TD 04 : Régression linéaire

Fondements du Machine Learning, L3 IM2D

Octobre-Novembre 2024 (V2 : 3 novembre 2024)

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 4.1 : Variation de données

(Adapté d'un exercice d'examen proposé en 2019-2020.)

On considère un jeu de données $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, où $\boldsymbol{x}_i \in \mathbb{R}^n$ et $y_i \in \mathbb{R}$ pour tout $i = 1, \dots, m$. On notera de manière plus compacte

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n} \quad \text{et} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m.$$

On suppose dans cet exercice que $\operatorname{rang}(\boldsymbol{X}) = n \leq m$, et qu'il existe un vecteur $\boldsymbol{\beta}^* \in \mathbb{R}^n$ tel que:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

où $\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 I_m\right)$, c'est-à-dire que les composantes de $\boldsymbol{\epsilon}$ sont i.i.d. selon une loi normale de moyenne 0 et de variance $\sigma^2 > 0$.

On considère ensuite le problème aux moindres carrés suivant :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} f(\boldsymbol{\beta}) := \frac{1}{m} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 = \frac{1}{m} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2. \tag{1}$$

a) Donner une solution du problème (1) sous les hypothèses considérées. Pourquoi cette solution est-elle unique ?

b) Soit $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. Montrer que $\boldsymbol{H}$ est symétrique et que $\boldsymbol{H}^2 = \boldsymbol{H}$.

c) Justifier qu'une valeur propre de $\boldsymbol{H}$ est nécessairement égale à 0 ou 1. Sachant que $\operatorname{rang}(\boldsymbol{H}) = n$, en déduire que $\operatorname{trace}(\boldsymbol{H}) = n$.

d) Montrer que $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$, où $\hat{\boldsymbol{\beta}}$ est la solution de (1) considérée en question a). Montrer alors que

$$f(\hat{\boldsymbol{\beta}}) = \frac{1}{m}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y}).$$

e) Justifier que $\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\boldsymbol{\beta}^*$, où $\boldsymbol{\beta}^*$ est la vérité terrain du problème. En déduire que :

$$f(\hat{\boldsymbol{\beta}}) = \tfrac{1}{m}(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\epsilon}).$$

f) On veut maintenant obtenir l'espérance de l'erreur représentée par f lorsque le jeu de données $\mathcal{D}$ (donc le vecteur $\boldsymbol{\epsilon}$) varie. Montrer que l'on a :

$$\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] = \sigma^2(1 - \frac{n}{m}).$$

Indication : $\mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{u}\right] = \sigma^2 \operatorname{trace}(\boldsymbol{M})$ *pour toute matrice carrée* $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ *et tout vecteur* $\boldsymbol{u} \sim \mathcal{N}(0, \sigma^2 I_n)$.

g) En déduire que

$$\mathbb{E}_{\mathcal{D}}\left[f(\hat{\boldsymbol{\beta}})\right] \longrightarrow \mathbb{E}_{(\boldsymbol{x},y)}\left[R(\boldsymbol{\beta}^*)\right] \text{ quand } m \to \infty,$$

où $R(\boldsymbol{\beta}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} - y)^2$ et $(\boldsymbol{x}, y)$ suit la même distribution que les éléments de $\mathcal{D}$, c'est-à-dire que l'on a $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Exercice 4.2 : Ajout de données

Soient $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$ avec $\operatorname{rang}(\boldsymbol{X}) = n \leq m$ et $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, où $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_m)$. On rappelle que dans ce contexte, l'estimateur du maximum de vraisemblance (ou l'estimateur des moindres carrés) vaut

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{\beta}^* + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\epsilon}.$$

a) Retrouver les formules $\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}^*$ ainsi que
$\operatorname{Cov}\left[\hat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])(\hat{\boldsymbol{\beta}} - \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right])^{\mathrm{T}}\right] = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$.

b) Pour tout vecteur aléatoire $\boldsymbol{z} \in \mathbb{R}^n$, montrer que

$$\mathbb{E}\left[\|\boldsymbol{z} - \mathbb{E}\left[\boldsymbol{z}\right]\|_2^2\right] = \operatorname{trace}\operatorname{Cov}\left[\boldsymbol{z}\right].$$

En déduire que $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right] = \sigma^2 \operatorname{trace}\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right)$.

c) On considère l'ajout d'une nouvelle donnée $(\boldsymbol{v}, w = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_{m+1}) \in \mathbb{R}^n \times \mathbb{R}$ avec $\epsilon_{m+1} \sim \mathcal{N}(0, \sigma^2)$, et on s'intéresse à l'impact de cet ajout sur l'estimateur du maximum de vraisemblance. Etablir la formule de cet estimateur pour le nouveau jeu de données (on le notera $\tilde{\boldsymbol{\beta}}$) en fonction de $\boldsymbol{X}, \boldsymbol{v}, \boldsymbol{y}, w$ et calculer sa matrice de covariance.

d) Pour toute matrice $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ définie positive et tout vecteur $\boldsymbol{u} \in \mathbb{R}^n$, on a :

$$(\boldsymbol{A}^{-1} + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^{-1} = \boldsymbol{A} - \frac{1}{1 + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}.$$

En utilisant cette relation, montrer que $\operatorname{trace}((\boldsymbol{A}^{-1} + \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})^{-1}) = \operatorname{trace}(\boldsymbol{A}) - \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{u}}{1+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}}$.

e) Déduire de la question précédente un lien entre $\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right]$ et $\mathbb{E}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right]$. Sous quelle condition est-il avantageux d'introduire une donnée supplémentaire ?

Exercice 4.3 : Maximum de vraisemblance

(Adapté d'un exercice d'examen proposé en 2020-2021.)

