Université Paris Cité
Master 2 Mathématiques et Applications
Parcours MMA

Classification

TP noté : Modèles de mélange et régression logistique
Date limite de rendu du compte-rendu : lundi 13 novembre 2023

Instructions : Un compte-rendu de TP est à rendre par étudiant. Il devra être rédigé en Rmarkdown et vous devrez me l'envoyer par mail à l'adresse olivier.bouaziz@parisdescartes.fr. Le compte-rendu pourra être rédigé au format **pdf** ou **html** et avoir pour nom *nometudiant_classif.pdf*.

Exercice 1

Dans cet exercice le but est d'implémenter la méthode LDA en utilisant les résultats du cours (sans utiliser la fonction `lda` du package `MASS`) et d'étudier ses performances sur des données simulées. Puis on comparera les résultats obtenus avec votre implémentation et celle obtenue avec la fonction `lda`.

1. Simulez un échantillon de taille $n = 200$ de données issues d'un mélange Gaussien homoscédastique ayant les caractéristiques suivantes : $Y \sim \mathcal{B}(p)$, $p = 0.8$, X est à valeurs dans $\mathbb{R}^2$ et

$$X|Y = 0 \sim \mathcal{N}_2(\mu_0, \Sigma), \qquad X|Y = 1 \sim \mathcal{N}_2(\mu_1, \Sigma),$$

où $\mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mu_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ et $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. On pourra pour cela utiliser la fonction `rmvnorm` du package `mvtnorm`.

2. Rappelez les formules du cours permettant d'obtenir les estimateurs de μ_0, μ_1, p, Σ et les implémenter sous R **sans utiliser la fonction** `lda`.
3. Rappelez l'équation de la frontière de classification (formule vue en cours). C'est la droite telle que les points au dessus de cette droite et ceux en dessous n'ont pas la même étiquette.
4. Tracez le nuage de points des observations en coloriant de deux couleurs différentes les données pour lesquelles $Y_i = 0$ et $Y_i = 1$. Sur ce même graphique, rajouter les estimations de μ_0 et μ_1 obtenus à la question 2. Puis y rajouter la frontière de classification à partir de l'équation donnée à la question précédente.
5. Donnez une règle de classification à partir des estimations de μ_0, μ_1, p et Σ. Simulez un nouvel échantillon test de taille 10 000 suivant les caractéristiques définies à la question 1. Calculez le taux de mauvaises classifications sur cet échantillon test.
6. Comparez vos résultats avec ceux obtenus en utilisant la fonction `lda` du package `MASS`. Vous devez en particulier montrer que vous obtenez les mêmes estimations de μ_0, μ_1, p et le même taux de mauvaises classification.
7. Calculez la distance de Mahalanobis entre les lois $\mathcal{N}(\mu_0, \Sigma)$ et $\mathcal{N}(\mu_1, \Sigma)$. Rappelez la définition du classifieur de Bayes dans le cadre de mélange Gaussien homoscédastique puis rappelez la formule du risque de ce classifieur de Bayes (formule vue en cours). Sous R, calculez la valeur numérique du risque du classifieur de Bayes en utilisant la fonction `pnorm` et comparez avec le résultat obtenu à la question précédente.

8. Sur les mêmes données d'apprentissage que précédemment, implémentez un nouvel algorithme de classification basé sur la régression linéaire à l'aide de la fonction `lm` de R. Sur les mêmes données tests que précédemment, calculez le taux de mauvaises classification avec ce nouvel algorithme et comparez avec vos résultats de la question 5.
9. Construire une fonction prenant en entrée les paramètres de taille d'échantillon (d'apprentissage), μ_0, μ_1, p et Σ permettant de retourner le taux de mauvaise classification obtenu à partir des deux méthodes (celle de l'algorithme LDA et celle basée sur la régression linéaire) évalué sur un échantillon test de taille 10 000. Faites varier les paramètres et résumez vos résultats dans un tableau contenant trois tailles d'échantillon n différents et pour chaque n trois jeux de paramètres μ_0, μ_1, p, Σ différents. Le but est de présenter des scénarios où la méthode a un faible taux de mauvaise classification et d'autres où le taux de mauvaise classification est élevé.
10. Existe-il des valeurs de p pour lesquels la méthode LDA et la méthode basée sur la régression linéaire donnent la même règle de classification ?

Exercice 2

Dans cet exercice, on s'intéresse à la prédiction du diabète chez les femmes, en fonction de plusieurs variables cliniques. Les données se trouvent dans la table **diabetes.csv**. La variable à prédire est la variable **Outcome**. Dans ce jeu de données, les variables **SkinThickness** et **Insulin** contiennent des données manquantes, codées par la valeur 0.

1. Importez le jeu de données sous R. Traitez les données manquantes et faites une analyse descriptive de la base de données (statistiques descriptives univariées et bivariées, ACP, ...).
2. Créez un échantillon d'apprentissage avec 80% de la base de données et un échantillon test avec les 20% restants.
3. Réalisez une analyse discriminante Gaussienne homoscédastique (à partir de la fonction `lda`) et hétéroscédastique (à partir de la fonction `qda`) sur la base d'apprentissage. Calculer le taux de mauvaise classification sur l'échantillon test.
4. Effectuez la même démarche en utilisant cette fois une règle de classification basée sur la régression logistique.
5. Effectuez la même démarche en utilisant cette fois une règle de classification basée sur les k-plus proches voisins, avec $k = 5$.
6. Pour le modèle de régression logistique, proposez une méthode de sélection de variables. Calculez le taux de mauvaise classification avec ce nouveau modèle.
7. Pour la méthode des k-plus proches voisins, proposez une méthode du choix de k basée sur le risque de classification.
8. Au final, quelle méthode donne le plus petit taux de mauvaise classification ?