Classification d'images de chats et de chiens.

Gautier Appert
gautier.appert.chess@gmail.com

Soit $\mathcal{D}_n = \{x_1, \dots, x_n\}$ une base de données d'images de chat et de chien où chaque image est représentée par un vecteur de pixels $x_i \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$. Notons que le nombre de pixels p est potentiellement très largement supérieur à n. On note $Y_i \in \{0, 1\}$ la variable aléatoire correspondant au label chat ou chien associé à l'image $x_i \in \mathbb{R}^p$. En pratique la base données $\mathcal{D}_n$ stock les images en vecteur lignes

$$\mathcal{D}_n = \left[\begin{array}{c|c|c|c} x_1 & x_2 & \dots & x_n \end{array} \right]^\top.$$

L'objet de ce tutoriel est de prédire le label chat ou chien à l'aide d'une analyse discriminante quadratique (QDA) pour une nouvelle image en dehors de la base d'apprentissage $x \notin \mathcal{D}_n$. L'implémentation doit être faite sous le langage `R`. Envoyer un mail au chargé de TD afin de pouvoir récupérer les données sur la dropbox.

1. Découverte de la base de données et réduction de la dimension

Deux bases de données intitulées X_{train}.RData et X_{test}.RData sont à disposition sur la dropbox. Ces bases contiennent des images de chiens et de chats stockées en vecteur ligne. En particulier on a $X_{\text{train}} \in \mathbb{R}^{315 \times 40000}$ et $X_{\text{test}} \in \mathbb{R}^{48 \times 40000}$. Deux autres bases de données Y_{train}.RData et Y_{test}.RData contiennent les labels associés au images X_{train} et X_{test}.

Question (1). Importer la base de données X_{train}.RData, X_{test}.RData, y_{train}.RData et y_{test}.RData à l'aide de la fonction `load`. Afficher deux à trois images des deux bases à l'aide de la fonction `image(...,col = grey(seq(0, 1, length = 256)))` en transformant préalablement les images en matrice de dimension 200×200 (fonction `matrix`). Enregistrer les images en format `pdf` ou `png` et les mettre dans votre rapport. A quoi correspond le label $y = 1$?

Question (2). On souhaite réduire la dimension des données $p = 40000$. Pour cela nous allons procéder à une analyse en composante principales (ACP) des images.

(a). Concatener X_{train} et X_{test} en utilisant la fonction `rbind` et centrer les vecteurs colonnes avec la fonction `scale`. On notera $X \in \mathbb{R}^{363 \times 40000}$ la matrice résultante. Construire une ACP en utilisant une décomposition en valeur singulière (SVD) de la matrice X à l'aide la fonction `svd`. On ne retiendra que les 15 premières composantes principales. On rappelle que la décomposition en valeur singulière permet de factoriser la matrice X de la manière suivante

$$X = UDV^\top,$$

où V est la matrice des vecteurs propres. Ainsi la matrice des composantes principales est donnée par $C = XV$.

(b). Quelle est la part de variance expliquée en ne retenant que 15 composantes principales ?

Désormais nous travaillerons sur les composantes principales $C \in \mathbb{R}^{363\times 15}$ au lieu des données d'origine X. Découper la base C en $C_{\text{train}} \in \mathbb{R}^{315\times 15}$ et $C_{\text{test}} \in \mathbb{R}^{48\times 15}$.

2. Analyse discriminante quadratique

On fait l'hypothèse du modèle suivant

- $Y_i \sim \mathcal{B}(\pi)$.
- $\mathbb{P}_{c_i|Y=1} = \mathcal{N}(\mu_1, \Sigma_1)$ et $\mathbb{P}_{c_i|Y=0} = \mathcal{N}(\mu_0, \Sigma_0)$ où c_i est le i-ième vecteur ligne de la matrice C.

Le paramètre inconnu est $\theta = (\pi, \mu_0, \mu_1, \Sigma_1, \Sigma_0)$ où $\pi \in]0,1[$, $(\mu_0, \mu_1) \in \mathbb{R}^{15} \times \mathbb{R}^{15}$ et $(\Sigma_0, \Sigma_1) \in \mathbb{R}^{15\times 15} \times \mathbb{R}^{15\times 15}$ sont des matrices définies positives. On definit $\mathbb{P}_\theta = \mathbb{P}_{c,Y}$ et on dispose d'un échantillon $(c_1, y_1), \dots, (c_n, y_n) \overset{\text{i.i.d}}{\sim} \mathbb{P}_\theta$.

