Fouille de données

Examen 2015/2016

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

1 Non, pas encore du Bayes ! (7 points)

Soit un problème de classification à deux classes. Chaque classe $\mathcal{C}_k, k \in \{1, 2\}$ est caractérisée par une probabilité a priori $P(\mathcal{C}_k)$ et une loi conditionnelle $p(x|\mathcal{C}_k)$ définie par

$$p(x|\mathcal{C}_k) = \frac{x}{a_k^2}\, e^{-\frac{x^2}{a_k^2}}$$

avec $a_k \in \mathbb{R}$ le paramètre inconnu et $x \in \mathbb{R}^+ - \{0\}$. On dispose d'un ensemble de données $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ avec $y \in \{\mathcal{C}_1, \mathcal{C}_2\}$

1. Pour une classe $\mathcal{C}_k$ donnée, on veut déterminer à partir des données le paramètre a_k de sa loi par la méthode du maximum de vraisemblance.
   (a) Donner l'expression de la log-vraisemblance.
   (b) En déduire l'estimation de $\hat{a}_k$ au sens du maximum de vraisemblance.
2. Pour classer les données avec l'approche de Bayes, on prend les coûts suivants : $\ell_{12} = 2\beta$, $\ell_{21} = \beta$ et $\ell_{11} = \ell_{22} = 0$ avec $\beta > 0$. On considère que les classes sont équiprobables c'est-à-dire $P(\mathcal{C}_1) = P(\mathcal{C}_2)$.
   (a) Donner l'expression des risques conditionnels $R(\mathcal{C}_k/x)$
   (b) Sous quelle condition sur $P(\mathcal{C}_1|x)$ et $P(\mathcal{C}_2|x)$ on décide d'affecter x à la classe $\mathcal{C}_1$?
   (c) Donner alors l'expression de la frontière de décision en fonction de x et des $\hat{a}_k$.
3. On introduit maintenant le rejet en ambigüité avec un coût $\alpha = r\beta$. Montrer qu'un point x sera rejeté si $r < P(\mathcal{C}_1|x) < \frac{2-r}{r}$. Analyser en fonction des valeurs de r comment se passe le rejet.

2 Logistique plus un - moins un (10 points)

On veut élaborer un module de reconnaissance de visages. Pour faire simple on considère deux personnes dont on a pris les photos dans différentes conditions. Chaque image $I_i \in \mathbb{R}^{50\times 50}$ est convertie en un vecteur $x_i \in \mathbb{R}^{2500}$. On a donc un ensemble de données $\mathcal{D} = \left\{(x_i, y_i) \in \mathbb{R}^{2500} \times \{-1, 1\}\right\}_{i=1}^n$.

On va utiliser la régression logistique avec des classes codées en -1 et 1. On modélise

$$P(y = -1/x) = \frac{\exp\left(\theta^\top x\right)}{1 + \exp\left(\theta^\top x\right)}$$

avec θ le vecteur de paramètres inconnus.

1. Donner l'expression de la log-vraisemblance $J(\theta)$ du problème de régression logistique.
2. Expliquer comment procéder pour déterminer θ et comment vous pouvez vérifier que le modèle appris est performant (*NB : aucun calcul n'est demandé*).

   Le modèle est lourd à apprendre et à utiliser car la dimension de θ, $D = 2500$ est grande. On souhaite donc mettre à zéro les paramètres les moins significatifs en pénalisant la valeur absolue des coefficients θ_j. On résoud le problème d'optimisation

$$\max_{\theta \in \mathbb{R}^D} J(\theta) - \lambda \sum_{j=1}^{D} |\theta_j| \quad \text{avec} \quad \lambda > 0$$

   $|\theta_j|$ n'étant pas différentiable on utilise la décomposition suivante

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

   Posons $V(\theta^+, \theta^-) = J(\theta^+ - \theta^-) = J(\theta)$. Le problème d'optimisation devient

$$\begin{array}{ll} \displaystyle\min_{\theta^+ \in \mathbb{R}^D, \theta^- \in \mathbb{R}^D} & -V(\theta^+, \theta^-) + \lambda \displaystyle\sum_{j=1}^{D} \left(\theta_j^+ + \theta_j^-\right) \\ \text{sous les contraintes} & \theta_j^+ \geq 0 \quad \forall j = 1, \cdots, D \\ & \theta_j^- \geq 0 \quad \forall j = 1, \cdots, D \end{array}$$

