Régression linéaire robuste

Gautier Appert gautier.appert@ensae.fr

On considère le modèle de régression linéaire robuste suivant

$$y_i = x_i^\top \beta + \sigma \varepsilon_i, \qquad i \in \{1, \dots, n\},$$

avec $\varepsilon_i \overset{i.i.d.}{\sim} t_\nu$. Les vecteurs $x_i \in \mathbb{R}^p$ sont supposés déterministes, et le degré de liberté $\nu > 0$ est supposé connu. Le paramètre d'intérêt du modèle est $\theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^*$. Dans ce modèle les variables dépendantes $\{y_i\}_{i=1}^n$ suivent donc une loi de Student décentrée avec un paramètre d'échelle $\sigma > 0$, autrement dit pour tout $i \in \{1, \dots, n\}$, $y_i \sim t_\nu(x_i^\top \beta, \sigma^2)$. On obtient donc la densité suivante pour chaque y_i

$$f_\theta(y_i) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y_i - x_i^\top \beta)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

où $\Gamma(a) \overset{def}{=} \int_0^{+\infty} x^{a-1} \exp\{-x\} \, \mathrm{d}x$, pour tout $a > 0$.

Préliminaires

Question (1). La loi de Student correspond-elle à un modèle exponentiel ? (justifier). Soit $X \sim \mathcal{N}(0,1)$ et $Z \sim \chi^2(\nu)$ tel que $X \perp\!\!\!\perp Z_\nu$. La Student centrée T_ν est définie par $T_\nu = X/\sqrt{Z_\nu/\nu}$. A l'aide de l'inégalité de Tchebychev, montrer que $Z_\nu/\nu \xrightarrow[\nu\to+\infty]{\mathbb{P}} 1$. En déduire que $T_\nu \xrightarrow[\nu\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$.

Question (2). Ecrire la vraisemblance associée au modèle linéaire robuste et dériver les équations de vraisemblance. Peut-on obtenir des estimateurs du maximum de vraisemblance sous forme explicite ? *Bonus:* Montrer que la vraisemblance n'est pas concave globalement. *Notons que la méthode de Newton Raphson n'est pas appropriée lorsque la vraisemblance n'est pas globalement concave.*

Modèle de mélange et Algorithme EM

Il est possible de représenter une loi de Student comme un mélange d'une loi normale, avec comme loi mélangeante la distribution du $\chi^2(\nu)$

$$y|z \sim \mathcal{N}\left(x^\top \beta, \frac{\nu\sigma^2}{z}\right), \qquad z \sim \chi^2(\nu) \equiv \gamma\left(\frac{\nu}{2}, \frac{1}{2}\right).$$

La variable aléatoire z joue ici le rôle de la variable latente. L'objectif de cette partie est de déterminer un estimateur $\hat{\theta}$ à l'aide de l'algorithme d'Espérance Maximisation (EM) en se basant sur cette représentation d'une loi de Student vue comme un mélange. *On suppose jusqu'à la question (5) que l'on dispose seulement d'une seule variable aléatoire y associée à la variable latente z.*

Question (3). Justifier la représentation de la Student $t_\nu(x^\top \beta, \sigma^2)$ vu comme le mélange décrit précédement. De plus, montrer que $z|y \sim \gamma\left(\frac{\nu+1}{2}, \frac{(y-x^\top\beta)^2+\nu\sigma^2}{2\nu\sigma^2}\right)$.

Question (4). On définit la vraisemblance complète comme étant la vraisemblance augmentée de la variable latente $L(y, z; \theta)$. En utilisant le fait que $f_\theta(z|y) = \frac{L(y,z;\theta)}{L(y;\theta)}$, montrer que pour tout θ_0 fixé

$$\log L(y;\theta) = \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log L(y,z;\theta)\Big] - \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log f_\theta(z|y)\Big].$$

QUESTION (5). On pose la fonction $Q(\theta;\theta_0,y) \overset{def}{=} \mathbb{E}_{z\sim f_{\theta_0}(z|y)}\Big[\log L(y,z;\theta)\Big]$. L'algorithme EM consiste à maximer cette fonction en θ, afin d'obtenir un premier estimateur $\hat{\theta}_1$. En remplaçant $\theta_0 \leftarrow \hat{\theta}_1$, et en maximisant à nouveau la fonction $Q(\theta;\theta_1,y)$, l'algorithme crée de manière itérative une séquence d'estimateurs $(\hat{\theta}_k)_{k\geq 1}$. Montrer que la vraisemblance augmente à chaque itération de l'algorithme EM, ie $L(y;\hat{\theta}_{k+1}) \geq L(y;\hat{\theta}_k)$.