Question (3). Ecrire le modèle statistique associé aux observations $(c_1, y_1), \dots, (c_n, y_n)$.

Question (4). On pose $N_1 = \sum_{i=1}^n y_i$ et $N_2 = n - N_1$. En utilisant le fait que $f_{c,Y}(c, y) = f_{c|Y=y}(c)\, f_Y(y)$, montrer que la log vraisemblance $\ell\big((c_1, y_1), \dots, (c_n, y_n); \theta\big)$ s'écrit

$$\ell\big((c_1, y_1), \dots, (c_n, y_n); \theta\big) = N_1 \log(\pi) + N_2 \log(1-\pi)$$
$$-\frac{N_1}{2}\log(\det(\Sigma_1)) - \frac{1}{2}\sum_{i:y_i=1} (c_i - \mu_1)^\top \Sigma_1^{-1}(c_i - \mu_1) - \frac{N_2}{2}\log(\det(\Sigma_0)) - \frac{1}{2}\sum_{i:y_i=0} (c_i - \mu_0)^\top \Sigma_0^{-1}(c_i - \mu_0).$$

Question (5). En utilisant les formules $\nabla_\Sigma \log(\det(\Sigma)) = \Sigma^{-1}$ et $\nabla_\Sigma\big(a^\top \Sigma^{-1} b\big) = -\Sigma^{-1} a b^\top \Sigma^{-1}$, écrire l'équation du premier ordre pour le maximum de vraisemblance et montrer que l'on obtient les estimateurs

$$\hat{\pi} = \frac{N_1}{n} \quad \hat{\mu}_1 = \frac{1}{N_1}\sum_{i:y_i=1} c_i \quad \hat{\mu}_0 = \frac{1}{N_0}\sum_{i:y_i=0} c_i$$

$$\widehat{\Sigma}_1 = \frac{1}{N_1}\sum_{i:y_i=1} (c_i - \hat{\mu}_1)(c_i - \hat{\mu}_1)^\top \quad \widehat{\Sigma}_0 = \frac{1}{N_0}\sum_{i:y_i=0} (c_i - \hat{\mu}_0)(c_i - \hat{\mu}_0)^\top.$$

Question (6). Montrer que la sous Hessienne $\nabla^2_{\pi,\mu_1,\mu_0}\ell(\theta)$ est bien définie négative. On ne regardera pas les conditions du second ordre avec Σ_1 et Σ_0.

Question (7). Montrer que $\hat{\pi}$ est sans biais et montrer que $\hat{\mu}_1$ et $\hat{\mu}_0$ sont sans biais (conditionner par rapport à l'échantillon $\{y_1, \dots, y_n\}$ via la loi des espérances itérées.)

Question (8). Montrer que les estimateurs issus de la méthode des moments coïncident avec les estimateurs du maximum de vraisemblance. (on pourra utiliser la définition de l'espérance conditionelle sachant un évenement $\mathbb{E}\big[C|Y=y\big] = \frac{\mathbb{E}[C\mathbb{1}(Y=y)]}{\mathbb{E}[\mathbb{1}(Y=y)]}$).

Question (9). Coder une fonction sous `R` intitulée `computeML(C,Y)` prenant en argument une matrice C et un vecteur Y, et qui renvoie sous forme de liste les estimateurs du maximum de vraisemblance $\hat{\pi}, \hat{\mu}_1, \hat{\mu}_0, \widehat{\Sigma}_1, \widehat{\Sigma}_0$. Lancer la fonction `computeML` sur `Ctrain,Ytrain`. Comparer les estimateurs obtenus avec la fonction `qda(Ctrain,Ytrain)` du package `MASS`. (La fonction `qda` ne fournit pas les estimateurs concernant les matrices de variances covariances mais fournit le log du determinant).

