Regression Linéaire

September 9, 2019

Mots clefs. Après avoir lu le cours, ces mots doivent vous être familiés.

- Regression,
- Transformations des variables,
- Données corrélées.

1 Regression linéaire

1.1 Régression linéaire pure

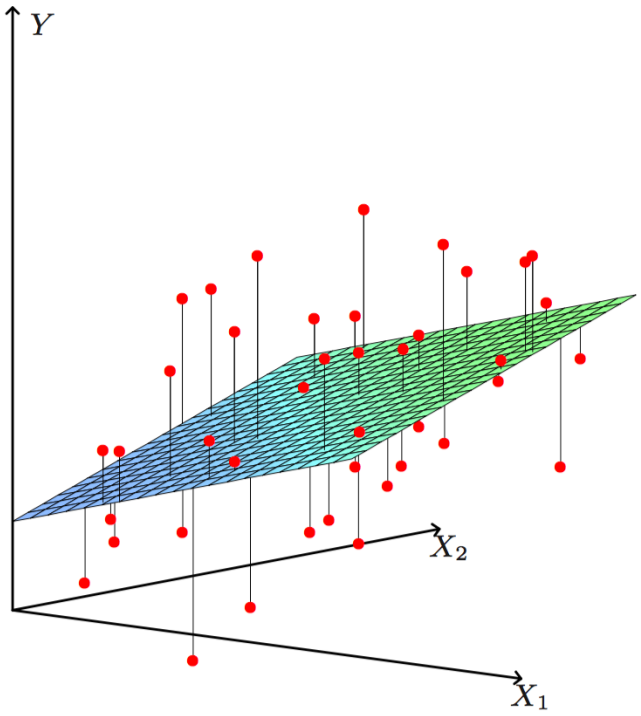
- p-Entrées (= descripteur) quantitives : $X = (X^1, ..., X^p) \in \mathbb{R}^p$.
- Une sortie quantitative : $Y \in \mathbb{R}$.

Nous parions que Y est proche d'une combinaison affine des X^j. On se donne une famille paramétrée

$$\mathcal{F} = \left\{ f_w(x) = w_0 + \sum_{j=1}^{p} w_j x^j \ : \ w \in \mathbb{R}^{p+1} \right\}$$

On cherche le meilleur dans cette famille :

$$\widehat{f} = f_{\widehat{w}}, \qquad \text{avec } \widehat{w} = \underset{w}{\operatorname{argmin}} \sum_{Train} \Big(Y_i - f_w(X_i) \Big)^2 := Loss(w)$$



Il y a deux techniques pour trouver $\hat{w}$.

- En minimisant $w \to Loss(w)$ par une méthode de gradient.
- Par un calcul direct que nous détaillons dans la prochaine section.

1.2 Calcul direct de $\hat{w}$

On pose $\mathbf{X} = \mathbf{X}_{train}$ la matrice dont la colonne 0 est constituée de 1, et dont les autres colonnes sont données par $\mathbf{X}_{ij} = X_i^j$ (i est l'indice qui fait parcourir $Train$ et j est l'indice des différentes variables explicatives). On pose $\mathbf{Y}$ la matrice colonne telle que $\mathbf{Y}_i = Y_i$. On considére $w = (w_0, w_1, ..., w_p)$ comme une matrice colonne.

Exprimons $Loss$ avec des multiplications matricielles :

$$\begin{aligned} Loss(w) &= \sum_i \Big(Y_i - f_w(X_i)\Big)^2 \\ &= \sum_i \Big(Y_i - \sum_j X_{ij} w_j)\Big)^2 \\ &= (\mathbf{Y} - \mathbf{X}w)^T (\mathbf{Y} - \mathbf{X}w) \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}w - w^T\mathbf{X}^T\mathbf{Y} + w^T\mathbf{X}^T\mathbf{X}w \end{aligned}$$

