INSTITUT NATIONAL des SCIENCES APPLIQUÉES INSA ROUEN

Data-Mining
Examen 2011/2012

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

1 Les triangles de Bayes (5 points)

Soit un problème de classification à 2 classes ω_1 et ω_2 de lois conditionnelles définies par

$$p(x|\omega_1) = \begin{cases} 2x & \text{si} \quad 0 \leq x \leq 1 \\ 0 & \text{autrement} \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 2(1-x) & \text{si} \quad 0 \leq x \leq 1 \\ 0 & \text{autrement} \end{cases}$$

1. Illustrer sur un graphique les lois conditionnelles des deux classes
   Donner intuitivement la frontière de décision entre les deux classes et justifier la réponse.

   On veut réaliser une classification des données. Le coût d'une bonne décision est 0 et une mauvaise décision coûte β. On décide d'utiliser l'approche bayésienne.

2. Montrer qu'on décidera de façon optimale la classe ω_1 si $P(\omega_1|x) > P(\omega_2|x)$ avec $P(\omega_k|x)$ la probabilité a posteriori de la classe ω_k.
3. Expliciter la fonction de décision pour $P(\omega_1) = 1/2$. Donner la probabilité d'erreur de cette fonction de décision.
4. On considère maintenant le rejet avec un coût $\alpha = 0.25$.
   Donner la nouvelle fonction de décision. Illustrer ce rejet sur le graphique de la question 1.

2 SVDD (10 points)

Partie I

Soit un ensemble de données $\mathcal{D} = \{x_i \in \mathcal{X}\}_{i=1}^n$. On veut englober ces points dans une sphère $\mathcal{S}$ de rayon minimal en autorisant néanmoins certains points à être en dehors de la sphère. Soit $C \in \mathcal{X}$ et $R \in \mathbb{R}$ respectivement le centre et le rayon de la sphère. Le problème d'optimisation correspondant est de la forme

$$\begin{aligned} \min_{C,R,\xi_i} \quad & R^2 + \lambda \sum_{i \in \mathcal{D}} \xi_i \\ \text{sous les contraintes} \quad & \|x_i - C\|^2 \leq R^2 + \xi_i \quad \forall i = 1, \cdots, n \\ & \xi_i \geq 0 \quad \forall i = 1, \cdots, n \end{aligned}$$

où les ξ_i sont des variables d'écart (permettant d'autoriser certains points à être en dehors de la sphère) et $\lambda \geq 0$ un paramètre de régularisation fixé par l'utilisateur.

1. Écrire le Lagrangien correspondant à ce problème d'optimisation.
2. Écrire les conditions d'optimalité par rapport aux variables primales C, R et ξ_i et en déduire l'expression du centre de la sphère.
3. Formuler le problème dual correspondant.
4. Des points suivants, lesquels peuvent être points supports ? Justifier la réponse.
   (a) Points à l'intérieur de la sphère.
   (b) Points sur la sphère.
   (c) Points à l'extérieur de la sphère.
5. En déduire une façon d'estimer le rayon R.

Partie II

On veut adapter le principe précédent à la classification binaire. On suppose disposer de deux ensembles de points (les positifs et les négatifs) :

$$\mathcal{D}_+ = \{(x_i, y_i) \in \mathcal{X} \times \{1\}\}_{i=1}^{N_+}, \quad \mathcal{D}_- = \{(x_i, y_i) \in \mathcal{X} \times \{-1\}\}_{i=1}^{N_-}.$$

On veut englober le plus possible les points positifs dans une sphère de centre C et de rayon R tout en rejetant les points négatifs à l'extérieur de cette sphère. Mathématiquement ceci revient à considérer

$$\begin{aligned} \|x_i - C\|^2 &\leq R^2 + \xi_i \quad \xi_i \geq 0, \quad \forall i \in \mathcal{D}_+ \\ \|x_i - C\|^2 &\geq R^2 - \xi_i \quad \xi_i \geq 0, \quad \forall i \in \mathcal{D}_- \end{aligned}$$

1. Montrer que le problème d'optimisation peut s'écrire

$$\begin{aligned} \min_{C,R,\xi_i} \quad & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\ \text{sous les contraintes} \quad & y_i \|x_i - C\|^2 \leq y_i R^2 + \xi_i \quad \forall i = 1, \cdots, N \\ & \xi_i \geq 0 \quad \forall i = 1, \cdots, N \end{aligned}$$

   avec $N = N_+ + N_-$
2. En vous inspirant des résultats de la partie I, donner l'expression du centre de la sphère et la formulation du problème dual.
3. On veut déterminer une fonction de décision $f(x)$ permettant de dire si un point est positif ou négatif. Exprimer $f(x)$ en fonction des variables duales.

3 Regroupons-nous (5 points)

On considère un ensemble de points 1D $\mathcal{D} = \{-5, -3.5, -2.75, -0.5, 0, 0.2, 0.5, 2, 3, 5, 7\}$. On veut réaliser le clustering de ces points en utilisant deux méthodes : la classification hiérarchique ascendante (CHA) et les K-means.

1. Appliquer le CHA aux données ; dessiner le dendogramme et montrer que le nombre raisonnable de clusters est 3. On utilisera comme métrique entre deux clusters

$$d(\mathcal{C}_i, \mathcal{C}_j) = \min_{x \in \mathcal{C}_i, z \in \mathcal{C}_j} \|x - z\|,$$

2. Appliquer l'algorithme des K-means à partir des initialisations suivantes $\mu_1 = -1, \mu_2 = -0.25, \mu_3 = 1$. Illustrer sur un graphique les étapes de l'algorithme.
3. Ecrire un programme Matlab permettant de calculer votre clustering par les K-means.