INSTITUT NATIONAL des SCIENCES APPLIQUÉES INSA ROUEN

Data-Mining

Examen 2009/2010

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 L'ordonnancement du Dr SVM (10 points)

On considère une application de recherche d'informations sur internet. Par exemple, un utilisateur à la recherche d'articles sur Paris tape la requête *Paris* et reçoit en retour des documents sur la ville de Paris, sur Paris Hilton et les sites de paris en ligne. Les documents sur la ville de Paris sont considérés comme pertinents et le reste non pertinent.

Chaque document est représenté par le vecteur $x \in \mathbb{R}^d$. Soit $\mathcal{S}_+ = \{x_i\}_{i=1}^{n_+}$ l'ensemble des documents pertinents et soit $\mathcal{S}_- = \{x_j\}_{j=1}^{n_-}$ l'ensemble des documents non pertinents. On note $x_i \succ x_j$ pour dire que le document x_i est plus pertinent que x_j et. On cherche une fonction $f(x) = \langle w, x\rangle$ permettant d'ordonner correctement les documents c'est-à-dire $f(x_i) > f(x_j)$ si $x_i \succ x_j$. On définit alors le problème d'optimisation suivant

$$\begin{aligned} \min_{w,\xi_{ij}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\sum_{j=1}^{m}\xi_{ij} \\ \text{sous les contraintes} \quad & \langle w, x_i\rangle - \langle w, x_j\rangle \geq 1 - \xi_{ij} \quad \forall i = 1, \cdots, n_+ \text{ et } \forall j = 1, \cdots, n_- \\ & \xi_{ij} \geq 0 \quad \forall i = 1, \cdots, n_+ \text{ et } \forall j = 1, \cdots, n_- \end{aligned}$$

avec C, le paramètre de régularisation et ξ_{ij} les variables d'écart.

1. Expliquer le sens du problème d'optimisation.
2. Ecrire le lagrangien correspondant au problème d'ordonnancement.
3. Ecrire les conditions d'optimalités par rapport aux variables primales w et ξ_{ij} et en déduire l'expression du vecteur de paramètres w.
4. A partir de ces éléments, exprimer le problème dual.
5. Parmi les situations suivantes, dites en justifiant votre réponse, les paires de points $(x_i, x_j) \in \mathcal{S}_+ \times \mathcal{S}_-$ qui correspondent à des points supports.
   a) (x_i, x_j) telle que $f(x_i) - f(x_j) = 1$,
   b) (x_i, x_j) telle que $f(x_i) - f(x_j) > 1$,
   c) (x_i, x_j) telle que $f(x_i) - f(x_j) = 0$,
   d) (x_i, x_j) telle que $f(x_i) - f(x_j) < 1$.
6. On définit maintenant les observations suivantes (z_ℓ, y_ℓ) avec

$$z_\ell = x_i - x_j, \qquad y_\ell = \begin{cases} 1 & \text{si} \quad x_i \succ x_j \\ -1 & \text{autrement} \end{cases}$$

   (a) Montrez que le problème d'ordonnancement précédent équivaut à un problème SVM particulier *sans le terme de biais* appliqué aux points z_ℓ. Formuler ce problème SVM.
   (b) Supposons que $n_+ = n_- = n$. Comparé à un SVM normal avec n points, quelle est la complexité en termes de nombre de paramètres du dual du problème d'ordonnancement.

2 Assurance Miner (6 points)

Une compagnie d'assurance s'est penchée sur les déclarations d'accidents de ses clients. Elle a extrait 1200 déclarations de sa base de données à partir desquelles elle a formé un jeu d'apprentissage $\{(x_i, y_i)\}_{i=1}^{1200}$ où $y_i = 1$ si la déclaration est jugée frauduleuse et $y_i = 0$ autrement. Les observations $x \in \mathbb{R}^4$ sont des informations sur les clients résumées par les variables suivantes :

- $x^{(1)}$: vaut 1 si le client habite une grande ville et 0 autrement,
- $x^{(2)}$: vaut 1 si le client est du genre masculin et 0 autrement,
- $x^{(3)}$: représente l'âge du client,
- $x^{(4)}$: représente la franchise du client.

