TP PROBA STATS

benoit.albert@uca.fr

November 2022

1 Régression linéaire

Le but de ce TP est d'explorer les méthodes de régression linéaire.

1.1 Introduction

Etudions l'ajustement affine, méthode qui consiste à rechercher la droite permettant d'expliquer le comportement d'une variable statistique y comme étant une fonction affine d'une autre variable statistique x. C'est une méthode supervisée. Il est indispensable d'avoir des données.

1.2 Données

Nous souhaitons comprendre la relation entre température et altitude. Voici des données en été (table 1) et en hiver (table 2) dans 8 stations d'une vallée alpine.

1.3 Objectif

Trouver une relation linéaire entre la température t et l'altitude h. Par exemple, $a, b \in$ R, tel que $t = f(h) = ah + b$. Pour entrainer nos modèles, nous pouvons générer un nuage de point aléatoire (*np.random.rand(., 1)*) suivant une relation

Altitude (m)	Tempéarture (°C)
3500	10
2800	13
1300	20
750	25
300	30
900	22
1800	18
3100	11

Table 1: Température en été

Altitude (m)	Tempéarture (°C)
3500	-15
2800	-11
1300	0
750	3
300	10
900	2
1800	-2
3100	-13

Table 2: Température en hiver

linéaire définie $y = \beta + \alpha * x +$ bruit. $\alpha, \beta \in$ R fixés et le bruit construit aléatoirement.

2 Méthodes

2.1 Moindres Carrés

Soit deux variables aléatoires, une variable à expliquer Y et une variable explicative X. On dispose de n réalisations de ces variables. Soit le modèle de régression linéaire $y_i = ax_i + b + \epsilon_i$. ϵ_i est le terme d'erreur. On recherche a, b, les estimateurs des Moindres Carrés Ordinaires les valeurs minimisant la quantité l'erreur totale,

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

Comme la fonction S est convexe, il suffit d'annuler le gradient. On obtient les formules :

$$\hat{a} = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \tag{2}$$

$$\hat{b} = \frac{\sum_i y_i}{n} - \hat{a} \frac{\sum_i x_i}{n}. \tag{3}$$

Ayant calculé les paramètres avec les données (training). On peut utiliser notre modèle pour faire de la prédiction :

$$\hat{y} = f(x) = \hat{a}x + \hat{b}. \tag{4}$$

2.2 Maximum de vraisemblance

Il est possible de faire des hypothèse sur le bruit, c'est à dire modéliser l'erreur par exemple en supposant que les erreurs sont distribués par une loi gaussienne. Ce type de méthode est appelé maximum de vraisemblance. Désormais la fonction à minimiser est

$$\min_{a,b,\sigma} L(a,b,\sigma) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{y_i - ax_i - b}{\sigma}\right)^2. \tag{5}$$

Après calculs on obtient :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2, \tag{6}$$

et a,b comme les moindres carrées vous trouverez la démonstration p 4-7.

2.3 Méthode d'optimisation

Pour un modèle décrit, on peut aussi trouver ces paramètres comme les estimateurs à l'aide des méthodes d'optimisation. Soit la fonction coût équivalente à celle des moindres carrés,

$$\min_{a,b} J(a,b) = \frac{1}{2n}\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}(ax_i + b - y_i)^2 \tag{7}$$

Comme J est une fonction convexe, pour obtenir a, b il suffit de résoudre l'équation $\nabla J = 0$. Soit le gradient ∇J,

$$\frac{\partial J}{\partial a} = \frac{1}{n}\sum_{i=1}^{n} x_i(ax_i + b - y_i), \tag{8}$$

$$\frac{\partial J}{\partial b} = \frac{1}{n}\sum_{i=1}^{n} ax_i + b - y_i. \tag{9}$$

La méthode du gradient est une méthode itérative. Partant d'un point de départ, par exemple $a_0 = b_0 = 0$, elle chaque itération $k > 0$, $a_k = a_{k-1} - \gamma\frac{\partial J(.,a_{k-1})}{\partial a_{k-1}}$, idem pour b_k.γ est le " learning rate ", on le prendra égal à 0.5. L'algorithme s'arrête lorsque $|J(a_k, b_k) - J(a_{k-1}, b_{k-1})| < \varepsilon = e^{-3}$.

2.4 Bibliothèque Sklearn

Nous pouvons également utiliser la bibliothèque de python*sklearn* est notamment la fonction *LinearRegression*.

3 Qualité des prédictions

3.1 RMSE

Root-mean-square deviation est la racine de l'erreur quadratique moyenne entre les données observés Y et les données estimées $\hat{Y}$:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}. \tag{10}$$

3.2 Coefficient de détermination

Le coefficient de détermination est le ratio entre la somme des carrés des écarts à la moyenne des valeurs prédites par la régression et la somme des carrés des écarts à la moyenne totale $\bar{y}$:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{11}$$

4 Questions

Implémentez toutes les méthodes et mesures d'évaluation.
Afficher les différents jeux de données et l'approximation par les modèles que vous souhaitez utiliser.
Quel est le modèle le plus précis ?
Quel est la température à 1000m en été, en hiver ?
Supposons qu'il fasse 15 degré à 300m, combien devrait-il faire à 1000m ?