Partiel de Classification et Reconnaissance des Formes

Lundi 25 janvier 2011, 8h00-10h00

INP ENSEEIHT

Partiel avec documents autorisés

On considère un vecteur $\mathbf{X}$ de $\mathbb{R}^n$ appartenant à l'une des deux classes ω_0 et ω_c représentant par exemple la présence et l'absence de visage dans une image. On suppose que les propriétés statistiques de $\mathbf{X}$ conditionnellement aux deux classes ω_0 et ω_c sont définies comme suit

$$\omega_0 \text{ (présence de visage)} : \mathbf{X} \sim \mathcal{N}\left(\mathbf{M}_0, \mathbf{\Sigma}_0\right) \text{ et } \omega_c \text{ (absence de visage)} : \mathbf{X} \sim \mathcal{N}\left(\mathbf{M}_c, \mathbf{\Sigma}_c\right) \tag{1}$$

où $\mathcal{N}\left(\mathbf{M}, \mathbf{\Sigma}\right)$ désigne la loi normale (à n dimensions) de moyenne $\mathbf{M}$ et de matrice de covariance $\mathbf{\Sigma}$. On suppose que les deux classes ω_0 et ω_c ont les probabilités a priori p_0 et p_c.

1) **Règle de décision Bayésienne :** montrer que la règle de décision Bayésienne associée au problème (1) consiste à accepter l'hypothèse ω_0 si

$$\left(\mathbf{X} - \mathbf{M}_c\right)^T \mathbf{\Sigma}_c^{-1} \left(\mathbf{X} - \mathbf{M}_c\right) - \left(\mathbf{X} - \mathbf{M}_0\right)^T \mathbf{\Sigma}_0^{-1} \left(\mathbf{X} - \mathbf{M}_0\right) > b \tag{2}$$

où b est un seuil que lon déterminera en fonction des probabilités a priori p_0 et p_c et des déterminants des matrices de covariances notés $|\mathbf{\Sigma}_0|$ et $|\mathbf{\Sigma}_c|$.

- Montrer que dans le cas où les deux classes sont équiprobables et où les matrices de covariances sont les mêmes sous les deux hypothèses, la règle (2) se réduit à la règle de la distance aux Barycentres. Quel est le nom de la distance utilisée dans cette règle de décision ?

- Dans le cas $n = 2$ et où les matrices de covariances sont égales à la matrice identité, représenter dans le plan (x_1, x_2) les zones d'acceptation des classes ω_0 et ω_c (on pourra prendra $\mathbf{M}_0 = (1, 1)$ et $\mathbf{M}_c = (0, 0)$). Montrer que la probabilité d'erreur du classifieur peut s'exprimer sous la forme de l'intégrale d'une fonction dépendant de la densité p et de la fonction de répartition F d'une loi normale $\mathcal{N}(0, 1)$, c'est-à-dire

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

2) **Apprentissage :** dans les applications pratiques, on ne connaît pas $\mathbf{M}_0, \mathbf{M}_c, \mathbf{\Sigma}_0$ et $\mathbf{\Sigma}_c$. Rappeler comment on peut estimer ces vecteurs moyennes et matrices de covariance à l'aide d'une base d'apprentissage constituée d'éléments des deux classes ω_0 et ω_c (on pourra noter $\mathbf{X}_0^{(1)}, ..., \mathbf{X}_0^{(q_0)}$ les éléments de la classe ω_0 et $\mathbf{X}_c^{(1)}, ..., \mathbf{X}_c^{(q_c)}$ les éléments de la classe ω_c).

3) **Analyse en composantes principales :** on désire effectuer une analyse en composantes principales avant la règle de décision Bayésienne. En reprenant les notations de l'article, montrer que la matrice de covariance des données de la base d'apprentissage $\mathcal{B} = \left\{\mathbf{X}_0^{(1)}, ..., \mathbf{X}_0^{(q_0)}, \mathbf{X}_c^{(1)}, ..., \mathbf{X}_c^{(q_c)}\right\}$ définie par

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

où $\mathbf{X}_i$ parcourt la base d'apprentissage $\mathcal{B}$ et $q = q_0 + q_c$ s'écrit

$$\mathbf{\Sigma}_l = \frac{q_0}{q} \mathbf{\Sigma}_0 + \frac{q_c}{q} \mathbf{\Sigma}_c + \mathbf{\Sigma}_m$$

où $\mathbf{\Sigma}_0$ et $\mathbf{\Sigma}_c$ sont les matrices de covariances des classes ω_0 et ω_c et où $\mathbf{\Sigma}_m$ est la matrice de covariance inter-classes de $\mathcal{B}$ (dont on donnera l'expression).