QUESTION (6). On dispose désormais d'un échantillon de variables aléatoires $\{y_i\}_{i=1}^n$ associées aux variables latentes $\{z_i\}_{i=1}^n$. Ecrire la log vraisemblance complète du modèle $\log L\left(\{(y_i,z_i)\}_{i=1}^n;\theta\right)$ et calculer la fonction $Q(\theta;\theta_0,\{y_i\}_{i=1}^n)$. On pourra utiliser le fait que si $X\sim\gamma(a,b)$ alors $\mathbb{E}[X]=a/b$ et $\mathbb{E}[\log(X)]=\Psi(a)-\log(b)$ où $\Psi(x)=\Gamma'(x)/\Gamma(x)$.

QUESTION (7). Ecrire les conditions du premier ordre concernant la fonction $Q\left(\theta;\theta_0,\{y_i\}_{i=1}^n\right)$ et montrer que les équations de récurrence liées à la maximisation sont données pour tout $k\geq 0$ par

$$\hat{\beta}_{k+1} = \left(\sum_{i=1}^n \frac{x_i x_i^\top}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}\right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}$$
$$\hat{\sigma}_{k+1}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(\nu+1)\hat{\sigma}_k^2(y_i - x_i^\top\hat{\beta}_{k+1})^2}{(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2}$$

QUESTION (8). *Question Bonus:* On pose la matrice de poid $W_k = \mathbf{diag}\left(1/[(y_i - x_i^\top\hat{\beta}_k)^2 + \nu\hat{\sigma}_k^2]\right)$ ainsi que la matrice $X = [x_1|\dots|x_n]^\top$. Réécrire les équations de récurrence précédentes en utilisant la matrice de poid W_k et la matrice X. A quelle type d'estimateur vu en cours d'Econométrie $\hat{\beta}_{k+1}$ correspond t'il ?

PROGRAMMATION AVEC R

On souhaite dans cette section appliquer l'algoritme EM sur le jeu de données `election` issu du package `LearnBayes` sous `R`. Ce jeu de données contient en particulier le nombre de votes lors des élections présidentielles américaines en 2000 pour le condidat *Pat Buchanan* et le nombre de votes en 1996 pour le candidat *Ross Perot*, dans chacun des 67 comtés de Floride *. On veut estimer un modèle linéaire simple $y_i = \beta_0 + \beta_1 x_i + \sigma\varepsilon_i$, où y_i représente la racine carré du nombre de votes pour le candidat Pat Buchanan dans le comté numéro i et x_i la racine carré du nombre de votes pour le candidat Ross Perot.

QUESTION (9). Faire un plot des données $\{y_i\}_{i=1}^n$ en fonction des $\{x_i\}_{i=1}^n$. Justifier l'utilisation d'un modèle linéaire simple robuste.

QUESTION (10). Estimer un modèle linéaire simple robuste à l'aide de l'algorithme EM en prenant plusieurs valeurs de ν, $\nu = 1; 10; 100$ (*on peut prendre un critère d'arrêt basé sur la différence des log vraisemblances prises entre deux itérations successives*). Prendre plusieurs initialisations différentes pour le paramètre $\theta = (\beta, \sigma^2)$. Faire un plot de l'évolution de la log vraisemblance à chaque itération.

QUESTION (11). Estimer le modèle par Moindres carrées ordinaires (MCO), et afficher sur le graphique de la question (9) les droites de régressions issues de l'estimation par MCO et de l'algorithme EM. Conclure.

*Variables intitulées `Buchanan` et `Perot` dans le jeu de données `election`.

