} f\left(t|\, C_0\right) \pi_0 dt,$$

où $f\left(t|\, C_1\right)$ et $f\left(t|\, C_0\right)$ sont les densités de $T(x)$ conditionnellement aux classes C_1 et C_0, $R_1 =]0, \infty[$ et $R_0 =]-\infty, 0[$. On a donc

$$\begin{aligned} P_e &= \pi_1 \int_{-\infty}^{0} f\left(t|\, C_1\right) dt + \pi_0 \int_0^{\infty} f\left(t|\, C_0\right) dt, \\ &= \pi_1 \int_{-\infty}^{-m_1/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt + \pi_0 \int_{-m_0/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt. \end{aligned}$$

Comme $-m_0/\sigma = m_1/\sigma > 0$, les deux intégrales intervenant dans cette expression sont égales, d'où

$$\begin{aligned} P_e &= (\pi_1 + \pi_0) \int_{-\infty}^{-m_1/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt, \\ &= F\left(\frac{m_0}{\sigma}\right) = F\left(-\frac{m_1}{\sigma}\right). \end{aligned}$$

Quand $\sigma \to 0$, $-\frac{m_1}{\sigma}$ tend vers moins l'infini et donc la probabilité d'erreur du classifieur Bayésien tends vers 0. Ceci est compréhensible car la performance du classifieur décroit quand la variance de la statistique de test diminue.

2) Apprentissage

On désire estimer β et β_0 à l'aide d'une base d'apprentissage constituée de p vecteurs $x_1, ..., x_p$ de $\mathbb{R}^n$dont les classes sont connues. Chaque vecteur x_i est muni d'une étiquette $z_i \in \{-1, 1\}$ telle que

$$P\left(z_i|\, x_i\right) = \frac{1}{1 + \exp\left[-z_i \left(x_i^T \beta + \beta_0\right)\right]}, \qquad z_i \in \{-1, 1\}.$$

- En supposant que les étiquettes $z_1, ..., z_p$ sont indépendantes conditionnellement à $x_1, ..., x_p$, on a

$$\begin{aligned} g(\beta, \beta_0) &= \ln P\left(z_1, ..., z_p|\, x_1, ..., x_p\right), \\ &= \ln \left[\prod_{i=1}^{p} \frac{1}{1 + \exp\left[-z_i \left(x_i^T \beta + \beta_0\right)\right]}\right], \\ &= -\sum_{i=1}^{p} \ln \left\{1 + \exp\left[-z_i \left(x_i^T \beta + \beta_0\right)\right]\right\}. \end{aligned}$$

- Pour déterminer β et β_0 qui maximisent $g(\beta, \beta_0)$, il suffit de dériver cette fonction par rapport à β et β_0

$$\begin{aligned}
\frac{\partial g(\beta,\beta_0)}{\partial \beta} &= 0 \Longrightarrow -\sum_{i=1}^{p} \frac{\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]\left[[-x_i(j)z_i]\right]}{1+\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]} = 0, \qquad \forall j = 1,...,n \\
\frac{\partial g(\beta,\beta_0)}{\partial \beta_0} &= 0 \Longrightarrow -\sum_{i=1}^{p} \frac{\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right](-z_i)}{1+\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]} = 0.
\end{aligned}$$

Le vecteur β et le scalaire β_0 doivent donc vérifier le système d'équations ($n+1$ équations à $n+1$ inconnues)

$$\begin{aligned}
\sum_{i=1}^{p} \frac{x_i(j)\, z_i \exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]}{1+\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]} &= 0, \qquad \forall j = 1,...,n \\
\sum_{i=1}^{p} \frac{z_i \exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]}{1+\exp\left[-z_i\left(x_i^T\beta+\beta_0\right)\right]} &= 0.
\end{aligned}$$

- En pratique, on peut déterminer β et β_0 en utilisant un algorithme d'optimisation comme l'algorithme du gradient qui nécessite de connaitre les dérivées de $g(\beta, \beta_0)$ par rapport à β et β_0 définies ci-dessus.

