Correction du Partiel de Classification du Mercredi 19 décembre 2012.

December 21, 2012

Exercice 1

Partie 1 (7 points)

1) (2 points) Puisque les lois conditionnelles associées aux classes ω_1 et ω_2 sont définies comme suit

$$\begin{array}{rcl} \omega_1 & : & f(\mathbf{x}\,|\omega_1) = \mathcal{N}(\mathbf{m}_1, \sigma^2\mathbf{I}_2) \\ \omega_2 & : & f(\mathbf{x}\,|\omega_2) = \mathcal{N}(\mathbf{m}_2, \sigma^2\mathbf{I}_2) \end{array}$$

on a

$$d_B(\mathbf{x}) = \omega_1 \Leftrightarrow P(\omega_1|\,\mathbf{x}) \geq P(\omega_2|\,\mathbf{x})$$

avec

$$P(\omega_i|\,\mathbf{x}) = \frac{f(\mathbf{x}|\,\omega_i)\,P(\omega_i)}{f(\mathbf{x})}.$$

Puisque les classes sont équiprobables, on obtient

$$d_B(\mathbf{x}) = \omega_1 \Leftrightarrow f(\mathbf{x}|\,\omega_1) \geq f(\mathbf{x}|\,\omega_2)$$

c'est-à-dire, après avoir remplacé par les densités gaussiennes

$$d_B(\mathbf{x}) = \omega_1 \Leftrightarrow (\mathbf{x}-\mathbf{m}_1)^T(\mathbf{x}-\mathbf{m}_1) \leq (\mathbf{x}-\mathbf{m}_2)^T(\mathbf{x}-\mathbf{m}_2).$$

On reconnait la règle de la distance aux Barycentres.

2) (2 points) On peut estimer les vecteurs $\mathbf{m}_1$et $\mathbf{m}_2$ à l'aide de la base d'apprentissage

$$\begin{array}{rcl} \omega_1 & : & \mathbf{x}_1 = (1,1)^T \\ \omega_2 & : & \mathbf{x}_2 = (-1,-1)^T, \mathbf{x}_3 = (1,0)^T, \mathbf{x}_4 = (0,1)^T \end{array}$$

en utilisant les estimateurs du maximum de vraisemblance. On obtient

$$\widehat{\mathbf{m}}_1 = \mathbf{x}_1 = (1,1)^T \text{ et } \widehat{\mathbf{m}}_2 = \frac{1}{3}(\mathbf{x}_2+\mathbf{x}_3+\mathbf{x}_4) = (0,0)^T.$$

Lorsque $\mathbf{m}_1$ et $\mathbf{m}_2$ sont remplacés par leurs estimateurs dans la règle de décision Bayésienne, si on note $\mathbf{x} = (x,y)^T$, on obtient finalement

$$d_B(\mathbf{x}) = \omega_1 \Leftrightarrow x + y \geq 1.$$

On affecte donc $\mathbf{x}$ à la classe ω_1 si $\mathbf{x}$ est localisé dans le demi plan supérieur délimité par la droite d'équation $x+y=1$. De même, on affecte $\mathbf{x}$ à la classe ω_2 si $\mathbf{x}$ est localisé dans le demi plan inférieur délimité par la droite d'équation $x+y=1$. Ce classifieur n'a pas l'air très intéressant car deux points de la base d'apprentissage sur quatre sont situés sur la frontière de décision.

3) (3 points) La matrice de dispersion interclasse est définie par

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

Si on remplace $\mathbf{m}_1$ et $\mathbf{m}_2$ par leurs estimateurs, on obtient

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

La matrice de dispersion intraclasse est définie par

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{x_i \in \omega_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{x_i \in \omega_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T.$$

Si on remplace $\mathbf{m}_1$et $\mathbf{m}_2$ par leurs estimateurs, des calculs élémentaires conduisent à

$$\mathbf{S}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ et } \mathbf{S}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

d'où

$$\mathbf{S} = \mathbf{S}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Remarque : puisque $\widehat{\mathbf{m}}_2$ est le vecteur nul, on a $\mathbf{S}_2 = \mathbf{x}_2\mathbf{x}_2^T + \mathbf{x}_3\mathbf{x}_3^T + \mathbf{x}_4\mathbf{x}_4^T$.
D'après le cours, la droite résultant de la maximisation du critère de Fisher est définie par le vecteur

$$\mathbf{u} = k\mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

Si on remplace $\mathbf{m}_1$ et $\mathbf{m}_2$ par leurs estimateurs, on obtient

$$\mathbf{u} = \frac{k}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{k}{3} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Si on veut un vecteur unitaire, on choisira par exemple

$$\mathbf{u} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

qui est un vecteur directeur unitaire de la droite d'équation $x = y$.

