Télécom ParisTech, CES Data scientist : TP d'évaluation du module 1
A. Gramfort / A. Sabourin / J. Salmon

Régression linéaire et classification

Format demandé : Rédiger les réponses aux questions suivantes et rendre le travail sous forme d'un document pdf (latex conseillé mais non obligatoire). Les questions numériques doivent être traitées en Python et le code utilisé pour générer les résultats doit être fourni dans un fichier `.py` à part, avec commentaires facilitant sa lecture. Les résultats doivent être reproductibles, c'est-à-dire : le code doit être exécutable par le correcteur. Les résultats numériques et graphiques doivent être inclus dans le fichier pdf. Le tout, rassemblé dans une archive, est à rendre pour le 12 avril 2015, 23h59.

- Classification binaire par moindre carrés -

On considère deux populations gaussiennes dans $\mathbb{R}^p$. On observe des points générés par un mélange de ces deux populations.

Les lois conditionnelles de X sachant $Y = +1$ (respectivement $Y = -1$) sont des gaussiennes multivariées $\mathcal{N}_p(\mu_+, \Sigma_+)$ (respectivement $\mathcal{N}_p(\mu_-, \Sigma_-)$). On notera leur densités respectives f_+ et f_-. Les vecteurs μ_+ et μ_- sont dans $\mathbb{R}^p$ et les matrices Σ_+, Σ_- sont symétriques de taille $p \times p$. On note également $\pi_+ = \mathbb{P}\{Y = +1\}$. On tire avec probabilité π_+ une étiquette $Y = +1$ ou $Y = -1$, qui indique si X est tiré selon la loi f_+ ou f_-. La densité du mélange est donc

$$f(\mathbf{x}) = \pi_+ f_+(\mathbf{x}) + \pi_- f_-(\mathbf{x})$$

On rappelle que la densité p-dimensionnelle de la loi $\mathcal{N}_p(\mu_i, \Sigma_i)$ est donnée par :

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma_i)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^\top \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\} .$$

et que la matrice de covariance d'un vecteur aléatoire X est définie par $\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top\big]$.

1. Générer un jeu de données simulées selon le modèle de mélange précédent. On prendra comme valeurs numériques : $\pi_+ = 0.5, p = 2, n = 500, \mu_- = (-1,-1), \mu_+ = (1,1), \Sigma_+ = 3\,\mathrm{Id}_p, \Sigma_- = 2\,\mathrm{Id}_p$ Détailler le pseudo-code (l'algorithme) utilisé. Afficher deux exemples de jeux de données de sorte que le label de chaque point soit visible (en s'inspirant éventuellement de la fonction `plot_2d` du TP d'introduction à la classification).
2. On suppose que l'échantillon considéré contient n observations notées $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, où $\mathbf{x}_i \in \mathbb{R}^p$ et $y_i \in \{-1, 1\}$ pour $1 \leq i \leq n$. Le nombre d'observations dans la classe « $y = +1$ » est $\sum_{i=1}^n \mathbb{1}\{y_i = +1\} = n_+$. On note $n_- = \sum_{i=1}^n \mathbb{1}\{y_i = -1\} = n - n_+$.
   En utilisant par exemple la méthode des moments (*i.e.,* on remplace les espérances par leurs contreparties empiriques), proposer des estimateurs non biaisés $\widehat{\pi}_+$, $\widehat{\mu}_+$, $\widehat{\mu}_-$ des paramètres π_+, μ_+, μ_-. Montrer que l'espérance théorique de X est $\mu = \pi_+\mu_+ + \pi_-\mu_-$.
3. Sur le même principe, donner des estimateur $\widehat{\Sigma}_+, \widehat{\Sigma}_-$ des variances au sein de chaque classe. Ces estimateurs sont-ils biaisés ?