3. Prédiction des labels sur la base test

On souhaite dans cette partie prédire les labels correspondant au données C_{test} à l'aide de l'analyse discriminante quadratique dont les paramètres ont été estimés sur l'échantillon d'apprentissage `(Ctrain, Ytrain)`. En effet, l'analyse discriminante quadratique permet de modéliser les probabilités $\mathbb{P}\big(Y=1|c\big)$ et $\mathbb{P}\big(Y=0|c\big)$. C'est pourquoi nous prendrons la règle de prédiction suivante $\hat{y} = \arg\max_{y\in\{0,1\}} \mathbb{P}\big(Y=y|c\big)$.

Question (10). A l'aide de la formule de Bayes, montrer que

$$\mathbb{P}\big(Y=1|c\big) = \frac{\pi\,\varphi(c;\mu_1,\Sigma_1)}{\pi\,\varphi(c;\mu_1,\Sigma_1) + (1-\pi)\,\varphi(c;\mu_0,\Sigma_0)}$$

où $\varphi(c;\mu,\Sigma)$ représente la densité de la Gaussienne multivariée $\mathcal{N}\big(\mu,\Sigma\big)$. Calculer de la même manière $\mathbb{P}\big(Y=0|c\big)$.

Question (11). En déduire que

$$\log\left(\frac{\mathbb{P}\big(Y=1|c\big)}{\mathbb{P}\big(Y=0|c\big)}\right) = -\frac{1}{2}\log(\det(\Sigma_1)) - \frac{1}{2}(c-\mu_1)^\top\Sigma_1^{-1}(c-\mu_1) + \log(\pi)$$
$$+\frac{1}{2}\log(\det(\Sigma_0)) + \frac{1}{2}(c-\mu_0)^\top\Sigma_0^{-1}(c-\mu_0) - \log(1-\pi).$$

Question (12). Montrer que

$$\mathbb{1}\left(\log\left(\frac{\mathbb{P}\big(Y=1|c\big)}{\mathbb{P}\big(Y=0|c\big)}\right) > 0\right) = \underset{y\in\{0,1\}}{\arg\max}\ \mathbb{P}\big(Y=y|c\big).$$

Ainsi, on utilisera la règle de prédiction suivante (méthode Plug-in): pour toutes lignes $c \in C_{\text{test}}$

$$\hat{y} = \mathbb{1}\bigg(-\frac{1}{2}\log(\det(\widehat{\Sigma}_1)) - \frac{1}{2}(c-\hat{\mu}_1)^\top\widehat{\Sigma}_1^{-1}(c-\hat{\mu}_1) + \log(\hat{\pi})$$
$$+\frac{1}{2}\log(\det(\widehat{\Sigma}_0)) + \frac{1}{2}(c-\hat{\mu}_0)^\top\widehat{\Sigma}_0^{-1}(c-\hat{\mu}_0) - \log(1-\hat{\pi}) > 0\bigg).$$

Question (13). En utilisant le fait que $(y,\hat{y}) \in \{0,1\}^2$, montrer la double égalité

$$\mathbb{E}\big[(y-\hat{y})^2\big] = \mathbb{E}\big[|y-\hat{y}|\big] = \mathbb{P}\big(\hat{y}\neq y\big).$$

Empiriquement on prendra

$$\widehat{\mathbb{E}}\big[|y-\hat{y}|\big] = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|,$$

ce qui correspond à l'erreur de classification.

Question (14). Ecrire une fonction `R` intitulée `computeLogRatio(c,pi,mu1,mu0,Sigma1,Sigma0)` prenant en argument un vecteur $c \in C_{\text{test}}$ et le paramètre θ, et qui renvoie la quantité: $\log\left(\frac{\mathbb{P}\left(Y=1|c\right)}{\mathbb{P}\left(Y=0|c\right)}\right)$. Puis coder une fonction `computePred(C,pi,mu1,mu0,Sigma1,Sigma0)` prenant en argument une matrice C et le paramètre θ et qui renvoie la prédiction des labels pour chaque ligne de la matrice C.

Question (15). Prédire les labels de la base de données test C_{test} avec `computePred` et donner l'erreur de classification à l'aide de Y_{test}. Comparer avec la prédiction en utilisant l'estimation du modèle faite avec la fonction `qda` de `R`. La prédiction est-elle meilleur que le prédicteur aléatoire ?

Référence

- ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***, *Trevor Hastie, Robert Tibshirani, Jerome Friedman.*