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

$$\mathcal{D}_n = \left[\; x_1 \; \middle| \; x_2 \; \middle| \; \dots \; \middle| \; x_n \; \right]^\top.$$

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

- $Y_i \sim \mathcal{B}\big(\pi\big)$.
- $\mathbb{P}_{c_i|Y=1} = \mathcal{N}\big(\mu_1, \Sigma_1\big)$ et $\mathbb{P}_{c_i|Y=0} = \mathcal{N}\big(\mu_0, \Sigma_0\big)$ 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 $(c_1, y_1), \ldots, (c_n, y_n) \overset{\text{i.i.d}}{\sim} \mathbb{P}_\theta$.

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

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 $\ell\big((c_1, y_1), \ldots, (c_n, y_n); \theta\big)$ s'écrit

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

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 $\{y_1, \ldots, y_n\}$ 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.*

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 $x_i \in \mathbb{R}^p$ est représentée par un vecteur $x_i \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$. Notons que 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 déterministe $x_i \in \mathbb{R}^p$. En pratique la base données $\mathcal{D}_n$ stock les images en vecteur lignes

$$\mathcal{D}_n = \left[\; x_1 \; \middle| \; x_2 \; \middle| \; \dots \; \middle| \; x_n \; \right]^\top.$$

L'objet de ce tutoriel est de prédire le label chat ou chien à l'aide d'un modèle logistique 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(.)`). Enregister 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 30 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 30 composantes principales ?

Désormais nous travaillerons sur les composantes principales $C \in \mathbb{R}^{363\times 30}$ au lieu des données d'origine. Découper la base C en $C_{\text{train}} \in \mathbb{R}^{315\times 30}$ et $C_{\text{test}} \in \mathbb{R}^{48\times 30}$.

2. Modèle logistique et pénalité Ridge

On suppose que les données $\{Y_i\}_{i=1}^n$ suivent un modèle logistique, c'est à dire

- $Y_i \sim \mathcal{B}\big(\pi(c_i)\big)$ avec $\pi(c_i) = \mathbb{P}\big(Y_i = 1\big)$, où c_i est le vecteur ligne de la matrice C.
- $\log\left(\frac{\pi(c_i)}{1-\pi(c_i)}\right) = c_i^\top \beta, \quad i \in \{1,\dots,n\}$.

Les Y_i sont indépendants.

Question (3). Montrer que la log vraisemblance $\ell(Y_1,\dots,Y_n;\beta)$ s'écrit

$$\ell(Y_1,\dots,Y_n;\beta) = \sum_{i=1}^n y_i c_i^\top \beta - \sum_{i=1}^n \log(1+e^{c_i^\top \beta}).$$

Question (4). Au lieu de maximiser la vraisemblance on va ajouté une pénalité Ridge sur les coefficients $\{\beta_j\}_{j\in\{1,\dots,p\}}$, ce qui aura pour effet de rétrécir les coefficients. On préfèrera donc maximiser la quantité

$$J(\beta;\lambda) \stackrel{\text{def}}{=} \ell(Y_1,\dots,Y_n;\beta) - \lambda\|\beta\|_2^2,$$

où $\lambda \in \mathbb{R}_+$ est un hyperparamètre à calibrer. Il est possible de faire un lien entre la vraisemblance pénalisée $J(\beta;\lambda)$ et la formulation Bayésienne. En effet, montrer que si on pose un prior $\pi(\beta) \propto \exp\{-\lambda\|\beta\|_2^2\}$ alors la loi a posteriori s'écrit

$$\pi(\beta|Y_1,\dots,Y_n) \propto \exp\{J(\beta;\lambda)\}.$$

Question (5). Montrer que le gradient $\nabla_\beta J(\beta;\lambda)$ et la Hessienne $\nabla^2_\beta J(\beta;\lambda)$ sont donnés par

- $\nabla_\beta J(\beta;\lambda) = \sum_{i=1}^n y_i c_i - \sum_{i=1}^n \frac{e^{c_i^\top\beta}}{1+e^{c_i^\top\beta}}\, c_i - 2\lambda\beta,$
- $\nabla^2_\beta(\beta;\lambda) = -\sum_{i=1}^n \frac{e^{c_i^\top\beta}}{(1+e^{c_i^\top\beta})^2}\, c_i c_i^\top - 2\lambda I,$

où I est la matrice identité. On pose $\pi \in \mathbb{R}^n$ le vecteur des probabilités $\{\pi(c_i)\}_{i\in\{1,\dots,n\}}$ et on pose la matrice de poid $W = \mathbf{diag}\Big(\pi(c_1)(1-\pi(c_1)),\dots,\pi(c_n)(1-\pi(c_n)\Big)$. Montrer que le gradient et la hessienne se réécrivent sous la forme

