Travaux Dirigés n°1
Régression Linéaire

KMAXPP05

2025-2026

Exercice 1

Nous disposons des données suivantes pour des joueurs de rugby. L'espace des caractéristiques d'entrée est défini par la taille et noté $\mathcal{X} \subseteq \mathbb{R}$. L'espace des caractéristiques de sortie est défini par le poids et noté $\mathcal{Y} \subseteq \mathbb{R}$:

Taille (cm)	**Poids** (kg)
190	112
172	85
186	135
180	92

On cherche à obtenir un modèle linéaire de la forme : $\text{poids}^{(i)} = \beta_0 + \beta_1 \times \text{taille}^{(i)} + \epsilon^{(i)}$ que l'on peut également écrire, sous forme matricielle comme $Y = \tilde{X}\beta + \epsilon$. Les observations $(\text{poids}^{(i)}, \text{taille}^{(i)})$ sont indépendantes. Les $\epsilon^{(i)}$ sont des variables aléatoires indépendantes toutes issues de la loi normale centrée et de variance σ_ϵ^2.

1. Quelle est l'application Φ permettant de passer de l'espace des données observées $\mathcal{X}$ à l'espace des données augmenté $\tilde{\mathcal{X}} \subseteq \mathbb{R}^2$?
2. Écrire la matrice de design $\tilde{X}$ de ce problème de régression. On rappelle que la matrice de design est l'empilement vertical des données d'entrée dans l'espace augmenté $\tilde{\mathcal{X}}$.
3. Calculer la matrice $\tilde{X}^T\tilde{X}$. Cette matrice est-elle inversible ?
4. En utilisant la méthode des moindres carrés, calculer l'estimation $\hat{\beta}$ de ce problème de régression.
5. Donner une interprétation de β_0 et de β_1.
6. A l'aide de cette estimation de modèle, quels seraient les poids de deux joueurs qui mesureraient 188 cm et 200 cm.

Exercice 2

Un ballon sonde a effectué des mesures de température et de pression dans l'atmosphère. Les résultats de l'expérience sont consignés dans la table ci-dessous :

Altitude (km)	**Pression** (hPa)	**Température** (°C)
2	869	2
6	581	-23
10	293	-51

Nous cherchons à obtenir un modèle linéaire permettant de prédire la température à partir de l'altitude et de la pression. Nous considèrerons l'application $\Phi : \mathcal{X} \subseteq \mathbb{R}^2 \rightarrow \tilde{\mathcal{X}} \subseteq \mathbb{R}^3$ telle que $\Phi : \left(\text{altitude}^{(i)}, \text{pression}^{(i)}\right) \mapsto \left(1 \quad \text{altitude}^{(i)} \quad \text{pression}^{(i)}\right)$

Pour ce problème, on donne la matrice suivante :

$$\tilde{X}^T\tilde{X} = \begin{bmatrix} 3 & 18 & 1743 \\ 18 & 140 & 8154 \\ 1743 & 8154 & 1178571 \end{bmatrix}$$

1. La matrice $\tilde{X}^T\tilde{X}$ est-elle inversible ? Que peut-on en déduire ?
2. Calculer la matrice (2×2) de corrélation de Pearson de l'altitude et de la pression. On rappelle que pour deux variables aléatoires X_i et X_j, le coefficient de corrélation de Pearson ρ est donné par $\rho(X_i, X_j) = \dfrac{\text{Cov}[X_i, X_j]}{\sqrt{\text{Var}[X_i]\,\text{Var}[X_j]}}$
3. Que conclure de cette matrice de corrélation ?
4. Quelle alternative proposer pour un modèle linéaire de prédiction de la température ?
5. En utilisant la méthode des moindres carrés, quel serait ce modèle ?
6. D'après ce modèle quelle serait la température dans la couche haute de la troposphère (11000 mètres d'altitude) ?

Exercice 3

Un second ballon sonde a effectué une autre série de mesures de température dans l'atmosphère. Voici les relevés de ce ballon :

Altitude (km)	**Température** (°C)
0	15
11	-56.5

Nous cherchons à obtenir un modèle linéaire permettant de prédire la température à partir de l'altitude de la forme suivante : $\text{température}^{(i)} = \beta_0 + \beta_1 \times \text{altitude}^{(i)} + \epsilon^{(i)}$. On suppose que les observations sont indépendantes. Les $\epsilon^{(i)}$ sont des variables aléatoires indépendantes toutes issues de la loi normale centrée et de variance σ_ϵ^2.

La résolution numérique d'un problème de régression linéaire avec la méthode des moindres carrés évite d'inverser la matrice $\tilde{X}^T\tilde{X}$. Une approche classique est de passer par une décomposition QR de la matrice de design $\tilde{X}$ avec Q une matrice orthogonale (i.e $Q^TQ = I$) et R une matrice triangulaire supérieure.

Il n'existe pas une décomposition unique. La méthode de Gram-Schmidt permet d'en obtenir une. Afin d'obtenir Q, on peut ré-écrire la matrice $\tilde{X} \in \mathcal{M}_{(2,2)}(\mathbb{R})$ sous une forme qui fait apparaitre les vecteurs colonnes $\tilde{X}_1 \in \mathcal{M}_{2,1}(\mathbb{R})$ et $\tilde{X}_2 \in \mathcal{M}_{2,1}(\mathbb{R})$. La méthode peut se résumer de la façon suivante :

— La première colonne de Q que l'on notera Q_1 peut être obtenue en normalisant le premier vecteur $\tilde{X}_1$: $Q_1 = \dfrac{\tilde{X}_1}{\|\tilde{X}_1\|_2}$

— Pour la deuxième colonne de Q que l'on notera Q_2, on retire au vecteur $\tilde{X}_2$ sa projection sur Q_1 pour rendre Q_2 orthogonale à Q_1 : $\tilde{X}_2 - (\tilde{X}_2 \cdot Q_1)Q_1$. En normalisant ce vecteur, nous obtenons Q_2.

1. Un modèle linéaire peut être obtenu de manière triviale dans cet exemple. Quel est-il ?

2. Ré-écrire le calcul de $\hat{\beta}$ avec la méthode des moindres carrés en utilisant la décomposition $\tilde{X} = QR$. Montrer que $\hat{\beta}$ peut être simplement calculé à partir d'une substitution arrière (ou remontée). Cela revient à dire que $\hat{\beta}$ est la solution du système linéaire $R\hat{\beta} = A$ où le vecteur A, fonction de Q et de Y, est à déterminer.
3. Montrer que la méthode de Gram-Schmidt permet d'obtenir une matrice Q orthogonale. On pourra par exemple montrer que le produit scalaire des vecteurs Q_1 et Q_2 est nul.
4. Avec les données du ballon sonde, calculer les matrices Q et R.
5. En déduire la valeur de $\hat{\beta}$ en utilisant la méthode des moindres carrés. Obtient-on bien la même estimation $\hat{\beta}$ que la solution triviale ?