UNIVERSITÉ LUMIÈRE LYON 2

Université Lumiére Lyon 2
Master Informatique - 2ème année
parcours Data Mining

Model-based learning
Examen du 01/02/2018

Durée : 2h - Tous documents autorisés

1 Nombre de SMS et qualité du sommeil

Une étude s'intéresse à la qualité du sommeil des étudiants en fonction du nombre de SMS qu'ils reçoivent. Pour cela, un échantillon $(\underline{X}, \underline{Z}) = (X_i, Z_i)_{i=1,\dots,n}$ a été relevé, où X_i est le nombre moyen de SMS reçus par jour pour le $i$ème étudiant et $Z_i = (Z_{i1}, Z_{i2})$ la variable indiquant si l'étudiant a des troubles du sommeil (dans ce cas $Z_{i1} = 1$ et $Z_{i2} = 0$) ou non (dans ce cas $Z_{i1} = 0$ et $Z_{i2} = 1$).

Le nombre moyen de SMS reçus par jour par un étudiant est modélisé par une distribution de Poisson $\mathcal{P}(\lambda)$, de densité $f(x)$ donnée par :

$$f(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{si } x \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

1.1 Classification (10 points)

Nous cherchons à estimer un modèle de classification permettant de prédire si un étudiant va ou non avoir des troubles du sommeil en fonction du nombre de SMS qu'il reçoit.
Nous considérons pour cela un modèle de mélange de loi de Poisson, à deux composantes. Nous notons λ_1 le paramètre de la loi exponentielle pour les étudiants n'ayant pas de troubles du sommeil ($Z_{i1} = 1$) et λ_2 celui pour les étudiants ayant des troubles du sommeil ($Z_{i2} = 1$). De même, nous noterons π_1 la proportion d'étudiants n'ayant pas de troubles du sommeil, et $\pi_2 = 1 - \pi_1$.

1. Quelle est la densité de probabilité de $X_i | Z_{i1} = 1$?
2. Quelle est la densité de probabilité marginale de X_i ?
3. Ecrire la vraisemblance du modèle de mélange de Poisson de paramètre $\theta = (\pi_1, \pi_2, \lambda_1, \lambda_2)$ pour l'échantillon $(\underline{X}, \underline{Z})$.
4. On suppose π_1 et π_2 connus. Calculer l'estimateur du maximum de vraisemblance λ_1 et λ_2.
5. Supposons cette fois que l'on dispose d'une estimation $\hat{\theta}$ de θ. Comment classer une nouvelle observation x^* par la règle du maximum a posteriori pour ce modèle ?
6. Que cela donne-t-il pour $x^* = 2$, $(\hat{\pi}_1, \hat{\pi}_2, \hat{\lambda}_1, \hat{\lambda}_2) = (0.3, 0.7, 1, 3)$?

1.2 Clustering (8 points)

Dans une seconde étude, nous n'avons relevé que les nombres de SMS moyens $\underline{X} = (X_i)_{i=1,\dots,n}$, et nous cherchons à faire un clustering de ces données en K clusters. Notons $Z_i = (Z_{ik})_{1 \leq k \leq K}$ avec $Z_{ik} = 1$ si l'observation i appartient au cluster k et 0 sinon.

1. Ecrire la vraisemblance du modèle de mélange de Poisson de paramètre $\theta = (\pi_k, \lambda_k)_{1 \leq k \leq K}$ pour l'échantillon $\underline{X}$.
2. Pourquoi cette vraisemblance ne peut-elle pas être maximisée analytiquement ?
3. Proposer un algorithme EM pour maximiser cette vraisemblance, en détaillant les calculs nécessaires aux étapes E et M.
4. Peut-on choisir le nombre K de clusters par maximum de vraisemblance ? Si oui, donner son estimateur. Si non, comment peut-on faire alors ?

2 Généralités sur les modèles de mélanges (2 points)

1. En statistique, quel est l'intérêt de considérer des modèles parcimonieux ? Dans cette optique, quel est l'avantage des modèles de mélanges ?