INSTITUT NATIONAL des SCIENCES APPLIQUÉES
INSA ROUEN

Data-Mining

Examen 2007/2008

DÉPARTEMENT **ARCHITECTURE** DES SYSTÈMES **D'INFORMATION**
4ème année
G. Gasso

- Durée : 3h
- *Documents autorisés : cours et calculatrice*
- La copie du voisin n'est pas un document autorisé
- **L'annexe 1 est à rendre avec la copie finale**

1 Ka-mines (4 points)

1. On veut regrouper les points (les ronds) des figures en annexe 1 en deux clusters. On utilise l'algorithme de K-means avec $K = 2$. L'initialisation des centres μ_1 et μ_2 est marquée sur la figure en haut à gauche.

   Détailler sur les figures en annexe chaque itération (affectation des points et calcul des nouveaux centres) de l'algorithme jusqu'à convergence.

   Aucun calcul numérique n'est demandé. Justifier graphiquement les résultats de chaque itération.

2 SVM à pénalisations différentes en fonction des points (7 points)

Soit un ensemble de données étiquettées $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1,\ldots,n}$ avec $\mathcal{Y} = \{-1, +1\}$. On cherche à résoudre un problème SVM où on utilise un terme de régularisation C_i spécifique pour chaque point x_i. Le problème d'apprentissage est alors

$$\begin{array}{lll} \min_{w,b,\xi_i} & \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} C_i \xi_i & \\ \text{s.c.} & y_i \left(\langle w, x_i \rangle + b\right) \geq 1 - \xi_i & \forall \quad i = 1, \ldots, n \\ & \xi_i \geq 0 & \forall \quad i = 1, \cdots, n \end{array}$$

Dans cette formulation du problème, les ξ_i représentent les variables d'écart, les paramètres C_i représentent les termes de régularisation fixés par l'utilisateur.

1. Exprimer le lagrangien $\mathcal{L}$ correspondant à ce problème.
2. Exprimer les conditions d'optimalité du lagrangien par rapport aux variables primales w, b, ξ_i.
3. Donner la formulation du problème dual. Quelle méthode connaissez-vous pour résoudre ce problème dual ?
4. Proposer une façon de calculer le paramètre b.
5. On note respectivement $\mathcal{D}^+ = \{(x_i, y_i), y_i = 1\}$ et $\mathcal{D}^- = \{(x_i, y_i), y_i = -1)\}$ les ensembles de points des classes "positive" et "négative". On considère $C_i = C_+, \ \forall i \in \mathcal{D}^+$ et $C_i = C_-, \ \forall i \in \mathcal{D}^-$.
   En s'inspirant de la question 1, donner la nouvelle formulation du problème SVM. Que devient le problème dual ?

3 Bayes et moindres carrés (9 points)

Soit un problème de classification à deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ avec des données $\{(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}\}_{i=1,\ldots,n}$ où $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2\}$. On veut construire un classifieur en utilisant l'analyse linéaire discriminante (LDA).

1. Montrer qu'un point quelconque x est affecté à la classe $\mathcal{C}_2$ si

$$w^\top x > \frac{1}{2}\mu_2^\top \Sigma^{-1}\mu_2 - \mu_1^\top \Sigma^{-1}\mu_1 + \log(\Pr(\mathcal{C}_1)) - \log(\Pr(\mathcal{C}_2))$$

avec $w = \Sigma^{-1}(\mu_2 - \mu_1)$, $\Pr(\mathcal{C}_1)$ et $\Pr(\mathcal{C}_2)$ les probabilités a priori des classes $\mathcal{C}_1$ et $\mathcal{C}_2$.

2. On veut estimer les moyennes μ_1, μ_2 et la matrice de covariance Σ à partir des données $\{(x_i, y_i)\}_{i=1,\ldots,n}$. On notera n_1 le nombre de points de la classe $\mathcal{C}_1$ et n_2 le nombre de points de la classe $\mathcal{C}_2$ ($n_1 + n_2 = n$).
Donner l'expression des estimations $\hat{\mu}_1$, $\hat{\mu}_2$ et $\hat{\Sigma}$.

3. Plutôt que de faire la LDA, on veut estimer directement une fonction de décision $f(x) = x^\top \beta + \beta_0$ avec $\beta \in \mathbb{R}^p$ et $\beta_0 \in \mathbb{R}$ en utilisant la méthode des moindres carrés. Pour cela, on fait l'encodage suivant :

$$\text{Classe } \mathcal{C}_1 : y_i = -n/n_1 \qquad \text{Classe } \mathcal{C}_2 : y_i = -n/n_2$$

et on minimise le critère

$$J = \sum_{i=1}^{n} (y_i - x_i^\top \beta - \beta_0)^2$$

Ecrire le critère sous la forme matricielle $J = (Y - X_a\beta_a)^\top (Y - X_a\beta_a)$
avec $\beta_a = [\beta^\top \;\; 1]^\top$. On précisera le vecteur Y et la matrice X_a.

4. Ecrire la condition d'optimalité de ce problème des moindres carrés par rapport à β_a.

5. Montrer que le vecteur β vérifie la relation

$$\left((n-2)\hat{\Sigma} + \frac{n_1 n_2}{n}\hat{M}\right)\beta = n(\hat{\mu}_2 - \hat{\mu}_1) \quad \text{avec} \quad \hat{M} = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^\top$$

6. Déduire que $\hat{M}\,\beta$ est colinéaire au vecteur $(\hat{\mu}_2 - \hat{\mu}_1)$ càd $\hat{M}\,\beta \propto (\hat{\mu}_2 - \hat{\mu}_1)$.

7. En déduire alors que la solution de β obtenue par moindres carrés est colinéaire au vecteur $\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$. Comparer par rapport au vecteur w de la LDA. Quelle conclusion peut-on tirer ?

Annexe 1