INSTITUT NATIONAL des SCIENCES APPLIQUÉES INSA ROUEN

Data-Mining

Examen 2010/2011

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 Bayes en log-normal majeur (7 points)

Soit un problème de classification à C classes. Chaque classe ω_k est caractérisée par une probabilité a priori $P(\omega_k)$ et une densité conditionnelle $p(x|\omega_k)$. On suppose que les données de chaque classe ω_k suivent une loi log-normale avec $x \in \mathbb{R}^+ - \{0\}$

$$p(x|\omega_k) = \frac{1}{x\sigma_k\sqrt{2\pi}} exp\left(-\frac{\ln(x)-\mu_k}{2\sigma_k^2}\right) \quad (1)$$

1. Soit $\left\{x_i^{(k)}\right\}_{i=1}^{n_k}$ les données (supposées i.i.d.) de la classe ω_k. Donner l'estimation des paramètres μ_k et σ_k au sens du maximum de vraisemblance.

   On veut réaliser une classification des données. Le coût d'une bonne décision est 0 et une mauvaise décision coûte λ_s. On décide d'utiliser l'approche bayésienne. On note a_k, l'action de décider la classe ω_k.

2. Donner l'expression des risques conditionnels $R(a_k/x)$.
3. En déduire que le risque minimum est obtenu en décidant a_k si $P(\omega_k|x) > P(\omega_\ell|x) \;\; \forall \ell \neq k$ avec $P(\omega_k|x)$ la probabilité a posteriori de la classe ω_k.
4. Expliciter les fonctions de décision dans le cas suivant : $C = 2$, $P(\omega_k) = 1/2$ et $p(x|\omega_k)$ donnée par l'équation (1).
5. On considère maintenant le rejet avec un coût λ_r.
   Montrer qu'on affectera une observation x à la classe ω_k si

   $$P(\omega_k|x) > P(\omega_\ell|x) \quad \forall k \neq \ell \quad \text{et} \quad P(\omega_k|x) > 1 - \frac{\lambda_r}{\lambda_s}$$

   Que se passe-t-il si $\lambda_r = 0$? Même question si $\lambda_r > \lambda_s$.

2 Logistique à tous les étages (10 points)

On désire classer des articles sur le web en deux catégories : sports et autres. La classe Sports a pour label "0" et la classe autres "1". Chaque document est représenté par un sac de mots regroupés dans le vecteur $x \in \mathbb{R}^D$ avec D très grand. On dispose d'une séries de données $\mathcal{D} = \left\{(x_i, y_i) \in \mathbb{R}^D \times \{0,1\}\right\}_{i=1}^N$. Pour classer ces documents, on utilise la régression logistique. On suppose que $P(Y=0|x) = \frac{1}{1+\exp\left(w^\top x\right)}$ avec w le vecteur de paramètres inconnus.

Partie I

1. Donner l'expression de la log-vraisemblance $L(w)$ du problème de régression logistique.
2. Pour estimer w, on décide d'optimiser la log-vraisemblance pénalisée c'est-à-dire

$$\max_{w\in\mathbb{R}^D} J(w) \quad \text{avec} \quad J(w) = L(w) - \frac{\lambda_2}{2}\|w\|^2$$

et $\lambda \geq 0$ un paramètre de régularisation choisi par l'utilisateur.
Expliquer la signification de cette pénalisation. Comment évolue $\|w\|^2$ si on fait varier λ_2 entre 0 et $+\infty$?

3. On veut estimer w par une méthode de Newton.
   (a) En s'inspirant de votre cours, donner l'expression du gradient $g(w) = \nabla J(w)$ et de la matrice hessienne $H(w) = \nabla_{ww^\top} J(w)$.

   Nota : $\nabla_w \|w\|^2 = 2w$ et $\nabla_{ww^\top}\|w\|^2 = 2I$.

   (b) Proposer alors l'algorithme complet d'optimisation de w

Partie II

En fait la pénalité $\|w\|^2$ est jugée non satisfaisante pour mettre à zéro les paramètres les moins significatifs. On la remplace par une pénalité de type $\|w\|_1 = \sum_{j=1}^D |w_j|$. On résoud alors le problème d'optimisation

$$\max_w L(w) - \lambda_1 \sum_{j=1}^D |w_j| \quad \text{avec} \quad \lambda_1 > 0$$

L'utilisation de la méthode de Newton n'est plus possible ici car la fonction $f(z) = |z|$ n'est pas différentiable en 0. Pour contourner le problème, on remarque que pour un paramètre w_j, on peut écrire les décompositions suivantes

$$w_j = w_j^+ - w_j^- \quad et \quad |w_j| = w_j^+ + w_j^- \quad \text{avec} \quad w_j^+ \geq 0, w_j^- \geq 0$$

Le problème d'optimisation devient

$$\begin{aligned} \min_{w^+,w^-} \quad & -L(w^+, w^-) + \lambda_1 \sum_{j=1}^D \left(w_j^+ + w_j^-\right) \\ \text{sous les contraintes} \quad & w_j^+ \geq 0,\ w_j^- \geq 0 \quad \forall j = 1, \cdots, D \end{aligned}$$

4. Ecrire le lagrangien correspondant à ce problème
5. On montre que $\nabla_{w_j^+} L(w) = \nabla_{w_j} L(w)$ et $\nabla_{w_j^-} L(w) = -\nabla_{w_j} L(w)$. En utilisant ce résulat, donner en fonction de $\nabla_{w_j} L(w)$ et λ_1 l'expression des conditions d'optimalité du lagrangien par rapport à w_j^+ et w_j^-, $\forall j$.
6. On veut analyser les propriétés de ce nouveau problème.
   (a) Ecrire les conditions KKT correspondant à ce problème.

(b) Montrer que $w_j^+ > 0 \Rightarrow w_j^- = 0$. De même établir que $w_j^- > 0 \Rightarrow w_j^+ = 0$.

(c) En déduire alors que pour tout paramètre w_j non nul ($w_j \neq 0$) on a la condition $|\nabla_{w_j} L(w)| = \lambda_1$

(d) En déduire aussi que pour tout paramètre w_j nul ($w_j^+ = 0$ et $w_j^- = 0$), on a la condition $|\nabla_{w_j} L(w)| \leq \lambda_1$

3 Le bon vieux SVM (3 points)

Soit un ensemble de données $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1}^N$. On veut classer les données par un SVM par résolution du problème

$$\begin{aligned} \min_{w,\xi_i} \quad & \frac{1}{2}\|w\|^2 + \sum_{i\in\mathcal{D}} C_i \xi_i \\ \text{sous les contraintes} \quad & y_i \left(\langle w, x_i \rangle\right) \geq 1 - \xi_i, \quad \forall i \in \mathcal{D} \\ & \xi_i \geq 0, \quad \forall i \in \mathcal{D} \end{aligned}$$

Donner le problème dual correspondant.