Correction TD 2

Chargé.e.s de TD

March 2020

Exercice 1

Supposons que l'on observe Y un vecteur aléatoire dans $\mathbb{R}^n$ vérifiant

$$\mathbf{Y} = \beta + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et $\beta = (\beta_1, \dots, \beta_n)^\top$ est le paramètre à estimer. Noter qu'il s'agit d'un cas particulier du modèle de régression linéaire $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ où $\mathbf{X} = I_n$. On définit l'estimateur LASSO

$$\hat{\beta} = \arg\min_b \{||\mathbf{Y} - b||_2^2 + \lambda ||b||_1\},$$

l'estimateur *ridge*

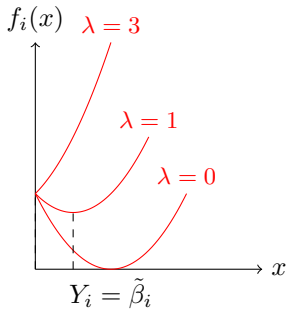
$$\check{\beta} = \arg\min_b \{||\mathbf{Y} - b||_2^2 + \lambda ||b||_2^2\},$$

et l'estimateur des moindres carrés

$$\tilde{\beta} = \arg\min_b \{||\mathbf{Y} - b||_2^2\}.$$

1. Montrer que l'on peut dans ce cas obtenir une formule explicite pour chacune des coordonnées β_i, $1 \leq i \leq n$, de β.

Solution:



Par définition de $\widehat{\beta}$,

$$\left(\widehat{\beta}_i\right)_{1\leq i\leq n} = \arg\min_{b_1,\dots,b_n} \sum_{1\leq i\leq n} (Y_i - b_i)^2 + \lambda |b_i|.$$

On remarque que la fonction objectif est séparable, et donc pour tout $i = 1, ..., n$,

$$\widehat{\beta}_i = \arg\min_{b_i} (Y_i - b_i)^2 + \lambda |b_i|.$$

Notons $f_i : x \to (Y_i - x)^2 + \lambda |x|$. Cette fonction est dérivable sur $\mathbb{R}^*$, et sa dérivée est donnée par

$$f_i'(x) = \begin{cases} 2x - (2Y_i + \lambda) \text{ si } x < 0 \\ 2x - (2Y_i - \lambda) \text{ si } x > 0 \end{cases}$$

Trois cas se présentent. On rappelle que le paramètre de régularisation λ est strictement positif.

Cas 1 : si $Y_i \leq -\lambda/2$ Alors $2Y_i + \lambda < 0$ et $2Y_i - \lambda < 0$, donc

$$\begin{cases} f_i'(x) < 0 \text{ si } x < (Y_i + \lambda/2) \\ f_i'(x) > 0 \text{ si } x > (Y_i + \lambda/2) \end{cases}$$

Comme f_i est continue, elle atteind son minimum en $x = (Y_i + \lambda/2) = \text{signe}(Y_i)\,(|Y_i| - \lambda/2)_+$.
Cas 2 : si $Y_i \in [-\lambda/2, \lambda/2]$ Alors $2Y_i + \lambda > 0$ et $2Y_i - \lambda < 0$, donc

$$\begin{cases} f_i'(x) < 0 \text{ si } x < 0 \\ f_i'(x) > 0 \text{ si } x > 0 \end{cases}$$

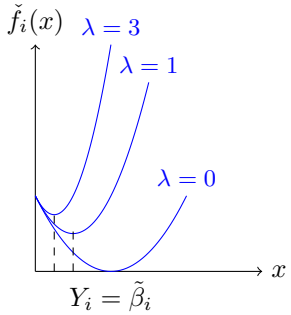
Comme f_i est continue, elle atteind son minimum en $x = 0 = \text{signe}(Y_i)\,(|Y_i| - \lambda/2)_+$.
Cas 3 : si $Y_i \geq \lambda/2$ Alors $2Y_i + \lambda > 0$ et $2Y_i - \lambda > 0$, donc

$$\begin{cases} f_i'(x) < 0 \text{ si } x < Y_i - \lambda/2 \\ f_i'(x) > 0 \text{ si } x > Y_i - \lambda/2 \end{cases}$$

Comme f_i est continue, elle atteind son minimum en $x = (Y_i - \lambda/2) = \text{signe}(Y_i)\,(|Y_i| - \lambda/2)_+$.
Conclusion
On a donc $\widehat{\beta}_i = \text{signe}(Y_i)\,(|Y_i| - \lambda/2)_+$.