4) **Fonctions radiales de base**

De manière similaire à un réseau de neurones, on considère un réseau de fonctions radiales qui à un vecteur d'entrée $\mathbf{X}$ associe

$$y = d(\mathbf{X}) = w_0 + \sum_{i=1}^{k} w_i f_i(\mathbf{X}) \text{ avec } f_i(\mathbf{X}) = \exp\left(\frac{-1}{2\sigma^2}\|\mathbf{X} - \mathbf{c}_i\|^2\right)$$

où $y \in \mathbb{R}$ est la sortie du réseau, $f_i(\mathbf{X})$ est la $i^{ème}$ fonction radiale de base de centre $\mathbf{c}_i \in \mathbb{R}^n$ et où σ^2 est un paramètre qui sera supposé connu. Le réseau de fonctions radiales est défini par ses centres $(\mathbf{c}_1, ..., \mathbf{c}_k)$ et par le vecteur $\mathbf{w} = (w_0, ..., w_k)^T \in \mathbb{R}^{k+1}$. A partir de centres fixés $(\mathbf{c}_1, ..., \mathbf{c}_k)$, on cherche un vecteur poids $\mathbf{w}$ de façon à avoir $y = d(\mathbf{X})$ proche de 1 lorsque $\mathbf{X}$ est issu de la classe ω_0 et $y = d(\mathbf{X})$ proche de -1 lorsque $\mathbf{X}$ est issu de la classe ω_c. La règle de décision issue de ce réseau est alors

$$\text{Affecter } \mathbf{X} \text{ à la classe } \omega_0 \text{ si } d(\mathbf{X}) > 0 \text{ et affecter } \mathbf{X} \text{ à la classe } \omega_c \text{ si } d(\mathbf{X}) < 0$$

- Pour montrer l'intérêt d'un tel réseau, on considère le problème simple où $\omega_0 = \{(0,0), (1,1)\}$ et $\omega_c = \{(0,1), (1,0)\}$ et on considère deux centres $\mathbf{c}_1 = (0,0)$ et $\mathbf{c}_2 = (1,1)$ avec $\sigma^2 = 1$. Représenter (dans le plan) les points $[f_1(\mathbf{X}), f_2(\mathbf{X})]$ lorsque $\mathbf{X} \in \{(0,0), (1,1), (0,1), (1,0)\}$. Déterminer un vecteur $\mathbf{w} = (w_0, w_1, w_2)^T$ tel que $d(\mathbf{X}) > 0$ si $\mathbf{X} \in \omega_0$ et $d(\mathbf{X}) < 0$ si $\mathbf{X} \in \omega_c$.
- Pour simplifier, on suppose dans cette question qu'on a uniquement deux centres $\mathbf{c}_1$ et $\mathbf{c}_2$ (comme dans la question précédente) et que ces centres sont connus. Afin d'estimer le vecteur poids $\mathbf{w} = (w_0, w_1, w_2)^T$, on utilise une base d'apprentissage constituée d'éléments des deux classes ω_0 et ω_c. Chaque vecteur de la base d'apprentissage $\mathbf{X}_i$ est muni d'une étiquette $e_i \in \{-1, 1\}$ indiquant son appartenance à l'une des deux classes ω_0 ou ω_c.
  - Montrer qu'il existe un unique vecteur $\mathbf{w}$ minimisant
  $$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{q}[e_i - d(\mathbf{X}_i)]^2$$
  (où q est le nombre d'éléments de la base d'apprentissage) qui est la solution d'un système d'équations linéaires que l'on précisera.
  - Donner les règles de mise à jour des poids w_i obtenues à partir de l'algorithme de plus profonde descente (on prendra soin de préciser la règle obtenue pour w_0, w_1 et w_2)
  - Donner les règles de mise à jour des poids w_i obtenues à partir de l'algorithme LMS
- Lorsque le paramètre σ est inconnu, on peut essayer de l'estimer conjointement avec le vecteur $\mathbf{w}$. Expliciter la règle de mise à jour de ce paramètre découlant de l'algorithme LMS.

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

- Expliquer comment on peut obtenir la règle (4) à partir de la règle (2) énoncée ci-dessus (qui correspond à l'expression (3) dans l'article).
- Quelle est la motivation pour utiliser la matrice de covariance asymétrique donnée dans l'équation (6) de l'article ?
- 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
$$(P_1) : \widehat{\Sigma}_0\Phi = \lambda\left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right)\Phi \text{ et } (P_2) : \widehat{\Sigma}_c\Phi = \mu\left(\widehat{\Sigma}_0 + \widehat{\Sigma}_c\right)\Phi$$
- Après l'équation (12), expliquer "in the APCA subspace".
- Après (13), les auteurs définissent une matrice $U = \widehat{\Phi}\widetilde{\Phi}$. Quelle est la différence entre $\widehat{\Phi}$ et $\widetilde{\Phi}$?
- Expliquer les acronymes PCA, APCA, PLCDA et APCDA en précisant le type d'analyse effectué pour chacune de ces méthodes.
- Dans la partie 4.3, de quoi est constitué le vecteur de données X que l'on veut classifier et quelle est sa dimension ?