INSTITUT NATIONAL des SCIENCES APPLIQUÉES INSA ROUEN

Data-Mining

Examen 2006/2007

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION
4ème année
G. Gasso - Ph. Leray

- Durée : 3h
- Documents autorisés : cours, notes personnelles et calculatrice
- **Les parties (1), (2-3) sont à rendre sur des copies séparées**

1 Clustering en entrée (6 points)

1. (1 point) **Kohonen**
   Quels sont les liens entre carte de Kohonen et ACP dont on aurait gardé les deux composantes principales seulement ?
2. (2 points) **K-means**
   Appliquez l'algorithme des k-means, à l'ensemble de points (en 1D) {1, 2, 3, 4, 6, 7, 8, 9}, avec comme initialisation les centres 1 et 2. Calculez à chaque itération l'inertie intra-clusters et l'intertie inter-clusters.
3. (3 points) **Classification hiérarchique ascendante**
   Appliquez maintenant l'algorithme de classification hiérarchique ascendante aux mêmes points (si plusieurs fusions sont possibles, commencez par celle qui concerne les points de moyenne la plus faible)
   (a) (1,5 points) en utilisant une ultramétrique de type saut maximal. Dessinez le dendogramme obtenu, et commentez le résultat
   (b) (1,5 points) en utilisant une ultramétrique de type saut minimal. Dessinez le dendogramme obtenu, commentez le résultat et comparez-le au précédent.

2 Du OneClass-SVM comme plat de résistance (10 points)

Soit un ensemble de données $\{x_i \in \mathbb{R}^2\}_{i=1,\ldots,n}$. On cherche à déterminer le cercle de plus petit rayon R et de centre a englobant ces données. Cette contrainte étant trop restrictive, on se donne alors le problème d'optimisation suivant :

$$\begin{array}{lll} & \min_{R,a\in\mathbb{R}^2,\xi_i} R^2 + \lambda \sum_{i=1}^n \xi_i & \\ \text{s.c.} & \|x_i - a\|^2 \leq R^2 + \xi_i & \forall i = 1,\ldots n \\ & \xi_i \geq 0 & \forall\, i = 1,\cdots,n \end{array}$$

Dans cette formulation du problème, ξ_i représente les variables d'écart (c'est-à-dire on autorise des points à être en dehors du cercle), λ est un *paramètre positif fixé par l'utilisateur* qui règle le compromis entre la minimisation du rayon du cercle et l'erreur liée aux variables d'écart ξ_i.

1. Exprimer le lagrangien $\mathcal{L}$ correspondant à ce problème.
2. Donner les conditions d'optimalité du lagrangien par rapport aux variables primales R, a, ξ_i. En déduire l'expression du centre du cercle a.

3. Donner la formulation du problème dual.
4. Dire pour les points suivants, ceux qui sont des points supports. On justifiera les réponses.
   (a) Points se trouvant à l'intérieur du cercle,
   (b) Points se trouvant sur le cercle,
   (c) Points à l'extérieur du cercle.
5. On veut estimer le rayon du cercle. En utilisant la condition KKT associée aux points se trouvant sur le cercle, donner l'expression de R.

3 Un petit dessert Bayesien ? (4 points)

Soit un problème de classification à deux classes C_1 et C_2. Les données $x \in \mathbb{R}^2$ des 2 classes suivent des lois conditionnelles $p(x/C_i), i \in \{1, 2\}$ qui sont des lois normales de paramètres respectifs :

– Classe C_1 : $\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$

– Classe C_2 : $\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

On suppose a priori que les deux classes sont équiprobables c'est-à-dire $P(C_1) = P(C_2) = 1/2$. On utilise l'approche bayésienne pour la classification avec des coûts 0-1.

1. Donner l'expression de la frontière de décision (On prendra soin de fournir l'expression littérale puis l'application numérique).
2. Illustrer sur un schéma les deux classes et la frontière de décision

   Rappel : soit une variable aléatoire $X \in \mathbb{R}^d$. La loi normale est donnée par

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}}$$

où $\mu \in \mathbb{R}^d$ est la moyenne, $\Sigma \in \mathbb{R}^{n \times n}$, la matrice de variance-covariance et $|\Sigma|$ son déterminant.