On considère un jeu de données sous la forme de deux matrices $\boldsymbol{X} \in \mathbb{R}^{m_1 \times n}$, $\boldsymbol{W} \in \mathbb{R}^{m_2 \times n}$, et de deux vecteurs $\boldsymbol{y} \in \mathbb{R}^{m_1}$, $\boldsymbol{z} \in \mathbb{R}^{m_2}$. On fait l'hypothèse d'un modèle linéaire sous-jacent affecté par un bruit gaussien, c'est-à-dire que l'on suppose qu'il existe deux vecteurs $\boldsymbol{\beta}_1^* \in \mathbb{R}^n$ et $\boldsymbol{\beta}_2^* \in \mathbb{R}^n$ tels que

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}_1 \quad \text{et} \quad \boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_2,$$

où $\boldsymbol{\epsilon}_1$ est un vecteur aléatoire de $\mathbb{R}^{m_1}$ distribué selon une loi gaussienne $\mathcal{N}(\mathbf{0}_{\mathbb{R}^{m_1}}, \sigma^2 \boldsymbol{I}_{m_1})$ et $\boldsymbol{\epsilon}_2$ est un vecteur de $\mathbb{R}^{m_2}$ distribué selon une loi gaussienne $\mathcal{N}(\mathbf{0}_{\mathbb{R}^{m_2}}, \zeta^2 \boldsymbol{I}_{m_2})$, avec $\sigma > 0$ et $\zeta > 0$. On supposera également que les matrices $\boldsymbol{X}$ et $\boldsymbol{W}$ sont de même rang n, avec $n \leq \min\{m_1, m_2\}$.

a) Quelles sont alors les distributions des vecteurs $\boldsymbol{y}$ et $\boldsymbol{z}$?

b) Rappeler la définition d'une fonction de vraisemblance pour m observations réelles $a_1, \ldots, a_m$ dépendant d'un vecteur de paramètres $\boldsymbol{\beta}$, notée $L(a_1, \ldots, a_m; \boldsymbol{\beta})$. En suivant cette définition, établir des formules pour les vraisemblances

$$L(y_1, \ldots, y_{m_1}; \boldsymbol{\beta}_1) \quad \text{et} \quad L(z_1, \ldots, z_{m_2}; \boldsymbol{\beta}_2)$$

c) On s'intéresse à l'estimateur du maximum de vraisemblance pour $\boldsymbol{\beta}_1^*$.

   i) Donner le problème de maximisation de la vraisemblance dont cet estimateur est solution. Pourquoi est-il équivalent de maximiser la vraisemblance et son logarithme ?

   ii) Donner le lien entre l'estimateur du maximum de vraisemblance et l'ensemble des solutions du problème aux moindres carrés

$$\min_{\boldsymbol{\beta}_1 \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{y}\|^2.$$

   iii) En déduire une formule pour l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}_1^*$, que l'on notera $\hat{\boldsymbol{\beta}}_1$.

d) En utilisant le même raisonnement qu'en question c), donner la formule de l'estimateur du maximum de vraisemblance pour $\boldsymbol{\beta}_2^*$, noté $\hat{\boldsymbol{\beta}}_2$.

e) Justifier que le vecteur $\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$ est un estimateur non biaisé du vecteur $\begin{bmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{bmatrix}$.

Exercice 4.4 : Maximum a posteriori

(Adapté d'un exercice d'examen proposé en 2019-2020.)

Soit un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. Dans le cadre de la régression linéaire, on recherche un vecteur $\boldsymbol{\beta}$ tel que

$$y_i \approx \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i,$$

où les ϵ_i sont i.i.d. selon une loi normale de moyenne 0 et de variance 1.

On s'intéresse ici à l'estimateur du maximum a posteriori obtenu avec un a priori uniforme sur un intervalle réel $[a, b]$ (avec $a < b$). On rappelle que la densité de probabilité d'une variable z suivant une loi uniforme dans un intervalle $[a, b]$ est la fonction p donnée par :

$$p(z) = \begin{cases} \frac{1}{b-a} & \text{si } z \in [a, b], \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

a) Rappeler la définition de la vraisemblance $\mathcal{L}(y_1, \ldots, y_m; \boldsymbol{\beta})$, et donner sa formule pour le problème considéré.

b) On suppose a priori que les entrées de $\boldsymbol{\beta}$ sont i.i.d. suivant une loi uniforme sur $[a, b]$. Donner alors la formule de la loi a priori de $\boldsymbol{\beta}$, notée $\mathcal{L}(\boldsymbol{\beta})$.

c) Formuler le problème d'estimation du maximum a posteriori en utilisant les logarithmes des lois de probabilité.

d) Montrer que si l'on ne tient pas compte des termes constants, le problème se ramène à celui ci-dessous :

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ -\frac{1}{2} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 + \sum_{j=1}^n \ln \left(\chi_{b-a}([\boldsymbol{\beta}]_j) \right) \right\}, \tag{3}$$

où $\chi_{b-a}(z) = 1$ si $a \leq z \leq b$ et $\chi_{b-a}(z) = 0$ sinon.

e) On peut montrer que l'ensemble des solutions du problème (3) correspond à celui du problème aux moindres carrés avec contraintes suivant :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i)^2 \quad \text{sous les contraintes } a \leq \beta_j \leq b \quad \forall j = 1, \ldots, n. \tag{4}$$

En quoi ce nouveau problème modélise-t-il l'a priori que l'on a mis sur le modèle linéaire ?

f) On note $\hat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance. Si cet estimateur vérifie $a \leq [\hat{\boldsymbol{\beta}}]_j \leq b$ pour tout $j = 1, \ldots, n$, est-il solution du problème (4) ?