On se propose d'étudier un classifieur (*i.e.,* une fonction qui à tout point associe l'étiquette prédite) obtenu en deux étapes :

(i) On résout un problème de minimisation d'un critère des moindres carrés, de solution $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_0, \widehat{\theta}_{1:p}) \in \mathbb{R}^{p+1}$, c'est à dire

$$\widehat{\boldsymbol{\theta}} = \underset{\theta_0 \in \mathbb{R}, \theta_{1:p} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^n (y_i - \theta_0 - \theta_{1:p}^\top \mathbf{x}_i)^2 .$$

(ii) On définit un classifieur $\widehat{y}(\mathbf{x})$ de la forme suivante :

$$\widehat{y}(\mathbf{x}) = \mathrm{sign}\left(\widehat{\theta}_0 + \widehat{\theta}_{1:p}^\top \mathbf{x}\right) ,$$

où la fonction sign est définie par

$$\text{sign}(z) = \begin{cases} 1 & \text{si } z > 0, \\ -1 & \text{si } z < 0, \\ 0 & \text{sinon.} \end{cases}$$

4. Construire numériquement le classifieur ci-dessus, et l'appliquer aux jeu de données simulées. Détailler le pseudo-code. Afficher le résultat en faisant apparaître les zones du plan correspondant à chaque classe, en même temps que les données d'apprentissage. On pourra par exemple s'inspirer de la fonction `frontiere` du TP d'introduction à la classification.
5. Écrire la fonction de coût $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_{1:p}^\top \mathbf{x}_i)^2$ sous forme matricielle, en faisant intervenir les matrices

$$Z = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \; ; \; \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} \; ; \; Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} .$$

   Ecrire la condition d'annulation du gradient sous forme d'une équation matricielle faisant intervenir les matrices précédentes.
6. Montrer que la condition d'annulation précédente s'écrit sous la forme

$$\left(\begin{array}{c|c} n & n\hat{\mu}^\top \\ \hline n\hat{\mu} & \mathbb{X}^\top \mathbb{X} \end{array}\right) \left(\begin{array}{c} \widehat{\theta}_0 \\ \hline \widehat{\theta}_{1:p} \end{array}\right) = \left(\begin{array}{c} n_+ - n_- \\ \hline n_+\widehat{\mu}_+ - n_-\widehat{\mu}_- \end{array}\right)$$

   où $\mathbb{X}$ est la matrice des données

$$\mathbb{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}$$

   et $\hat{\mu} = \frac{n_+}{n}\hat{\mu}_+ + \frac{n_-}{n}\hat{\mu}_-$.
7. On rappelle que $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^\top\mathbb{X} - \hat{\mu}\hat{\mu}^\top = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \mu)^\top$ est un estimateur classique de la variance totale du jeu de données. Mettre en évidence le fait que la solution $\widehat{\boldsymbol{\theta}}$ doive satisfaire une équation de la forme :

$$\widehat{\Sigma}\,\widehat{\boldsymbol{\theta}}_{1:p} = 2\frac{n_+ n_-}{n^2}(\widehat{\mu}_+ - \widehat{\mu}_-)$$

8. Montrer que l'estimateur de variance totale $\widehat{\Sigma}$ de la question précédente peut s'écrire

$$\widehat{\Sigma} = \widehat{S} + \alpha\widehat{\Sigma}_B$$

   où $\widehat{\Sigma}_B = (\widehat{\mu}_+ - \widehat{\mu}_-)(\widehat{\mu}_+ - \widehat{\mu}_-)^\top$ (estimateur de la variance inter-classe), et $\widehat{S} = \frac{n_+}{n}\widehat{\Sigma}_+ + \frac{n_-}{n}\widehat{\Sigma}_-$ (estimateur de la variance intra-classe), avec α un facteur dépendant de n, n_+, n_- à préciser.
9. Montrer alors que $\widehat{\Sigma}_B\widehat{\boldsymbol{\theta}}_{1:p}$ est porté par la direction $(\widehat{\mu}_+ - \widehat{\mu}_-)$. En déduire que $\widehat{\boldsymbol{\theta}}_{1:p}$ est proportionnel à $\widehat{S}^{-1}(\widehat{\mu}_+ - \widehat{\mu}_-)$, dans le cas où $\hat{S}$ est inversible.
10. Vérifier cette dernière propriété numériquement avec un jeu de données généré à la question 1.