1. Un modèle de régression logistique a été utilisé par un ingénieur de la compagnie pour traiter ce problème de classification binaire. Ce modèle est

$$\log \frac{Pr(fraude/x)}{Pr(non-fraude/x)} = [1 \;\; x^T]\theta \quad \text{avec} \quad \theta = \left[\begin{array}{ccccc} 0.9 & -0.081 & 0.47 & 0.04 & -0.01 \end{array} \right]^T.$$

   Classer alors les déclarations dont les données sont regroupées dans le tableau ci-dessous. Justifier tous vos éléments de réponse.

Identifiant déclaration	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
Cl 1	1	0	22	550
Cl 2	1	1	27	450
Cl 3	0	0	30	150
Cl 4	0	1	42	100

TAB. 1 – Données relatives aux déclarations à classer

2. Un autre ingénieur a choisi de traiter le problème en utilisant la règle de décision bayésienne. Pour ce faire, il s'est contenté de retenir seulement deux variables, l'âge du client et la franchise et a formé une nouveau jeu de données $\{(z_i, y_i)\}_{i=1}^{1200}$ avec $z_i = [x_i^{(3)} \;\; x_i^{(4)}]^T$.
   (a) Cet ingénieur désire représenter la densité conditionnelle de chaque classe par une loi normale multidimensionnelle $f(x/C_k) \equiv N(\mu_k, \Sigma_k)$, $k = 1, 2$ (l'indice 1 désigne la classe des fraudes et l'indice 2 la classe des non-fraudes). Proposer et expliciter une méthode d'estimation des paramètres des densités conditionnelles.
   (b) L'ingénieur a finalement retenu les paramètres suivants pour les densités conditionnelles

$$\mu_1 = \left[\begin{array}{c} 25 \\ 500 \end{array} \right], \quad \Sigma_1 = \left[\begin{array}{cc} 2 & 0 \\ 0 & 50 \end{array} \right] \quad \text{et} \quad \mu_2 = \left[\begin{array}{c} 36 \\ 125 \end{array} \right], \quad \Sigma_2 = \left[\begin{array}{cc} 4 & 0 \\ 0 & 25 \end{array} \right].$$

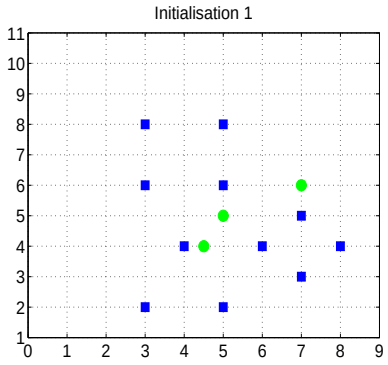
   Donner alors l'expression de la règle de décision pour des coûts 0-1 et des classes équiprobables.
   (c) Classer les déclarations des clients en utilisant cette fois-ci la règle de Bayes.

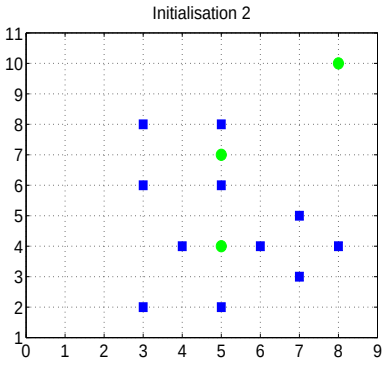
3 K-démineurs (4 points)

Les réponses à cet exercice sont à porter sur les graphiques de la feuille 4. Cette feuille est à rendre avec votre copie

On veut regrouper les points représentés sur la feuille 4 en utilisant l'algorithme de K-means avec K=3 et une *distance de Manhattan*.
Deux initialisations des centres des clusters vous sont proposées. Les centres initiaux sont les ronds.

Détailler sur les figures en annexe chaque itération (affectation des points et calcul des nouveaux centres) de l'algorithme jusqu'à convergence pour chaque initialisation. Comment justifiez-vous la différence entre les résultats ?


Initialisation 1


Initialisation 2