2. Même question pour les coordonnées de $\check{\beta}_i$ et $\tilde{\beta}$.

Solution:



Par définition de $\check{\beta}$,

$$\left(\check{\beta}_i\right)_{1\leq i\leq n} = \underset{b_1,\ldots,b_n}{\arg\min} \sum_{1\leq i\leq n} (Y_i - b_i)^2 + \lambda b_i^2.$$

Là encore, la fonction objectif est séparable, et donc pour tout $i = 1, ..., n$,

$$\begin{aligned} \check{\beta}_i &= \underset{b_i}{\arg\min}\, (Y_i - b_i)^2 + \lambda b_i^2 \\ &= \frac{Y_i}{1+\lambda}. \end{aligned}$$

Ce résultat s'obtient facilement en étudiant la dérivée des fonctions $\check{f}_i : x \to (Y_i - x)^2 + \lambda x^2$.
De même, on montre que pour $i = 1, ..., n$, $\tilde{\beta}_i = Y_i$ (il suffit d'appliquer le résultat précédent pour $\lambda = 0$).

Conclusion

L'estimateur de la question 1 est appelé estimateur LASSO. Il est *parcimonieux* (*sparse* en anglais), puisque beaucoup des coefficients $\widehat{\beta}_i$ sont nuls (ceux qui correspondent aux signaux Y_i trop faibles, tels que $|Y_i| \leq \lambda/2$).

Ce phénomène de parcimonie est souhaitable pour éviter le *surapprentissage* (en anglais *overfitting*). Par exemple, un exercice classique montre que la variance de l'estimateur des moindres carrés diminue comme d/n lorsque le modèle comporte d dimensions et l'échantillon n observations. Lorsque le nombre de de variables

explicatives est grand, on a intérêt à réduire la variance de notre estimateur en contraignant un grand nombre de ses paramètre à être nuls, et en réduisant ainsi la taille du modèle.

En général, il n'existe pas de formule explicite pour les coefficients de l'estimateur LASSO, mais on peut efficacement calculer des approximations de cet estimateur.

L'estimateur régularisé correspondant à la question 2 est l'estimateur RIDGE. Contrairement à l'estimateur LASSO, il ne possède pas de coefficient nul (sauf si $Y_i = 0$). Cependant ses coefficients sont plus petits d'un facteur $1/(1+\lambda)$ que ceux de l'estimateur des moindres carrés, ce qui réduit le risque de surapprentissage.

Exercice 2

1. Expliquez à quelle condition sur δ l'hypothèse de "séparation presque linéaire" est satisfaite.

Solution: On rappelle que pour $i = 1, ..., n$, $X_i \in \mathbb{R}^d$ et $Y_i = \pm 1$. Les données $(X_i, Y_i)_{1 \leq i \leq n}$ sont linéairement séparables s'il existe ω de norme 1 et $\rho > 0$ tel que pour tout $i = 1, ..., n$, $Y_i \omega X_i \geq \rho$. L'hyperplan perpendiculaire au vecteur ω sépare les données (Y_i est du signe de ωX_i), et ρ caractérise la marge, c'est à dire la distance des points à l'hyperplan qui les sépare.

On considère un jeu de données presque linéairement séparables, c'est à dire tel qu'il existe ω et $\rho > 0$ séparant presque les données. Pour $i = 1, ..., n$, on définit $d_i = (\rho - Y_i \omega X_i)_+$. Si le couple (X_i, Y_i) fait partie des données "séparables", c'est à dire si $Y_i \omega X_i \geq \rho$, alors $d_i = 0$. Réciproquement, $d_i > 0$ si le couple (X_i, Y_i) ne vérifie pas la condition de séparation (c'est à dire si le signe de Y_i est différent de celui de ωX_i, ou si il est à une distance inférieure à ρ de l'hyperplan séparateur).

Ainsi, l'hypothèse "quasi linéairement séparable" est satisfaite lorsque le nombre de couples "non séparés" par (ω, ρ) est petit, et donc que $\delta = \sqrt{\sum d_i^2} = o(\sqrt{n})$.

