Data-Mining
Examen 2014/2015

– Durée : 1h30
– *Documents autorisés : cours, notes de cours et calculatrice*
– La copie du voisin n'est pas un document autorisé

1 La loi des réseaux (7 points)

Soit $K = 3$ réseaux (informatiques) $\mathcal{C}_k, k = 1, \cdots, K$ qu'on souhaite classifier automatiquement à partir de leurs caractéristiques. Chaque classe $\mathcal{C}_k$ est caractérisée par une probabilité a priori $P(\mathcal{C}_k)$ et une densité conditionnelle $p(x|\mathcal{C}_k)$. Les données (représentant ici les dates d'apparition d'évènements) de chaque classe $\mathcal{C}_k$ suivent une loi d'Erlang

$$p(x|\mathcal{C}_k) = \theta_k^2\, x\, e^{-x\,\theta_k}\, \Gamma(x) \tag{1}$$

où $\Gamma(x)$ est la fonction qui vaut 1 si $x > 0$ et 0 autrement.

1. On dispose de données d'apprentissage $\{(x_i, y_i)\}_{i=1}^N$. Pour une classe donnée, on cherche à estimer le paramètre θ_k par maximum de vraisemblance.
   (a) Donner l'expression de la log-vraisemblance.
   (b) En déduire l'estimation de θ_k au sens du maximum de vraisemblance.

2. On veut réaliser une classification des données par la règle de Bayes. Le coût d'une bonne décision est 0 et une mauvaise décision coûte α.
   (a) Donner l'expression des risques conditionnels $R(\mathcal{C}_k/x)$
   (b) Montrer qu'on décide la classe $\mathcal{C}_k$ si $P(\mathcal{C}_k|x) > P(\mathcal{C}_\ell|x)\ \ \forall \ell \neq k$
   (c) Expliciter les fonctions de décision si les classes $\mathcal{C}_k$ sont équiprobables.

2 Enveloppe englobante et sa variante quadratique (9 points)

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 qu'on veut résoudre est

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

où les ξ_i sont des variables d'écart 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. Connaissant la solution du problème dual, proposer une façon de calculer le rayon R ?

3 Métro ou voiture ? (4 points)

Une communauté d'agglomérations a réalisé une étude sur le mode de déplacement au travail (voiture personnelle ou transport en commun) de ses usagers. Elle a récolté pour chaque personne les données suivantes :

y_i	indique préférentiellement transport en commun ($y_i = 1$) ou en voiture ($y_i = 2$)
x_i	= temps mis en transport en commun - temps mis en voiture (en minutes)

A partir de ces données, le modèle de régression logistique suivant a été trouvé :

$$P(y_i = 1/x) = \frac{\exp(w_0 + w_1 x_i)}{1 + \exp(w_0 + w_1 x_i)} \quad \text{avec} \quad w_0 = -0,24 \text{ et } w_1 = 0,053$$

1. Que signifie le fait que w_1 est positif ?
2. Quelle est la probabilité qu'une personne se déplace en voiture personnelle si le trajet en transport en commun dure 30 minutes de plus ?
3. Répondre à la même question si le trajet en transport en commun dure 10 minutes de plus.
4. L'agglomération a également des informations sur le coût journalier du déplacement de chaque personne. Comment faire évoluer le modèle pour tenir compte de cette variable ?