

Devoir surveillé : 1h30

Exercice 1

Le tableau suivant décrit l'expérience de 5 étudiants avant l'examen d'un module et la note qu'ils ont obtenu en conséquence **du nombre d'heures qu'ils ont passé à étudier** et **du nombre d'heures qu'ils ont dormi la veille de l'examen**. La première colonne contient E1 jusqu'à E5 : qui est l'identifiant de chaque étudiant. La deuxième colonne définit le nombre d'heures total passé par chaque étudiant à étudier le module, la troisième colonne définit le nombre d'heures que chaque étudiant a dormi la veille de l'examen et la dernière colonne définit la note obtenue pour ce module.

Etudiant	Nombre d'heures d'études	Nombre d'heures de sommeil la veille de l'examen	Note
E1	1	8	3
E2	20	8	18
E3	5	5	7
E4	15	3	14
E5	25	8	19

Nous voulons définir la fonction qui exprime la note en fonction du nombre d'heures d'études et du nombre d'heures de sommeil la veille de l'examen en utilisant **la régression linéaire** :

Note = $f(\text{Nombres d'heures d'études}, \text{Nombre d'heures de sommeil la veille de l'examen})$

1-A quelles colonnes correspondent x_1 , x_2 et y dans le tableau précédent ? (1pt)

2-Exprimer $h_{\theta}(x^{(i)})$ en fonction des θ_i , des $x^{(i)}$ et des $y^{(i)}$ pour le tableau précédent. (1 pt)

3-Si on vous donne 2 propositions de $\theta = \begin{bmatrix} \theta_0 = 0 \\ \theta_1 = 1 \\ \theta_2 = 0.5 \end{bmatrix}$ et $\theta = \begin{bmatrix} \theta_0 = 0 \\ \theta_1 = 1 \\ \theta_2 = 1 \end{bmatrix}$, laquelle des deux permet de prédire le mieux la note pour l'étudiant E3 ? Justifier la réponse. (1pt)

4-Pour le choix de $\theta = \begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix}$, exprimer l'erreur $J(\theta)$ pour les étudiants sur le tableau en exprimant la formule et les calculs d'une manière claire. (2pts)

5-Pour le choix de $\alpha = 0.1$ et de $\theta = \begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix}$ et la formule de mise à jour :

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Calculer la nouvelle valeur de θ après une mise à jour, détaillez le résultat. (2pts)

Exercice 2

Pour adapter le tableau précédent au problème de **régression logistique** nous proposons un tableau où la note est remplacée par la validation : **V**. Le but est de trouver un modèle de régression logistique qui doit prédire si un étudiant a validé un module en utilisant le nombre d'heure qu'il a passé à étudier ce module et le nombre d'heures qu'il a dormi la veille de l'examen. Le tableau est le même en changeant uniquement la dernière colonne par l'utilisation de la règle suivante. **Si Note ≥ 10 alors V=1 sinon V=0.**

Etudiant	Nombre d'heures d'études	Nombre d'heures de sommeil la veille de l'examen	V
E1	1	8	0
E2	20	8	1
E3	5	5	0
E4	15	3	1
E5	25	8	1

6-A quelles colonnes correspondent x_1 , x_2 et y dans le tableau précédent ? (1pt)

7-Un nouvel étudiant a étudié pendant 10 heures et a dormi pendant 6 heures. Selon le modèle

a-t-il validé le module quand $\theta = \begin{bmatrix} -10 \\ 6 \\ 10 \\ 4 \\ 6 \end{bmatrix}$, la règle est la suivante : si $\sigma(\dots) \geq 1/2$ alors l'étudiant

a validé sinon il n'a pas validé, écrire la formule et les calculs correspondants. (1pt)

8-Vérifier si pour la même valeur θ si le modèle prédit bien si un étudiant a validé ou pas dans le cas de l'étudiant à la première ligne et celui à la quatrième ligne. Justifier en appliquant les règles et les calculs correspondants. (1pt)

Exercice 3

Les données du premier tableau de l'exercice 1 sont stockées dans un fichier nommé 'notes.txt' du répertoire '/etudiant/input '

9-Comment transférer ces données vers le système de fichiers hdfs dans le répertoire '/etudiant/output avec le même nom (2pts)

10-Comment créer le répertoire '/etudiant/input dans le système HDFS ? (2pts)

11- Comment supprimer fichier créé. (2pts)

Exercice 4

1-Quel est l'intérêt est le besoin d'utiliser l'approche Big Data dans un projet informatique ? (2pts)

2-Décrire un exemple d'application de Map Reduce non cité dans le cours, l'intérêt de cette application et la manière d'appliquer les deux fonctions Map et Reduce à ce cas de figure. (2pts)