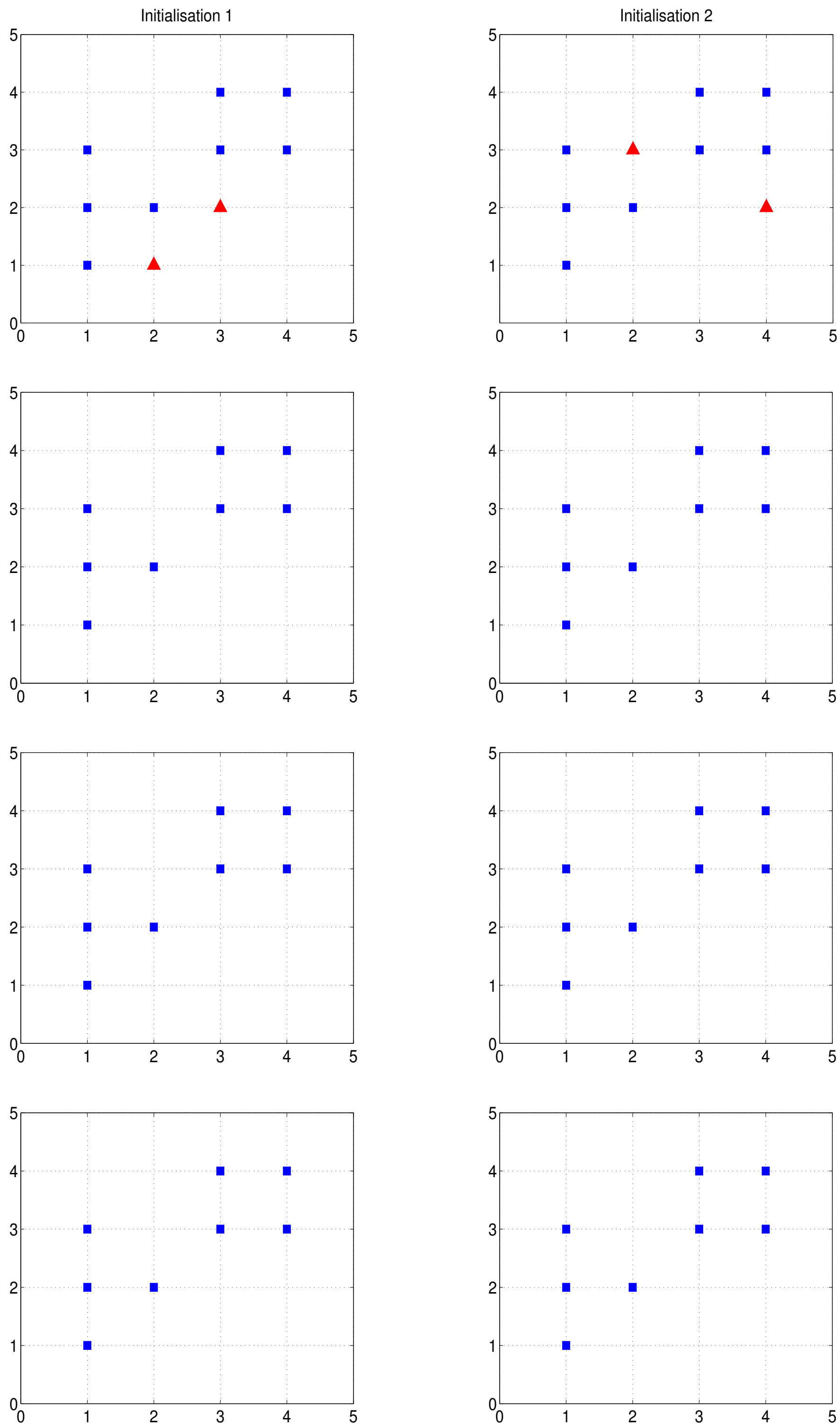
3. Ecrire le lagrangien correspondant à ce problème
4. On montre que $\nabla_{\theta_j^+} V(\theta^+, \theta^-) = \nabla_{\theta_j} J(\theta)$ et $\nabla_{\theta_j^-} V(\theta^+, \theta^-) = -\nabla_{\theta_j} J(\theta)$.

   En utilisant ce résultat, donner en fonction de $\nabla_{\theta_j} J(\theta)$ et λ les conditions KKT d'optimalité du lagrangien par rapport à θ_j^+ et θ_j^-, $\forall j$.
5. On veut déterminer sous quelle condition cette méthode peut mettre des coefficients à 0
   (a) Ecrire les conditions KKT de complémentarité correspondant à ce problème.
   (b) Montrer que $\theta_j^+ > 0 \Rightarrow \theta_j^- = 0$. De même établir que $\theta_j^- > 0 \Rightarrow \theta_j^+ = 0$.
   (c) En déduire alors que pour tout paramètre $\theta_j \neq 0$ on a la condition $|\nabla_{\theta_j} J(\theta)| = \lambda$ et que seuls ces paramètres restent

3 Coloriage (3 points)

Les réponses à cet exercice sont à porter sur les graphiques de la feuille 3. Cette feuille est à rendre avec votre copie

On veut regrouper les points représentés sur la feuille 3 en utilisant l'algorithme de K-means avec K=2. Deux initialisations des centres des clusters vous sont proposées. Les centres initiaux sont les triangles. Détailler sur les figures en annexe chaque itération de l'algorithme jusqu'à convergence pour chaque initialisation.


Initialisation 1
Initialisation 2