Examen TIAD - Master 1 IM
Vendredi 23 Mai 2014

Durée 3 h. Calculatrices autorisées. Seul un formulaire sur feuille double est autorisé.

Tran Viet Chi, `chi.tran@univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Analyse discriminante sur SAS)
Nous avons réalisé une analyse discriminante des 150 iris que Fisher a mesurées en 1936. Ces fleurs proviennent des espèces Setosa, Versicolor et Virginica et sont décrites par 4 variables explicatives : les longueurs et largeurs des sépales et pétales. L'analyse discriminante que l'on a réalisée se fonde sur une analyse factorielle des variances inter et intra-classes. Les résultats SAS se trouvent en annexe.

1. En vous appuyant sur l'annexe 2, combien y a-t-il de fleurs de chaque espèce ?

2. Quels sont les centres de gravités de chaque espèce ? Calculer les carrés des distances euclidiennes entre ces centres. Comparer avec les résultats de l'annexe 3 qui montrent le carré des distances de Mahalanobis qui accorde un poids moindre aux composantes les plus dispersées. Les distances de Mahalanobis entre les centres des différentes espèces sont-elles significatives ?

3. A partir des tableaux de l'annexe 1, calculer les variances intra et inter pour la variable `SepalLength`. Une analyse de la variance est donnée à l'annexe 4. En considérant les R^2, quelles sont les variables les plus discriminantes pour comprendre les classes ? Y a-t-il des variables dont la variance n'est pas significativement expliquée par l'espèce ?

4. Les vecteurs propres associés aux deux plus grandes valeurs propres de $W^{-1}B$, où W est la matrice de variance-covariance intra-classe et B la matrice de variance-covariance inter-classe sont donnés dans le tableau `Raw Canonical Coefficients` de l'annexe 5. Exprimer le facteur 1 (`Can1`) en fonction des 4 variables initiales. Le facteur 1 a-t-il des composantes en accord avec les remarques faites à la question précédente ?

Exercice 2 (Estimateur à noyaux)
Soient $X_1, \dots X_n$, n variables aléatoires i.i.d. de densité f strictement positive par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On suppose que f est bornée, 2 fois dérivable et de dérivée seconde Lipschitzienne :

$$\forall a, b \in \mathbb{R}, \qquad |f''(a) - f''(b)| \leq L|a - b|.$$

On se donne un noyau K pair ($K(x) = K(-x)$) tel que

$$\int_{\mathbb{R}} K(u)du = 1, \qquad \int_{\mathbb{R}} u^2 K(u)du = 0, \qquad C_K = \int_{\mathbb{R}} |u|^3 |K(u)|du < +\infty$$
$$\int_{\mathbb{R}} K^2(u)du < +\infty, \qquad \forall x \in \mathbb{R},\ K(0)f(x) > 0. \tag{1}$$

On rappelle que l'estimateur à noyau de f associé à K et à la fenêtre $h > 0$ est :

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big), \qquad \text{pour } x \in \mathbb{R}.$$

1. Rappeler la définition du biais $b(x)$ en un point $x \in \mathbb{R}$ donné, de $\widehat{f}$ et montrer que ce biais est majoré par $C_1 h^3$ où C_1 est une constante que l'on exprimera en fonction de L et C_K.

2. Montrer qu'il existe $C_2 > 0$ telle que pour tout $x \in \mathbb{R}$,

$$\mathrm{Var}_f\big(\widehat{f}(x)\big) \leq \frac{C_2}{nh}.$$

3. En déduire un majorant du risque quadratique ponctuel de $\widehat{f}$ au point x.

4. Montrer que la fenêtre optimale h^* minimisant le majorant du risque quadratique obtenu à la question **3.** est en $n^{-1/7}$.

5. Pour α et $\beta > 0$, soit

$$K_{\alpha,\beta}(u) = \frac{\alpha}{2}\mathbf{1}_{[-1,1]}(u) + \frac{\beta}{2}\mathbf{1}_{[-2,2]}(u).$$

Déterminer α et β pour que $K_{\alpha,\beta}$ satisfasse les conditions (1). Dessiner $K_{\alpha,\beta}$.

6. Ecrire $\widehat{f}_{\alpha,\beta}$ l'estimateur à noyau associé à $K_{\alpha,\beta}$. Quelles sont les défauts de cet estimateur ? Comment peut-on les corriger ?

Exercice 3 (Estimateur d'une régression par un polynômes locaux)
On observe n couples d'observations $(X_1, Y_1), \dots (X_n, Y_n)$ i.i.d. tels que les X_i sont à valeurs dans $[0, 1]$ et tels que :

$$\forall i \in \{1, \dots, n\},\ Y_i = f(X_i) + \varepsilon_i,$$

où les ε_i sont des résidus i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et où f est une fonction à déterminer que l'on suppose de classe $\mathcal{C}^1$ à dérivée Höldérienne d'ordre α et de constante L :

$$\forall a,\ b \in [0, 1],\ |f'(a) - f'(b)| \leq L|a - b|^\alpha.$$

On pose

$$U(u) = \begin{pmatrix} 1 \\ u \end{pmatrix} \qquad \text{et} \qquad \theta(x) = \begin{pmatrix} f(x) \\ f'(x)\ h \end{pmatrix}$$

et on notera ${}^t\theta(x)$ la transposée de $\theta(x)$. On définit pour un noyau K et une fenêtre $h > 0$:

$$\widehat{\theta}(x) = \arg \min_{\theta = {}^t(\theta_1, \theta_2) \in \mathbb{R}^2} \sum_{i=1}^{n} \Big(Y_i - {}^t\,\theta U\big(\frac{X_i - x}{h}\big)\Big)^2 K\big(\frac{X_i - x}{h}\big).$$

1. Montrer que $\widehat{\theta}(x)$ est l'$\arg\min$ de $\theta =^t (\theta_1, \theta_2) \mapsto -2^t\theta\mathbf{a} +^t \theta\mathbf{B}_{nx}\theta$ où $\mathbf{a}$ et $\mathbf{B}_{nx}$ sont un vecteur et une matrice à préciser.

2. En déduire que $\mathbf{B}_{nx}\widehat{\theta}(x) = \mathbf{a}$. Résoudre ce système de 2 équations à 2 inconnues, θ_1 et θ_2 pour déterminer $\widehat{\theta}(x)$.

3. Quel est l'estimateur $\widehat{f}(x)$ de $f(x)$ que l'on peut proposer en fonction de $\widehat{\theta}(x)$ et $U(0)$? En utilisant la question **2.** , déduire $\widehat{f}(x)$ qu'on pourra exprimer sous la forme $\widehat{f}(x) = \sum_{i=1}^n Y_i W_{ni}^*(x)$ où $W_{ni}^*(x)$ est une fonction à préciser.

4. Vérifier que $\sum_{i=1}^n W_{ni}^*(x) = 1$ et que $\sum_{i=1}^n (X_i - x)W_{ni}^*(x) = 0$. En supposant de plus que K est à support dans $[-1, 1]$, justifier que $W_{ni}^*(x) = 0$ si $|X_i - x| > h$.

5. Dans la suite, on admettra que $\mathbb{E}\big(\sum_{i=1}^n |W_{ni}^*(x)|\big) < +\infty$. En utilisant la question 4 obtenir une majoration du biais de $\widehat{f}(x)$ en $C_2 h^{1+\alpha}$ où C_2 est une constante que l'on précisera.