Pour trouver le minimum de cette fonction convexe, on calcule sa différentielle (cf. annexe) :

$$dLoss(w) = -2\mathbf{Y}^T\mathbf{X} + 2w^T\mathbf{X}^T\mathbf{X}$$

Cette différentielle s'annule lorsque :

$$w^T\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{X} \Leftrightarrow \mathbf{X}^T\mathbf{X}w = \mathbf{X}^T\mathbf{Y} \Leftrightarrow w = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Conclusion :

$$\hat{w} = \underset{w}{\operatorname{argmin}}\, Loss(w) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Et donc l'estimation est donnée par

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{w} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Rappelons que $\mathbf{X} = \mathbf{X}_{train}$ est formée de donnée train. Donc l'estimation ci-dessus est sur les données train. Maintenant, si l'on dispose de descripteurs *test* que l'on met dans une matrice $\mathbf{X}_{test}$. On prédit l'output correspondant par

$$\hat{\mathbf{Y}}_{test} = \mathbf{X}_{test}\hat{w} = \mathbf{X}_{test}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

2 Pénalisation (shrinkage)

2.1 Ridge

Il arrive souvent que les variables explicatives $X^1, ..., X^p$ soient corrélées. Par exemple pour une maison

- variables corrélées : X^1 = surface, X^2 =nombre de pièces.

- variables très corrélées : $X^1 =$ surface totale, $X^2 =$ surface habitable.

Ces corrélations créent des problèmes : Avec la méthode directe, la matrice $\mathbf{X}^T\mathbf{X}$ peut-être difficile à inverser (on dit qu'elle est mal conditionnée). Avec la méthode du gradient : le minimum de $Loss$ peut-être atteint (ou presque atteint) pour de nombreux w.

Par exemple, considérons deux variables quasi égaux $X^1 \simeq X^2$

$$Loss(w) = \sum_{Train} \Big(w_0 + w_1 X_i^1 + w_2 X_i^2 - Y_i\Big)^2 \simeq \sum_{Train} \Big(w_0 + (w_1 + w_2) X_i^1 - Y_i\Big)^2$$

Ainsi, seule la somme des paramètres w_1, w_2 compte dans $Loss$. Ainsi, en fonction de l'initialisation, l'algorithme de la descente du gradient peut donner $w_1 = 1001, w_2 = -1000$ aussi bien que $w_1 = 1, w_2 = 0$.

Pour réduire cette instabilité on rajoute une pénalisation :

$$Loss_\alpha(w) = \sum_{i \in Train} \Big(y_i - w_0 - \sum_j w_j X_i^j\Big)^2 + \alpha \sum_j (w_j)^2$$

Ainsi la descente du gradient préférera toujours $w_1 = 1, w_2 = 0$ à $w_1 = 1001, w_2 = -1000$, et si l'on prend α assez petit, le premier terme de $Loss_\alpha$ sera lui aussi minimisé.

Exercice 2.1 Ré-exprimer $Loss_\alpha$ en terme matricielle. Montez que la solution qui minimise cette loss est:

$$\hat{w} = (\mathbf{X}^T\mathbf{X} + \alpha I)^{-1}\mathbf{X}^T\mathbf{Y}$$

2.2 Lasso

On peut aussi utiliser la fonction

$$Loss_\alpha(w) = \sum_{i \in Train} \Big(y_i - \sum_j w_j X_i^j \Big)^2 + \alpha \sum_j |w_j|$$

Le terme $\sum_j |w_j|$ pénalise même les petits w_j et les pousse à devenir 0. Ainsi le $\widehat{w}$ calculé aura peu de $\widehat{w}_j \neq 0$ ce qui le rend facile à interpréter ; les variables explicatives peu utiles ont été évacuées.

2.3 Régressions avec transformation des variables explicatives

Sortie quantitative $Y \in \mathbb{R}$. Entrées quelconques

$$X = (X^1, ..., X^p) \in E_1 \times ... \times E_p$$

On se donne des fonctions $\varphi_j : E_1 \times ... \times E_p \to \mathbb{R}$. Ensuite on fait de la régression linéaire classique avec les variables $\varphi_j(X)$:

On pose :

$$f_w(x) = w_0 + \sum_j w_j \varphi_j(x)$$

$$\widehat{f} = f_{\widehat{w}}, \qquad \widehat{w} = \operatorname*{argmin}_w \sum_{i \in Train} \Big(Y_i - f_w(X_i) \Big)^2$$

Là encore, on a intérêt à minimiser le nombre de w_j non nul, en imposant des pénalisations, ou tout simplement en supprimant les $\widehat{w}_j$ tels que $\widehat{w}_j X^j$ est "très petit". Dans le modèle linéaire (et dans le GLM) on dispose d'un teste pour savoir si l'influence d'un coefficient est négligeable ou pas (la p-value de ce teste est traduite en étoile dans R).