On fixe deux paramètres Δ et Z pour l'instant. Pour $X_i \in \mathbb{R}^d$, on définit

$$\tilde{X}_i = (X_i, 0, \ldots, 0, \Delta, 0, \ldots, 0) \in \mathbb{R}^{d+n}$$

où le Δ est en $(d+i)$-eme position. De même, à partir de $\omega \in \mathbb{R}^d$, on définit

$$\tilde{\omega} = (\frac{\omega}{Z}, \frac{Y_1 d_1}{\Delta Z}, \ldots, \frac{Y_n d_n}{\Delta Z}) \in \mathbb{R}^{d+n}$$

2. Que doit valoir Z pour que $\tilde{\omega}$ soit de norme 1 ?

Solution: Par définition de $\widetilde{\omega}$, on a

$$\|\widetilde{\omega}\|^2 = \frac{\|\omega\|^2}{Z^2} + \frac{\sum Y_i^2 d_i^2}{\Delta^2 Z^2}.$$

De plus $\|\omega\| = 1$, $Y_i^2 = 1$, et $\sum d_i^2 = \delta^2$. Ainsi,

$$\|\widetilde{\omega}\|^2 = \frac{\Delta^2 + \delta^2}{\Delta^2 Z^2}.$$

Pour que $\|\widetilde{\omega}\| = 1$, il faut choisir la normalisation $Z = \sqrt{\frac{\Delta^2 + \delta^2}{\Delta^2}}$.

3. Montrez que l'on peut appliquer l'analyse du perceptron vue en cours à $(\tilde{X}_i, Y_i)$ (hint : il suffit de montrer la séparation linéaire des données)

Solution: On montre que les données augmentées $(\widetilde{X}_i, Y_i)_{1 \leq i \leq n}$ sont linéairement séparables, et en particulier qu'elles sont séparables par l'hyperplan perpendiculaire à $\widetilde{\omega}$ et telles que pour tout $i = 1, ..., n$, $Y_i \widetilde{\omega}^T \widetilde{X}_i \geq \widetilde{\rho}$ où $\widetilde{\rho} = \rho / Z$.

En effet, soit (X_i, Y_i) un couple "séparables" pour les paramètres ω, ρ, c'est à dire tel que $d_i = (\rho - Y_i \omega^T X_i)_+ = 0$. Alors

$$Y_i \widetilde{\omega}^T \widetilde{X}_i = \frac{Y_i \omega X_i}{Z} \geq \frac{\rho}{Z},$$

et donc $(\widetilde{X}_i, Y_i)$ est séparable pour les paramètres $\widetilde{\omega}, \widetilde{\rho}$.

De même, soit (X_i, Y_i) un couple "non séparables" pour les paramètres ω, ρ, c'est à dire tel que $d_i = (\rho - Y_i \omega^T X_i)_+ > 0$. Alors

$$Y_i \widetilde{\omega}^T \widetilde{X}_i = \frac{Y_i \omega^T X_i}{Z} + \frac{Y_i^2 d_i \Delta}{\Delta Z} = \frac{Y_i \omega^T X_i}{Z} + \frac{\rho - Y_i \omega^T X_i}{Z} \geq \frac{\rho}{Z},$$

et donc $(\widetilde{X}_i, Y_i)$ est séparable pour les paramètres $\widetilde{\omega}, \widetilde{\rho}$.
Ainsi, les données augmentées $(\widetilde{X}_i, Y_i)_{1 \leq i \leq n}$ sont linéairement séparables par $\widetilde{\omega}$, et de marge $\widetilde{\rho}$

4. Utilisez le résultat du perceptron du cours pour obtenir une borne sur le nombre d'itérations en fonction de Δ.

Solution: Les hypothèses de séparabilité et de marge nécessaires à l'analyse du perceptron étant vérifiées, on applique le résultat du théorème vu en cours. On remarque que pour tout $i = 1, ..., n$, $\left\| \widetilde{X}_i \right\|^2 = \|X_i\|^2 + \Delta^2 \leq r^2 + \Delta^2$. Ainsi, le nombre d'itérations nécessaires pour trouver un hyperplan séparateur est borné par

$$k = \frac{r^2 + \Delta^2}{\rho^2 / Z^2} = \frac{r^2 + \Delta^2}{\rho^2} \times \frac{\Delta^2 + \delta^2}{\Delta^2}.$$

5. Optimisez en fonction de Δ.

Solution: On montre facilement que cette borne est minimale pour le choix $\Delta^2 = r\delta$, et vaut dans ce cas $k = \frac{(r+\delta)^2}{\rho^2}$.
On remarque que le choix optimal de Δ dépend de r (qu'on peut facilement calculer) et de δ (qu'on peut calculer en connaissant ω et ρ, mais dans ce cas a-t-on encore besoin du perceptron ?).