A. Rakotomamonjy
Ph. Leray

Data-Mining
Examen 2004/2005

Architecture des Systèmes d'Information
4^{eme} année

- Durée : 3h
- Calculatrice autorisée
- Documents autorisés :
  - supports de cours et notes personnelles
  - votre voisin n'est pas un document
- Barême indicatif sur 20 points
- **Les parties (1), (2) et (3) sont à rendre sur des feuilles séparées**

1 Lagrangien pour tous ! (7 points)

Un problème de régression consiste à chercher une fonction $f(x)$ qui estime la relation entrée-sortie d'un système à partir d'exemples $\{x_i, y_i\}_{i=1,\cdots,n}$. Dans notre cas, on suppose que $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$ et que la forme de $f(x)$ sera de type $f(x) = \langle w, x\rangle$ On cherche à résoudre un problème de regréssion décrit sous la forme du problème d'optimisation suivant :

$$\min_{w,\xi_i} \lambda\|w\|^2 + \sum_{i=1}^n \xi_i^2$$
sous contraintes
$$y_i - \langle w, x_i\rangle = \xi_i \qquad i = 1\cdots n$$

1. Donner une relation intuitive entre ce problème et un problème de minimisation des moindres carrés et expliquer le role de λ.
2. Montrer et justifier que le Lagrangien associé à ce problème s'écrit :
$$\mathcal{L}(w, \xi_i, \alpha_i) = \lambda\|w\|^2 + \sum_i \xi_i^2 + \sum_{i=1}^n \alpha_i(y_i - \langle w, x_i\rangle - \xi_i)$$
où les α_i sont les multiplicateurs associés aux contraintes d'égalité.
3. Donner toutes les conditions d'optimalité du problème et expliciter ceux associés à w et ξ_k
4. Montrer que le problème dual s'écrit :
$$\min_{\alpha_i} -\sum_i y_i\alpha_i + \frac{1}{4\lambda}\sum_i\sum_j \alpha_i\alpha_j\langle x_i, x_j\rangle + \frac{1}{4}\sum_i \alpha_i^2$$
où sous la forme matricielle suivante :
$$\min_{\alpha} -y^t\alpha + \frac{1}{4\lambda}\alpha^t K\alpha + \frac{1}{4}\alpha^t\alpha$$
où K est la matrice de terme général $K_{i,j} = \langle x_i, x_j\rangle$, et y et α des vecteurs de $\mathbb{R}^n$.
5. Montrer que la solution du dual en α s'ecrit :
$$\alpha = 2\lambda(K + \lambda I)^{-1}y$$

2 Du Bayes sinon rien (6 points)

Soit un problème de décision monodimensionnel à deux classes où la loi conditionnelle pour chaque classe $p(x|C_i)$ suit une loi normale (donc gaussienne) centrée en m_i et de variance σ^2. On considère que $P(C_1) = P(C_2) = 1/2$ et pour simplifier le problème on considere que $m_2 > m_1$ Dans le cadre de la théorie Bayesienne de la décision :

1. Montrer que la règle de décision de Bayes à minimum d'erreur (ie coût 0-1) est la suivante :
$$\mathcal{R}_1 = \{x : x \leq \frac{m_1 + m_2}{2}\} \qquad \mathcal{R}_2 = \{x : x > \frac{m_1 + m_2}{2}\}$$
2. Représenter schématiquement les $p(x|C_i)$ ainsi que les régions $\mathcal{R}_i$ et justifier intuitivement le résultat obtenu ci-dessus.

3. Montrer que la probabilité minimale d'erreur est donnée par :

$$P_E = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du$$

où $a = (m_2 - m_1)/2\sigma$

4. En utilisant l'inégalité :

$$P_E = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du \leq \frac{1}{\sqrt{2\pi}a} e^{-a^2/2} du$$

montrer que la probabilité d'erreur P_E tend vers 0 qd a tend vers ∞. Donner une justification intuitive de cette propriété.

On rappelle qu'une loi gaussien centrée en m_i de variance σ^2 s'écrit

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-m_i)^2}{\sigma^2}}$$

3 Nuées dynamiques (7 points)

Un cluster dans $\mathbb{R}^n$ n'est plus représenté seulement par son barycentre, mais par la plus petite sphère englobant les points du cluster. Cette sphère est définie par son centre $\mu_i \in \mathcal{R}^n$ et son rayon $R_i \in \mathcal{R}$.

1. (1 point) Proposer une règle d'affectation d'un point $X_j \in \mathcal{R}^n$ à l'un des K clusters $\{\mu_i, R_i\}_{1:K}$.
2. (2 points) Proposer une règle de calcul des caractéristiques de K clusters à partir d'un ensemble de points étiquetés $\{X_j, C_j\}_{1:N}$
3. (0.5 points) Comment initialiser judicieusement les K clusters $\{\mu_i, R_i\}_{1:K}$?
4. (1.5 points) Appliquer votre version des nuées dynamiques, avec $K = 3$, à l'ensemble de points de la dernière page (une itération par figure). *Aucun calcul n'est demandé ici, juste une exécution manuelle "approximative" de l'algorithme.*
5. (2 points) Proposer une extension à la fonction d'affectation permettant de rejeter l'affectation d'un nouveau point dans des conditions que vous préciserez. Comment utiliser cette fonction pour la découverte de nouveaux clusters ?