Souvent $(\varphi_1, \varphi_2, ...)$ est appelé un "dictionnaire de fonctions". Le but est de décrire notre sortie avec un minimum de fonction du dictionnaire. Cela a un goût de compression de signal n'est-ce pas ?

2.4 Ridge et PCA: même combat

Ecrivons la décomposition SVD de X.

$$X = VSW$$
$$X^T = W^T S V^T$$

$$X^T X + \alpha I = W^T (S^2 + \alpha I) W$$
$$(X^T X + \alpha I)^{-1} = W^T (S^2 + \alpha I)^{-1} W$$

$$(X^T X + \alpha I)^{-1} X^T Y = W^T (S^2 + \alpha I)^{-1} S V^T Y$$
$$\hat{Y}_i = X(X^T X + \alpha I)^{-1} X^T Y = VS(S^2 + \alpha I)^{-1} S V^T Y$$

Notons D le nombre de colonne de X est N le nombre de ligne. Détaillons le calcul précédent:

$$\hat{Y}_i = \sum_{j=1}^{D} \sum_{n=1}^{N} V_{ij} \frac{s_j^2}{s_j^2 + \alpha} V_{nj} Y_n$$

Analysons: Augmenter α désavantage les petites valeurs de s_j.

Maintenant fixons la pénalité: $\alpha = 1$. Mais effectuons une PCA: il s'agit de choisir $d < D$ puis d'annuler toutes les petites valeurs propres s_j pour $j \in \{d+1, ..., D\}$, ainsi l'estimation est :

$$\hat{Y}_i = \sum_{j=1}^{d} \sum_{n=1}^{N} V_{ij} \frac{s_j^2}{s_j^2 + 1} V_{nj} Y_n$$

3 Annexe

3.1 Différentielle

Soit $f : \mathbb{R}^p \to \mathbb{R}, w \to f(w)$.

Définition 3.1 La différentielle de f en w est l'application linéaire $\ell : \mathbb{R}^p \to \mathbb{R}$ telle que

$$f(w + \varepsilon) = f(w) + \ell(\varepsilon) + o(\varepsilon)$$

Faites les deux exercices suivant en utilisant directement la définition ci-dessus de la différentielle. C'est très facile et cela aide à comprendre.

Exercice 3.1

- Soit S une matrice symétrique, vérifiez que la différentielle en w de l'application $f(w) = w^T S w$ est :
$$\ell(\varepsilon) = 2w^T S\varepsilon$$
- Soit A une matrice ligne. Vérifiez que la différentielle en w de $f(w) = Aw$ est donnée par :
$$\ell(\varepsilon) = A\varepsilon$$

Notons L la matrice ligne associée à ℓ càd $\ell(\varepsilon) = L\varepsilon$. La différentielle ℓ de f se calcul en général via la formule :

$$L = [\frac{\partial f}{\partial w_1}, ..., \frac{\partial f}{\partial w_p}]$$

3.2 Hessienne

La matrice Hessienne H de f en w est la matrice qui vérifie (en notant L la matrice colonne de la différentielle) :

$$f(w + \varepsilon) = f(w) + L\varepsilon + \varepsilon^T H\varepsilon + o(\|\varepsilon\|^2)$$

Vous en déduirez facilement que la matrice hessienne de $f(w) = w^T Sw$ est tout simplement ...

La matrice hessienne se calcule aussi avec les dérivées croisée : $H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j}$.

Exercice 3.2 bonus Appliquer la méthode de Newton sur la loss du modèle linéaire. Que constatez-vous ?