Université de Paris
Master 2 Mathématiques et Applications
Parcours MMA

Classification

Exercice supplémentaire : lien entre le modèle de mélange homoscédastique et le modèle de régression linéaire

Exercice 1

On rappelle que l'algorithme LDA suppose que l'on observe $(X_1, Y_1), \dots, (X_n, Y_n)$ n paires de variables aléatoires indépendantes de même loi que (X, Y) suivant un modèle de mélange Gaussien homoscédastique, avec $Y \in \{0, 1\}$, $X \in \mathbb{R}^d$, $X \mid Y = 0$ suit une loi $\mathcal{N}(\mu_0, \Sigma)$ et $X \mid Y = 1$ suit une loi $\mathcal{N}(\mu_1, \Sigma)$. On note n_0 le nombre d'individus tels que $Y_i = 0$ et n_1 le nombre d'individus tels que $Y_i = 1$, avec $n_0 + n_1 = n$.

1. Montrer que l'algorithme LDA classifie une observation $x \in \mathbb{R}^d$ dans le groupe 1 si

$$x^t \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) > \frac{1}{2}\hat{\mu}_1^t \hat{\Sigma}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_2^t \hat{\Sigma}^{-1}\hat{\mu}_2 + \log\left(\frac{n_0}{n}\right) - \log\left(\frac{n_1}{n}\right),$$

où

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^n Y_i X_i, \quad \hat{\Sigma} = \frac{1}{n-2}\sum_{k=0}^1 \sum_{i:Y_i=k} (X_i - \mu_k)(X_i - \mu_k)^t.$$

2. On considère à présent le modèle de régression linéaire qui consiste à minimiser, par rapport à $\beta \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$, le critère suivant :

$$\sum_{i=1}^n \left(Y_i - \beta_0 - \beta^t X_i\right)^2.$$

Montrer que les solutions $\hat{\beta}$, $\hat{\beta}_0$ vérifient

$$\hat{\beta}_0 = \frac{n_1}{n} - \frac{1}{n}\sum_{i=1}^n X_i \hat{\beta} \quad \text{et}$$

$$\hat{\beta}_0 X_i \sum_{i=1}^n X_i + \sum_{i=1}^n X_i X_i^t \hat{\beta} = n_1 \hat{\mu}_1$$

3. Montrer que

$$n(\hat{\mu}_0 - \hat{\mu}_1) = \frac{n}{n_0}\left(\sum_{i=1}^n X_i - n\hat{\mu}_1\right),$$

et en déduire que l'on a

$$\frac{n_1 n_0}{n}(\hat{\mu}_1 - \hat{\mu}_0) = \left(\sum_{i=1}^n X_i X_i^t - \frac{1}{n}\sum_{i=1}^n X_i \sum_{i=1}^n X_i^t\right)\hat{\beta}$$

4. On introduit à présent

$$\hat{\Sigma}_B = (\hat{\mu}_1 - \hat{\mu}_0)(\hat{\mu}_1 - \hat{\mu}_0)^t .$$

Montrer que

$$(n-2)\hat{\Sigma} = \sum_{i=1}^{n} X_i X_i^t - n_1 \hat{\mu}_1 \hat{\mu}_1^t - n_0 \hat{\mu}_0 \hat{\mu}_0^t,$$

et en déduire que l'on a

$$(n-2)\hat{\Sigma} + \frac{n_1 n_0}{n} \hat{\Sigma}_B = \left(\sum_{i=1}^{n} X_i X_i^t - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i^t \right).$$

En combinant ce résultat avec la question 3., montrer que

$$n(\hat{\mu}_0 - \hat{\mu}_1) = \left((n-2)\hat{\Sigma} + \frac{n_1 n_0}{n} \hat{\Sigma}_B \right) \hat{\beta}$$

5. Montrer que $\hat{\Sigma}_B \hat{\beta}$ est proportionnel à $\hat{\mu}_1 - \hat{\mu}_0$ et déduire de la question précédente qu'il existe $\lambda \in \mathbb{R}$ tel que

$$\hat{\beta} = \lambda \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)$$

6. Montrer que l'on peut exprimer $\hat{\beta}_0$ de la façon suivante :

$$\hat{\beta}_0 = \frac{n_1}{n} - \left(\frac{n_1}{n} \hat{\mu}_1 + \frac{n_0}{n} \hat{\mu}_0 \right) \hat{\beta}.$$

On utilise comme règle de classification par la méthode des moindres carrés la règle suivante : si $\hat{\beta}_0 + \hat{\beta}^t x > 1/2$ alors y est classifié comme valant 1, sinon comme valant 0. Montrer alors que la règle de classification par méthode des moindres carrés est équivalente à la règle de classification par la méthode LDA présentée en 1., si et seulement si $n_1/n = n_0/n = 1/2$.