INSTITUT NATIONAL des SCIENCES APPLIQUÉES INSA ROUEN

Data-Mining

Examen 2012/2013

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION

$4^{\text{ème}}$ année

G. Gasso

- Durée : 3h
- *Documents autorisés : cours et calculatrice*
- La copie du voisin n'est pas un document autorisé

1 Gruppetto (4 points)

1. **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.
2. **Classification hiérarchique ascendante**
   Appliquer le CHA aux données et dessiner le dendogramme obtenu. On utilisera comme métrique entre deux clusters $d(\mathcal{C}_i, \mathcal{C}_j) = \min \quad_{x\in\mathcal{C}_i, z\in\mathcal{C}_j} \|x - z\|$.

2 Aidons-nous (9 points)

Il a été demandé à deux groupes d'étudiants de faire des classifieurs SVM pour discriminer des caractères manuscrits. Le premier groupe a pour tâche de discriminer des $''1''$ contre des $''3''$ en utilisant des données $\mathcal{D}_1 = \{(x_{i1}, y_{i1}), i1 = 1, \cdots, n\}$ et le deuxième groupe des $''3''$ contre des $''8''$ avec des données $\mathcal{D}_2 = \{(x_{i2}, y_{i2}), i2 = 1, \cdots, n\}$. Chaque $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$.

Soient $f_1(x) = w_1^\top x$ et $f_2(x) = w_2^\top x$ les modèles qu'on cherche avec $w_1 \in \mathbb{R}^d$ et $w_2 \in \mathbb{R}^d$ les paramètres. Les problèmes étant presque similaires, les étudiants ont supposé que les solutions seront assez similaires. Ils ont défini les paramètres comme

$$w_1 = w_0 + v_1 \qquad \text{et} \qquad w_2 = w_0 + v_2.$$

Ici $w_0 \in \mathbb{R}^d$ est un vecteur de paramètres commun. v_1 et $v_2 \in \mathbb{R}^d$ représentent la variation autour de w_0. Le problème d'optimisation que les étudiants ont posé est alors

$$\begin{array}{ll}
\displaystyle\min_{w_0, v_1, v_2, \xi_{i1}, \xi_{i2}} & \displaystyle\frac{1}{2}\left(\|v_1\|^2 + \|v_2\|^2\right) + \frac{C_0}{2}\|w_0\|^2 + C\sum_{i=1}^{n}(\xi_{i1} + \xi_{i2}) \\
\text{sous les contraintes} & y_{i1}\left(w_0 + v_1\right)^\top x_{i1} \geq 1 - \xi_{i1} \quad \forall i = 1, \cdots, n \\
& y_{i2}\left(w_0 + v_2\right)^\top x_{i2} \geq 1 - \xi_{i2} \quad \forall i = 1, \cdots, n \\
& \xi_{i1} \geq 0 \quad \forall i = 1, \cdots, n \\
& \xi_{i2} \geq 0 \quad \forall i = 1, \cdots, n
\end{array}$$

1. Combien de contraintes comporte le problème d'optimisation ?
2. Montrer que le Lagrangien correspondant à ce problème d'optimisation peut s'écrire

$$\mathcal{L} = \frac{C_0}{2}\|w_0\|^2 + \frac{1}{2}\sum_{t=1}^{2}\|v_t\|^2 + C\sum_{t=1}^{2}\sum_{i=1}^{n}\xi_{it} - \sum_{t=1}^{2}\sum_{i=1}^{n}\alpha_{it}\left[y_{it}\left(w_0 + v_t\right)^\top x_{it} - 1 + \xi_{it}\right] - \sum_{t=1}^{2}\sum_{i=1}^{n}\gamma_{it}\xi_{it}$$

   avec α_{it} et γ_{it}, $\forall i, \forall t$ les paramètres de Lagrange. Quelles contraintes doivent satisfaire ces paramètres ?

3. Écrire les conditions d'optimalité par rapport aux variables primales w_0, v_t et ξ_{it}.
   (a) En déduire l'expression de w_0, v_t.
   (b) En déduire également que les α_{it} vérifient les contraintes $0 \leq \alpha_{it} \leq C$.
4. Formuler le problème dual correspondant.
5. Après résolution du problème dual on cherche l'expression des fonctions de décision. Exprimer $f_1(x)$ et $f_2(x)$ en fonction des variables duales α_{it}.
6. Proposer une procédure pour choisir les valeurs optimales des hyper-paramètres C et C_0.

3 Isométrie locale (7 points)

Soit $\{\mathbf{x}_i\}_{i=1}^n$ un ensemble de points de $\mathbb{R}^d$. Afin de les visualiser on cherche une projection $\{z_i \in \mathbb{R}^2\}_{i=1}^n$ de ces points dans un espace de dimension 2. L'objectif est de réaliser une projection non-linéaire maximisant la variance des points z_i (i.e. les points sont bien séparés) et préservant localement la distance entre les points (i.e. $\|z_i - z_j\|^2 = \|x_i - x_j\|^2$ si les points (i, j) sont voisins). On connait la matrice des distances $M \in \mathbb{R}^{n\times n}$ telle que $M_{ij} = \|x_i - x_j\|^2$.

1. Le voisinage d'un point x_i est l'ensemble de ses k plus proches voisins (kppv). Ecrire une fonction N = `kvoisins(`k, M`)` dont la sortie est une matrice $N \in \mathbb{R}^{n\times n}$ telle que
$$N_{ij} = \begin{cases} 1 & \text{si } x_j \text{ fait partie des kppv de } x_i \\ 0 & \text{autrement} \end{cases}$$
2. Soit la variance $J = \frac{1}{n}\sum_{i=1}^n (z_i - \mu)^\top (z_i - \mu)$ avec $\mu = \frac{1}{n}\sum_{i=1}^n z_i$. Montrer que
$$J = \frac{1}{n}\sum_{i=1}^n \|z_i\|^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n z_i^\top z_j$$
3. Soit la matrice $Z = [z_1\ z_2\ \cdots\ z_n]$. Le problème de projection se formule comme
$$\begin{array}{ll} \max\limits_{Z} & J \\ \text{sous les contraintes} & \|z_i - z_j\|^2 = M_{ij} \quad \forall (i,j) \quad \text{tel que} \quad N_{ij} = 1 \end{array}$$
   Comme il est difficile à résoudre, on formule un problème équivalent comme suit.
   (a) On définit $H = Z^\top Z$. Etablir que le terme général de cette matrice est $H_{ij} = z_i^\top z_j$. Exprimer alors J en fonction des termes H_{ij}.
   (b) Exprimer également $\|z_i - z_j\|^2$ en fonction des termes H_{ij}.
   (c) Montrer que H est définie positive.
   (d) Formuler alors un problème d'optimisation dépendant uniquement de la matrice H.
4. On décompose la solution $H^* = UDU^\top$ avec U la matrice des vecteurs propres et D celle des valeurs propres. Comment pouvez-vous obtenir la matrice Z connaissant H^* ?.