Traitement du signal avancé

TD - Régression linéaire

Rémi Flamary

Exercice 1 Régression linéaire simple

La régression linéaire simple vise à estimer les paramètres d'un modèle linéaire :

$$y = ax + b \tag{1}$$

Ceci se fait en minimisant l'erreur quadratique du modèle sur les données d'apprentissage $\{x_i, y_i\} \in \mathbb{R}^2, \forall i \in 1, \ldots, n$:

$$\min_{a,b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - ax_i - b)^2 \tag{2}$$

Le problème est convexe et différentiable, le minimum peut donc être déduit en calculant la dérivée le la fonction par rapport aux deux variables.

1. Calculer la dérivée de la fonction de coût par rapport à a et b.
2. Exprimer le système d'équations lorsque les dérivées sont égales à zéro.
3. Donner la valeur de $\hat{b}$ en fonction de $\bar{x} = \frac{1}{n}\sum_i x_i$ et $\bar{y} = \frac{1}{n}\sum_i y_i$ et $\hat{a}$.
4. Obtenir la solution $\hat{a}$ en fonction des données et des moyennes $\bar{x}, \bar{y}$.
5. Nous allons maintenant étudier une relation linéaire entre le chiffre d'affaire et le nombre de salariés d'une entreprise. Les données mesurées sont les suivantes :

Année	*Nombre de salariés*	*Chiffre d'affaire*
1957	294	634
1959	314	728
1961	383	819
1963	402	938
1965	475	1136
1967	786	1317

a) Représenter le nuage de points, poser le modèle et estimer les paramètres.
b) Quelle est la qualité de ce modèle ? Calculer le coefficient de corrélation.
c) Estimer le chiffre d'affaire d'une entreprise l'an prochain si elle emploie 800 salariés.
d) Discuter des exemples d'apprentissage, y a-t-il des points aberrants ?
e) Quel est le rôle du temps ?

Exercice 2 Régression linéaire multiple pondérée

Lorsque les données d'apprentissage proviennent de capteurs différents et que les caractéristiques de ces capteurs sont différentes, il peut être intéressant de prendre en compte la confiance que l'on a dans les exemples d'apprentissage lors de l'estimation.

Ceci est fait en ajoutant un poids pour chaque exemple dans la fonction de coût des moindres carrés :

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n} p_i \left(y_i - \mathbf{w}^\top \mathbf{x}_i - b\right)^2$$

1. Montrer que si $p_i = 1, \forall i$ alors le problème est un problème classique de régression linéaire multiple. Donner les paramètres estimés de la fonction linéaire dans ce cas là.
2. Exprimer le problème de minimisation sous une forme matricielle (avec $\mathbf{p}$ le vecteur contenant les poids p_i et $\mathbf{P} = diag(\mathbf{p})$).
3. Calculer le gradient de la fonction de coût et en déduire la solution de l'estimation des moindres carrés pondérés.
4. En suivant la même démarche, donner la solution des moindre carrés pondérés avec régularisation ridge.

Exercice 3 Régression Polynômiale

La régression linéaire étant un outil maîtrisé, nous proposons de traiter la régression polynômiale. Ainsi, on cherche à construire une relation entre une valeur réelle x et une valeur y comme :

$$y = \sum_{k=0}^{p} c_k x^k \quad \text{avec : } \mathbf{c} = [c_p, \ldots, c_k, \ldots, c_0] \in \mathbb{R}^{p+1} \tag{3}$$

On considère que $\mathbf{c}$ est un vecteur colonne. Pour apprendre les coefficients $\{c_k\}$ et donc pour apprendre la fonction de régression ci-dessus, on a à notre disposition, N couples (x_i, y_i) que l'on utilisera sous la forme de deux vecteurs colonne X et Y .

1. Donner le vecteur colonne $\mathbf{x}$ dépendant du réel x qui permet d'écrire l'équation comme $y = \mathbf{x}^T\mathbf{c}$, en déduire la relation matricielle correspondant à l'équation 2 entre le vecteur Y et la matrice Z avec :

$$Z = \begin{bmatrix} x_1^p & \ldots & x_1 & 1 \\ \vdots & \ddots & & \vdots \\ x_N^p & \ldots & x_N & 1 \end{bmatrix} = [X^p \; X^{p-1} \; \ldots X \; 1] \tag{4}$$

   construit à partir des données d'apprentissage (x_i, y_i). On admet que la puissance sur un vecteur s'applique composante par composante.
2. En appliquant le principe de minimisation des moindres carrés non-regularisés entre les données réelles et les données prédites par l'équation , exprimer $\mathbf{c}$ en fonction de Y et Z.
3. Que se passe-t-il si $p > N$ au niveau du calcul de $\mathbf{c}$?

Exercice 4 Régression non-linéaire

Dans les modèles suivants :
— y désigne la valeur réelle à prédire.
— w_i pour $i = 1, \ldots, d'$ désignent les paramètres du modèle.
— x_i pour $i = 1, \ldots, d$ désignent les variables
— ϵ désigne le bruit.

1. $y = w_1 x_1 + w_2 x_1^2 + \epsilon$, $d' = 2, d = 1$
2. $log(y) = w_1 x_1 + w_2 x_2 + w_3 + \epsilon$, $d' = 3, d = 2$
3. $log(y + \epsilon) = w_1 x_1 + w_2 x_2 + w_3$, $d' = 3, d = 2$
4. $log(y + w_4) = w_1 x_1 + w_2 x_2 + w_3 + \epsilon$, $d' = 4, d = 2$
5. $y = \epsilon (x_1)^{w_1} (x_2)^{w_2} (10)^{w_3}$, $d' = 3, d = 2$

Pour tous les modèles répondre aux questions suivantes :
— Est-il possible de reformuler le problème sous la forme d'une régression linéaire ?
— Si oui, décrire la procédure de linéarisation (transformation $y \to \tilde{y}$ et $x \to \tilde{x}$).
— Discuter de l'impact d'une éventuelle linéarisation sur le bruit. Le modèle linéaire avec bruit additif est-il toujours valable ?