INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Fouille de données

Examen 2016/2017

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION

— Durée : 2h
— *Documents autorisés : cours, notes de cours et calculatrice*

1 C'est quoi ce SVM ? (9 points)

Soit un ensemble de données $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$ avec $\mathcal{D}_+ = \{(\mathbf{x}_i, y_i = 1)\}_{i=1}^{N_+}$ l'ensemble des points avec le label $+1$ et $\mathcal{D}_- = \{(\mathbf{x}_i, y_i = -1)\}_{i=1}^{N_-}$ celui des points de label -1. Les données de la classe des "négatifs" sont majoritaires c'est-à-dire $N_+ << N_-$. Chaque point $\mathbf{x}_i$ est un vecteur de taille d.

On veut classer les données par un SVM linéaire $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ avec $\mathbf{w} \in \mathbb{R}^d$. Pour gérer le déséquilibre entre N_+ et N_-, on considère le problème suivant

$$
\begin{aligned}
\min_{w,b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N_-} \xi_i \\
\text{sous les contraintes} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 \quad \forall i = 1, \cdots, N_+ \\
& y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, N_- \\
& \xi_i \geq 0, \quad \forall i = 1, \cdots, N_-
\end{aligned}
$$

avec $\xi_i, i = 1, \cdots, N_-$ des variables d'écart.

1. Donner une interprétation du problème en termes d'erreurs sur les points positifs ou négatifs
2. Écrire le Lagrangien correspondant à ce problème d'optimisation
3. Donner les conditions d'optimalité par rapport aux variables primales $\mathbf{w}$, b et ξ_k ($\forall k = 1, \cdots, N_-$) et en déduire l'expression de $\mathbf{w}$.
4. Formuler alors le problème dual correspondant.
5. Des points suivants, lesquels peuvent être points supports ? Justifier la réponse.
   (a) Points positifs mal classés.
   (b) Points positifs sur la marge.
   (c) Points négatifs mal classés
   (d) Points négatifs bien classés et sur la marge.
6. Afin de choisir le paramètre C du SVM, quelqu'un propose la procédure suivante : i) choisir plusieurs valeurs possibles de C et apprendre un modèle SVM pour chaque C ; ii) évaluer chaque modèle sur les données de test ; iii) retenir le modèle SVM avec la plus petite erreur. Que pensez-vous de cette procédure ? Justifier la réponse.

2 Valse à trois (4 points)

Une agglomération s'intéresse au mode de transport de ses habitants. Elle dispose d'informations diverses comme le niveau de revenus, le temps de trajet du domicile au travail. L'agglomération s'intéresse à 3 modes de transport (les classes) : 1) bus/métro, 2) voiture, 3) co-voiturage.

Elle a donc des données $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^2 \times \{1,2,3\}\}_{i=1}^N$. On se propose de réaliser une classification des usagers en utilisant la régression logistique à 3 classes. Pour cela on définit la probabilité a posteriori de la classe k par :

$$P(y = k|\mathbf{x}) = \frac{\exp^{\mathbf{w}_k^\top \mathbf{x}}}{\sum_{\ell=1}^3 \exp^{\mathbf{w}_\ell^\top \mathbf{x}}}, \quad \forall\, k = 1,2,3$$

avec $\mathbf{w}_k \in \mathbb{R}^2$. Les paramètres $\mathbf{w}_k$ sont estimés par minimisation du critère suivant (inverse de la vraisemblance) :

$$J = \sum_{i=1}^N \left[-\mathbf{w}_{y_i}^\top \mathbf{x}_i + \log\left(\sum_{\ell=1}^3 \exp^{\mathbf{w}_\ell^\top \mathbf{x}_i} \right) \right]$$

1. Donner l'expression du gradient de J par rapport à $\mathbf{w}_k$, $\forall k = 1,2,3$
2. Proposer une méthode d'estimation des paramètres $\mathbf{w}_k$
3. Supposons que la solution obtenue est $\mathbf{w}_1 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, $\mathbf{w}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ et $\mathbf{w}_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$. Quel est le taux d'erreur sur les données test suivantes ?

$\mathbf{x}^{(1)}$	1	-2	0
$\mathbf{x}^{(2)}$	1	-1	3
y	2	1	3

3 Good Bayes (7 points)

Soit un problème de classification à deux classes dont les données $(\mathbf{x}_i, y_i) \in \mathbb{R}^2 \times \{1,2\}$ sont consignées dans le tableau ci-dessous On suppose que les points de chaque classe $\mathcal{C}_k, k \in \{1,2\}$

$\mathbf{x}^{(1)}$	2	10	6	14	6	10
$\mathbf{x}^{(2)}$	8	6	4	10	12	8
y	1	1	1	2	2	2

suivent une loi normale $\mathcal{N}(\mu_k, \Sigma_k)$ avec $\mu_k \in \mathbb{R}^2$ et $\Sigma_k \in \mathbb{R}^{2\times 2}$.

1. Pour une classe $\mathcal{C}_k, k = 1,2$ donnée, on veut déterminer les paramètres de sa loi conditionnelle et sa probabilité a priori à partir des données.
   (a) Calculer la probabilité a priori $P(\mathcal{C}_k)$ de chaque $\mathcal{C}_k$
   (b) Déterminer les paramètres $\hat{\mu}_k, \hat{\Sigma}_k$ de chaque loi conditionnelle
2. Pour classer les données avec l'approche de Bayes, on prend les coûts suivants : $\ell_{12} = \beta$, $\ell_{21} = 2\beta$ et $\ell_{11} = \ell_{22} = 0$ avec $\beta > 0$.
   (a) Donner l'expression des risques conditionnels $R(\mathcal{C}_k|\mathbf{x})$
   (b) Sous quelle condition sur $P(\mathcal{C}_1|\mathbf{x})$ et $P(\mathcal{C}_2|\mathbf{x})$ on décide d'affecter $\mathbf{x}$ à la classe $\mathcal{C}_1$?
   (c) Donner alors l'expression de la frontière de décision en fonction de $\mathbf{x}$ et des $\hat{\mu}_k, \hat{\Sigma}_k$.