Partie 2 (7 points)
1) (1 point) L'hyperplan séparateur donné par le classifieur SVM est clairement défini par la droite d'équation

$$x + y = \frac{3}{2}.$$

Les vecteurs supports sont $\mathbf{x}_1, \mathbf{x}_3$ et $\mathbf{x}_4$.
2)

2.1) (2 points) D'après le cours, la fonction $U(\boldsymbol{\alpha})$ à optimiser pour déterminer l'hyperplan séparateur associé au classifieur SVM est

$$U(\alpha) = -\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Y} \left(\mathbf{x}\mathbf{x}^T\right) Y\boldsymbol{\alpha} + \sum_{i=1}^{4} \alpha_i$$

avec

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ et } x = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Des calculs élémentaires conduisent à

$$\begin{aligned} U(\boldsymbol{\alpha}) &= -\frac{1}{2}\boldsymbol{\alpha}^T \begin{pmatrix} 2 & 2 & -1 & -1 \\ 2 & 2 & -1 & -1 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix} \boldsymbol{\alpha} + \sum_{i=1}^{4} \alpha_i \\ &= -\left(\alpha_1^2 + \alpha_2^2 + \frac{1}{2}\alpha_3^2 + + \frac{1}{2}\alpha_4^2\right) - 2\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_1\alpha_4 + \alpha_2\alpha_3 + \alpha_2\alpha_4 + \sum_{i=1}^{4} \alpha_i \end{aligned}$$

2.2) (1 point) Les contraintes que doivent satisfaire les coefficients $\alpha_i, i = 1, ..., 4$ sont

$$\alpha_i \geq 0 \text{ et } \sum_{i=1}^{4} \alpha_i y_i = 0$$

soit

$$\alpha_i \geq 0 \text{ et } \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 = 0.$$

2.3) (1 point) Puisque $\mathbf{x}_2$ n'est pas un vecteur support, on a

$$\alpha_2 = 0.$$

De plus, la contrainte égalité fournit la relation

$$\begin{aligned} \alpha_1 &= \alpha_2 + \alpha_3 + \alpha_4 \\ &= \alpha_3 + \alpha_4. \end{aligned}$$

La détermination de l'hyperplan séparateur peut donc se ramèner à l'optimisation du critère $V(\alpha_3, \alpha_4)$ défini comme suit

$$\begin{aligned} V(\alpha_3, \alpha_4) &= -(\alpha_3 + \alpha_4)^2 - \frac{1}{2}\alpha_3^2 - \frac{1}{2}\alpha_4^2 + (\alpha_3 + \alpha_4)^2 + 2(\alpha_3 + \alpha_4) \\ &= -\frac{1}{2}\alpha_3^2 - \frac{1}{2}\alpha_4^2 + 2(\alpha_3 + \alpha_4) \end{aligned}$$

2.4) (1 point) Le vecteur $(\alpha_3, \alpha_4)^T$ maximisant $V(\alpha_3, \alpha_4)$ annule le gradient de V, d'où

$$\begin{aligned} \frac{\partial V}{\partial \alpha_3} &= 0 \Leftrightarrow -\alpha_3 + 2 = 0 \\ \frac{\partial V}{\partial \alpha_4} &= 0 \Leftrightarrow -\alpha_4 + 2 = 0 \end{aligned}$$

et donc

$$(\alpha_3, \alpha_4) = (2, 2) \text{ et } \alpha_1 = \alpha_3 + \alpha_4 = 4.$$

Finalement

$$\boldsymbol{\alpha} = (4, 0, 2, 2)^T.$$

2.5) (1 point) Le vecteur $\mathbf{w}$ définissant l'hyperplan séparateur vérifie

$$\begin{aligned} \mathbf{w} &= \sum \alpha_i y_i x_i = 4 \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ 2 \end{pmatrix} \end{aligned}$$

Pour déterminer l'ordonnée à l'origine (paramètre b) de l'hyperplan séparateur, il suffit d'utiliser un vecteur support de chaque classe (notés $\mathbf{z}^+$ et $\mathbf{z}^-$) et d'appliquer la relation

$$b = \frac{1}{2}\left(\mathbf{w}^T\mathbf{z}^+ + \mathbf{w}^T\mathbf{z}^-\right)$$

Si on choisit $\mathbf{z}^+ = \mathbf{x}_1$ (on n'a pas le choix pour cette classe) et $\mathbf{z}^+ = \mathbf{x}_3$ (on pourrait prendre aussi $\mathbf{z}^+ = \mathbf{x}_4$, ce qui donnerait le même résultat), on obtient

$$b = \frac{1}{2}(4+2) = 3.$$

On a alors l'équation de l'hyperplan séparateur

$$w_1 x + w_2 y - b = 0 \Leftrightarrow 2x + 2y - 3 = 0$$

et on retrouve le résultat de la première question de cette exercice.