- $\nabla_\beta J(\beta;\lambda) = C^\top(y-\pi) - 2\lambda\beta,$
- $\nabla^2_\beta(\beta;\lambda) = -C^\top W C - 2\lambda I.$

Montrer que les équations de récurrences liées à la maximisation pour l'algorithme de Newton-Raphson sont données par

$$\beta^{new} = (C^\top W C + 2\lambda I)^{-1} C^\top W z,$$

où $z = W^{-1}(y-\pi) + C\beta^{old}$. A quelle type d'estimateur vu en TD β^{new} vous fait t'il penser ? Ecrire l'agorithme de Newton-Raphson en pseudocode pour maximiser $J(\beta;\lambda)$ avec un λ fixé.

Question (6). Coder une fonction sous `R` qui calcule le maximum de vraisemblance pénalisé $\arg\max J(\beta, \lambda)$ en utilisant l'algorithme de Newton-Raphson. La condition d'arrêt peut être basée sur la stabilité de la vraisemblance et l'hyperparamètre λ sera un argument de la fonction.

3. Validation croisée et prédiction des labels sur la base test

On souhaite dans cette partie calibrer le paramètre λ sur la base de données d'entrainement C_{train} par validation croisée et prédire les labels correspondant au données C_{test} à l'aide de l'estimation du modèle logistique pénalisée. La validation croisée consiste à découper la base C_{train} en K parties aléatoires (environ égales) puis à selectionner $K-1$ parties pour constituer un nouveau échantillon d'apprentissage sur lequel on estime le modèle logistique pénalisé sur une plage de valeur $\lambda \in \Lambda$. Puis on calcule l'erreur quadratique moyenne de prédiction des y_i en fonction de $\lambda \in \Lambda$ sur la dernière partie de la base (échantillon de validation) qui n'a pas été sélectionné pour l'estimation du modèle logistique. On répète cette prodédure K fois en prenant à chaque fois un autre échantillon de validation. On peut ainsi faire une moyenne des K erreurs quadratiques moyennes et selectionner le $\lambda \in \Lambda$ qui la minimise.

Notons que le modèle logistique permet de prédire des probabilités $\pi(c_i)$ et non pas les labels y_i directement. C'est pourquoi nous prendrons la règle de décision suivante $\hat{y}_i = \mathbb{1}\big(\hat{\pi}(c_i) \geq 0.5\big)$.

Question (7). On note $\hat{Y}_i$ la prédiction faite par le modèle. En utilisant le fait que $(y_i, \hat{Y}_i) \in \{0,1\}^2$, montrer la double égalité

$$\text{MSE}(\hat{Y}_i) = \mathbb{E}\big[|\hat{Y}_i - y_i|\big] = \mathbb{P}\big(\hat{Y}_i \neq y_i\big).$$

Empiriquement on prendra

$$\widehat{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|,$$

ce qui correspond à l'erreur de classification.

Question (8). Calculer λ par validation croisée en prenant $\Lambda = \{0, 100, 1000, 2000, 10000, 15000, 45000\}$ et $K = 15$, c'est à dire calculer

$$\lambda_{CV} = \arg\min_{\lambda \in \Lambda} \frac{1}{K}\sum_{k=1}^{K} \widehat{\text{MSE}}_{\lambda,k}.$$

On pourra utiliser `createFolds(.)` du package `caret` pour la création des K échantillons. Faut t'il sélectionner le modèle logistique sans pénalité ?

Question (9). Réestimer le modèle logistique sur toute la base de données C_{train} avec λ_{CV} calculé précédement, puis prédire les labels de la base de données test C_{test}. Donner l'erreur de prédiction à l'aide de Y_{test} . La prédiction est telle meilleur que le prédicteur aléatoire ?

Référence

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