3) Perceptron Multi-couches

- Le perceptron multi-couches pour la classification des signaux EEG issus de patients normaux ou épileptiques est défini par
  - quatre entrées qui sont les moyennes des coefficients en ondelettes associés aux bandes A5, D3, D4 et D5. Ce sont les moyennes des signaux A5, D3, D4 et D5 représentés sur la figure 4.
  - une couche cachée comprenant 21 neurones
  - une couche de sortie comprenant un noeud de sortie.
  - la sortie désirée vaut 1 si le signal EEG est issu d'un patient épileptique et 0 si le patient est normal

- La sortie du réseau de neurones s'écrit

$$y(n) = f\left(\sum_{i=1}^{M} u_i y_i(n) + a\right)$$

où $M = 21$ désigne le nombre de noeuds de la couche cachée et $y_i(n)$ est la sortie du *ième* noeud de cette couche cachée. La règle de mise à jour des poids u_i et du biais a s'écrit donc

$$\begin{aligned}
u_i(n+1) &= u_i(n) - \frac{\delta}{2} \left.\frac{\partial e^2(n)}{\partial u_i}\right|_{u_i=u_i(n)}, \\
a(n+1) &= a(n) - \frac{\delta}{2} \left.\frac{\partial e^2(n)}{\partial a}\right|_{a=a(n)},
\end{aligned}$$

avec $e(n) = d(n) - y(n)$, où $d(n)$ est la sortie désirée du réseau à l'instant n définie par

$$d(n) = \begin{cases} 1, \text{ si } x \in C_1, \\ 0, \text{ si } x \in C_0. \end{cases}$$

On a donc

$$\begin{aligned} u_i(n+1) &= u_i(n) + \delta e(n) y(n) \left[1 - y(n)\right] y_i(n), \\ a(n+1) &= a(n) + \delta e(n) y(n) \left[1 - y(n)\right]. \end{aligned}$$

La sortie du *ième* noeud de la couche cachée s'écrit

$$y_i(n) = f\left(\sum_{j=1}^{M} w_{ji} x_j(n) + b_j\right).$$

Donc la règle de mise à jour des poids w_{ji} et du biais b_j s'écrit

$$\begin{aligned} w_{ji}(n+1) &= w_{ji}(n) - \frac{\delta}{2} \left.\frac{\partial e^2(n)}{\partial w_{ji}}\right|_{u_i = u_i(n)}, \\ &= w_{ji}(n) + \delta e(n) y(n) \left[1 - y(n)\right] u_i(n) \left.\frac{\partial y_j(n)}{\partial w_{ji}}\right|_{w_{ji} = w_{ji}(n)}, \\ &= w_{ji}(n) + \delta e(n) y(n) \left[1 - y(n)\right] u_i(n) y_i(n) \left[1 - y_i(n)\right] x_j(n). \end{aligned}$$

.De même

$$\begin{aligned} b_j(n+1) &= b_j(n) - \frac{\delta}{2} \left.\frac{\partial e^2(n)}{\partial b_j}\right|_{b_j = b_j(n)}, \\ &= b_j(n) + \delta e(n) y(n) \left[1 - y(n)\right] u_i(n) \left.\frac{\partial y_j(n)}{\partial b_j}\right|_{b_j = b_j(n)}, \\ &= b_j(n) + \delta e(n) y(n) \left[1 - y(n)\right] u_i(n) y_i(n) \left[1 - y_i(n)\right]. \end{aligned}$$

4) Questions diverses sur l'article

- Comment les auteurs de l'article motivent-ils l'utilisation de la transformée en ondelettes pour détecter les EEG issus de patients épileptiques (plutôt par exemple que d'utiliser le spectre des signaux EEG) ? *Réponse :* les auteurs expliquent que les signaux EEG sont non-stationnaires et donc qu'il est préférable d'utiliser une représentation temps-fréquence ou temps-échelle plutôt qu'une simple analyse spectrale pour la classification de ces signaux.

- A quoi correspondent les quatre signaux représentés sur les figures 1 ? *Réponse :* ce sont des signaux enregistrés par quatre capteurs différents associés à quatre bandes de fréquences différentes.

- Comment peut-on déterminer A_i à partir de A_{i+1} et D_{i+1} sur la figure 3 ? *Réponse :* on a

$$A_i = A_{i+1} + D_{i+1}.$$

- Comment les auteurs ont-ils procédé pour déterminer le nombre de noeuds du perceptron représenté sur la figure 5 ? *Réponse :* ils ont fait plusieurs essais avec un neurone, puis deux, puis trois etc ... et ont retenu la structure qui donnait l'erreur minimale en sortie du réseau. La configuration optimale a été obtenue pour 21 neurones dans la couche cachée.

- Á plusieurs reprises, les auteurs parlent de l'algorithme "Backpropagation". De quel algorithme s'agit-il ? *Réponse :* il s'agit de l'algorithme qui permet d'estimer les poids du réseau de neurones.

- p. 95, les auteurs parlent de "Receiver operating characteristic (ROC) analysis". De quoi s'agit-il ? *Réponse :* ce sont des courbes qui représentent la probabilité de détection (ici probabilité de détecter une épilepsie sachant que le patient est épileptique) pour différentes probabilités de fausse alarme (ici la probabilité de décider que le patient est épileptique alors qu'il ne l'est pas).