Master 2 DS
Science des données 1
TD 1

T. Artières, V. Emiya et H. Kadri
QARMA team - LIS lab

2025-2026

I. On dispose des mesures suivantes du temps d'exécution d'un algorithme en secondes (y) en fonction de la taille d'un tableau de données (x).

x	110	125	152	172	190	208	220	242	253	270	290
y	187	225	305	318	367	365	400	435	450	506	558

On donne les résultats suivants:

$$\sum x_i = 2232, \sum y_i = 4116, \sum x_i^2 = 487750, \sum x_i y_i = 900961, \sum y_i^2 = 1666782.$$

1. Représentez sur un graphique ces données par un nuage de points.
2. Trouvez la fonction de régression $y = \alpha x + \beta$ obtenue par la méthode des moindres carrés sur ce jeu de données. Tracez la droite de régression sur le graphique.
3. Utilisez la fonction de régression obtenue pour prédire le temps d'exécution de l'algorithme pour un tableau de taille 500. Interprétez le résultat obtenu à partir du graphique que vous avez tracé.
4. Supposons qu'on dispose du temps d'exécution de l'algorithme pour chacune des $n = 20$ valeurs suivantes $x_1 = 110, x_2 = 115, \ldots, x_{20} = 290$. Si l'objectif principal est d'estimer α le plus précisément possible, serait-il préférable d'utiliser le jeu de données avec n=20 ou celui avec n=11 décrit par le tableau ci-dessus?

II. En prenant les exemples du perceptron pour la classification, et de la régression linéaire, on identifie des éléments récurrents dans la formalisation d'un problème d'apprentissage automatique: un modèle, un critère, un algorithme d'optimisation.

1. Identifiez ces trois éléments dans le cas du perceptron, puis dans le cas de la régression linéaire. Quel est la nature du modèle et quelle hypothèse fait-il sur les données ? Quel est le critère optimisé pour apprendre le modèle ? Comment le critère est-il optimisé ?
2. Comment pourrait-on transformer la formalisation de l'apprentissage du perceptron pour utiliser une stratégie similaire à celle utilisée pour la régression (résolution analytique) plutôt qu'un algorithme itératif ? Y voyez-vous un avantage ? Un inconvénient ?
3. Concernant le perceptron, voyez-vous d'autres critères qui seraient intéressants à optimiser ?

III. On considère le problème d'apprentissage d'une fonction à valeur réelle $h : \mathbb{R}^d \to \mathbb{R}$ à partir d'un ensemble de données d'apprentissage $S = \{(x_i, y_i), 1 \leq i \leq n\}$, $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$. On considère uniquement le cas où la fonction h est linéaire et qui s'écrit sous la forme $h(x) = \langle w, x \rangle$, avec $w \in \mathbb{R}^d$ le vecteur de pondération solution du problème d'optimisation suivant :

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|_2^2, \tag{1}$$

où $\lambda \in \mathbb{R}_+$ est le paramètre de régularisation.

Soit X la matrice de taille $n \times d$ telle que $\mathrm{X}_{i,j} = (x_i)_j$ et Y le vecteur colonne de taille n tel que sa i ème composante est y_i. Finalement, soit W le vecteur de dimension d correspondant au vecteur de pondération w.

1. Montrez que les prédictions du modèle pour l'ensemble des données dans X peuvent être calculées par $\mathrm{X} \times \mathrm{W}$.
2. Soient deux vecteurs u et v. Quelle est la dimension de $\frac{\partial u^T v}{\partial u}$. Montrez que $\frac{\partial u^T v}{\partial u} = v$.
3. Soit $\|w\|^2$ la norme d'un vecteur w. Que vaut $\frac{\partial \|w\|^2}{\partial w}$?
4. Exprimer la fonction objective du problème d'optimisation considéré ci-dessus (équation (1)) en fonction des matrices $\mathrm{X}, \mathrm{Y}, \mathrm{W}$ et le paramètre de régularisation λ.
5. Déterminer une forme analytique de la solution optimal W^* du problème d'optimisation en fonction de X, Y et λ. Vous pourriez avoir besoin d'utiliser $\frac{\partial \|A\|^2}{\partial A} = 2A$, pour une matrice A, ce que vous pouvez redémontrer facilement, et d'autres éléments d'aide dans le *matrix cookbook* pointé sur le site du master.
6. Quelle est la complexité en temps pour calculer le vecteur de pondération optimal W^* comme une fonction du nombre d'attributs d et du nombre d'exemples n? Quelle est la complexité de calculer $h(x)$ pour une nouvelle donnée $x \in \mathbb{R}^n$?
7. La matrice $\mathrm{X}\mathrm{X}^\top$ est appelé la matrice de Gram. Utilisant le fait que:

$$\mathrm{X}^\top(\mathrm{X}\mathrm{X}^\top + \lambda I)^{-1} = (\mathrm{X}^\top\mathrm{X} + \lambda I)^{-1}\mathrm{X}^\top,$$ [1]

déterminer une nouvelle expression de la solution optimale W^*? Quelle est la complexité en temps de calcul de W^* utilisant cette forme analytique de la solution?

IV. On utilise les mêmes notations que l'exercice précédent.

Montrer que l'estimateur de la régression Ridge (Equation 1) peut être obtenu à partir de l'estimateur des moindres carrés sur un jeu de donnée augmenté. Pour cela, augmenter la matrice des variables explicatives (données d'entrée) $X \in \mathbb{R}^{n\times d}$ par d lignes en ajoutant la matrice $\sqrt{\lambda}I$ et le vecteur de réponses (sorties) Y par d valeurs nulles.

V. Considérons le problème d'optimisation "elastic-net":

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} (\langle w, x_i\rangle - y_i)^2 + \lambda\big[\alpha\|w\|_2^2 + (1-\alpha)\|w\|_1\big].$$

Montrer que résoudre ce problème revient à résoudre le problème lasso sur un jeu de données augmenté.
Pour rappel, le problème Lasso s'écrit sous la forme :

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} (\langle w, x_i\rangle - y_i)^2 + \beta\|w\|_1.$$

[1] Voir The Matrix Cookbook, Eq. 167. `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`