Exercice 3

Questions diverses sur l'article (9 points)

1) (1 point) Dans le problème d'optimisation défini par (1), que traduit la contrainte $\|\boldsymbol{\alpha}_l\|_0 \leq L$? Déterminer $\|\boldsymbol{\alpha}_l\|_0$ lorsque $K = 5$ et $\boldsymbol{\alpha}_l = (1, 0, 0, \sqrt{3}, 0)^T$.

Réponse : cette contrainte traduit le fait que chaque vecteur peut se décomposer sur les éléments du dictionnaire (les atomes d_i) avec au plus L coefficients non-nuls. La norme l_0 est la norme comptant le nombre d'éléments non-nuls d'un vecteur. Par exemple, dans $\boldsymbol{\alpha}_l = (1, 0, 0, \sqrt{3}, 0)^T$, il y a deux éléments non nuls donc $\|\boldsymbol{\alpha}_l\|_0 = 2$.

2) (1 point) Que représente $R^*(\mathbf{x}_l, D_j)$?

Réponse : $R^*(\mathbf{x}_l, D_j)$ représente l'erreur de reconstruction entre x_l et les atomes du dictionnaire D_j. Plus précisément, D_j étant fixé, on recherche le vecteur $\boldsymbol{\alpha}_l$ qui est parcimonieux et qui minimise l'erreur de reconstruction $\|\mathbf{x}_l - D_j\boldsymbol{\alpha}_l\|$. L'erreur minimale obtenue est notée $R^*(\mathbf{x}_l, D_j)$.

3) (1 point) Dans la référence [21], la fonction de coût $c_i^\lambda(y_1, ..., y_N)$ utilisée dans (2) est précisée

$$c_i^\lambda(y_1, ..., y_N) = \log\left[\sum_{j=1}^{N} e^{-\lambda(y_j - y_i)}\right].$$

Quand est ce que le terme

$$c_i^\lambda\left[R^*(\mathbf{x}_l, D_1), ..., R^*(\mathbf{x}_l, D_N)\right]$$

est "petit" ?

Réponse : la fonction de coût est proche de zéro lorsque y_i est égal au minimum de $y_1, ..., y_N$. Donc le terme $c_i^\lambda\left[R^*(\mathbf{x}_l, D_1), ..., R^*(\mathbf{x}_l, D_N)\right]$ est petit lorsque le vecteur $\mathbf{x}_l$ est le mieux représenté sur le dictionnaire D_i.

4) (2 points) Expliquer comment les dictionnaires représentés sur la figure 2 ont été obtenus.

Réponse : pour construire les éléments de la figure 2.(a), l'image est convertie en niveaux de gris et les contours de cette image sont extraits à l'aide du détecteur de Canny. On découpe ensuite cette image de contours en patches de tailles 16×16 et on ne conserve que les patches contenant des contours. On applique ensuite l'algorithme d'apprentissage de dictionnaire de la référence [21] qui fournit un dictionnaire "texte"

contenant 512 atomes. L'algorithme de [21] est itératif et à chaque itération, on ne conserve que les 90% de patches qui ont été les mieux classifés. Il n'est pas dit exactement comment ces 90% de patches ont été déterminés mais on peut par exemple garder ceux pour lesquels les différences d'erreurs de reconstruction (7) sont les plus grandes.

5) (2 points) Comment les auteurs de l'article proposent-ils de détecter les contours en utilisant la transformée en ondelettes ?

Réponse : les auteurs proposent de calculer les transformées en ondelettes discrètes dans les directions horizontale et verticale à différentes échelles, de calculer les dérivées numériques de ces transformées en ondelettes et de seuiller ces dérivées.

6) (1 point) Après avoir déterminés les deux dictionnaires D_1 et D_2 associés respectivement aux images de texte et de background, comment peut-on détecter les zones de texte dans une image ?

Réponse : étant donnée une image de test, on découpe cette image en patches de tailles 16×16. Chaque patch est décomposé sur les atomes des deux dictionnaires. Si l'erreur de reconstruction associée au dictionnaire D_1 est plus faible que celle associée au dictionnaire D_2, c'est-à-dire $R_1^* < R_2^*$, alors le patch est classifié comme un patch de texte. Dans le cas contraire, le patch est classifié comme un patch de background.

7) (1 point) Expliquer l'analyse effectuée pour montrer que la méthode de l'article est sensible à la taille des caractères et la modification apportée par les auteurs pour que la méthode fonctionne pour différentes tailles de caractères.

Réponse : Sur la figure 11, on voit que les performances du détecteur de texte chutent de manière significative lorsque la taille du texte est plus grande que 130 pixels. Les auteurs proposent alors d'utiliser une opération de sous-échantillonnage avant d'effectuer la classification de zones de textes.