

Veuillez prendre connaissance des consignes ci-dessous :

- Une feuille de note recto verso 8.5" X 11" ou A4 et une calculatrice programmable sont permises.
- Il est interdit d'avoir sur soi un appareil de communication allumé ou éteint, y compris les téléphones, les tablettes, les montres intelligentes, et les ordinateurs portables.
- Toute forme de communication avec toute autre personne que le surveillant est interdite, incluant les paroles, les gestes, et l'échange de documents.
- Le barème est donné à titre indicatif et peut être sujet à modification.
- Vous devez encadrer les résultats et donner les valeurs avec deux chiffres après la virgule en arrondissant au plus proche. Par exemple, 1,947 doit être noté **1,95**.
- Vous pouvez répondre aux questions en français ou en anglais.

Please read carefully the following instructions :

- A handwritten double-sided sheet of size 8.5" X 11" or A4 and a programmable calculator are allowed.
- It is forbidden to have an electronic communication device on your person, including smartphones, tablets, smartwatches, and computers.
- All forms of communication with anyone other than the supervisor are forbidden, including speaking, making gestures, and exchanging documents.
- The marking of the questions is given for references and may be subject to change.
- You must draw a box around the results and give values with two decimal places by rounding to the nearest value. For instance, 1.947 must be given as **1.95**.
- You are allowed to answer in French or in English.

(425)

1. (2 points) Soit le jeu de données suivant sur lequel nous souhaitons faire une régression linéaire afin de prédire le montant du prêt (attribut "Montant du prêt"). Quelles étapes du prétraitement des données sont nécessaires afin d'obtenir un meilleur modèle ? Justifiez chaque étape en une ligne.

Consider the following dataset on which we want to apply a linear regression to predict the loan amount (feature "Montant du prêt"). What preprocessing steps are necessary in order to obtain a better model? Justify each step in one line.

Age	Poids	Sexe	Ville	Diplôme	Date de la demande	Montant du prêt
26	44 kg	F	Montréal	Baccalauréat	01/07/2021	\$18,000
44	140 lbs	M	montreal	Baccalauréat	01/02/2003	\$52,000
27	71 kg	F	Laval		01/09/2017	\$63,000
0	58 kg	M	Quebec	Maitrise	01/01/1999	\$21,000
56	122 lbs	M	Montréal	Baccalauréat	01/10/2019	\$41,000
34	56 kg	F	Quebec	Doctorat	01/06/2020	\$8,000
29	82 kg	M	Brossard	Baccalauréat	01/03/2014	\$5,000
61	98 lbs	F	Montréal	Maitrise	01/12/2014	\$20,000

(1,18)

- 1) Affecter valeur manquante : remplacer la valeur vide dans diplôme par la valeur majoritaire (Baccalauréat). La régression linéaire ne supporte pas les valeurs manquantes.
- 2) Normaliser les attributs numériques (âge, date demande, montant prêt...) afin d'éviter les instabilités numériques et de pouvoir comparer les poids de la régression.
- 3) Convertir les dates (chaîne de caractères) en temps UTC/Unix interprétable par la régression linéaire.
- 4) Binarisier l'attribut diplôme, car la régression linéaire n'accepte pas les chaînes de caractères en entrée, mais seulement les valeurs numériques.
- 5) Identifier les attributs hautement liés et les éliminer car ils ne contribuent pas à la prédiction.
- 6) Ajuster les montants du prêt en fonction de la date de prêt afin de tenir compte de l'inflation (on nous avons des dates en 1999 et en 2021 par exemple).
- 7) Binarisier l'attribut sexe.
- 8) Mettre le poids sur les mêmes unités de mesure (ex: tout mettre en kg) car cela crée un problème de compatibilité. Les derniers ont besoin d'être comparables pour faire des comparaisons.

Voir page suivante

-0,11
Mauvais
ordre

-0,11
ordre
(à la fin)

-0,11
Mauvais
ordre

(4)

Manque;

- 0,22 : Uniformiser les noms des Villes
- 0,05 : Binariser les noms des Villes
- 0,22 : Remplacer l'outlier Age = 0

2. (2 points) Appliquez l'algorithme d'emballage pour sélectionner les attributs avec la plus grande performance de classification tels que la somme de leur corrélations soit strictement inférieure à 1. La matrice de corrélation des attributs et la précision des modèles entraînés avec chaque ensemble d'attributs vous sont données. La sélection des attributs est-elle optimale ? Justifiez en une ligne.

Apply the sequential forward selection algorithm to select the features with the highest performance such that the sum of their correlations is strictly lower than 1. The feature correlation matrix and the accuracy of the models trained on each set of features are given. Is the feature selection optimal? Justify in one line.

La matrice de corrélation pour les 4 attributs :

	X^1	X^2	X^3	X^4
X^1	1.0	0.3	0.2	0.4
X^2	0.3	1.0	0.5	0.1
X^3	0.2	0.5	1.0	0.4
X^4	0.4	0.1	0.4	1.0

Le tableau des performances en fonction des attributs considérés :

X^1	X^2	X^3	X^4	Précision
✓				61%
	✓			44%
		✓		12%
			✓	29%
✓	✓			65%
✓		✓		78%
✓			✓	71%
	✓	✓		81%
	✓		✓	55%
		✓	✓	49%
✓	✓	✓		88%
✓	✓		✓	79%
✓		✓	✓	77%
	✓	✓	✓	91%
✓	✓	✓	✓	94%

①

Décomposition par valeur propre : toute matrice est carrée symétrique de taille 4×4 et décomposée de la façon $A = \sum_{i=1}^3 \lambda_i v_i v_i^T$. Les produits les + importants sont ceux associés aux valeurs propres les + grandes.

?

	X_1	X_2	X_3	X_4
1.0	0.61	0.44	0.23	0.4
0.3	0.44	1.0	0.5	0.12
0.2	0.23	0.5	1.0	0.4
0.4	0.4	0.12	0.4	1.0
Σ	1.9	1.9	2.1	1.9

Normalisation

	X_1	X_2	X_3	X_4
X_1	1,0	0,3	0,2	0,4
X_2	0,3	1,0	0,5	0,1
X_3	0,2	0,5	1,0	0,4
X_4	0,4	0,1	0,4	1,0
Σ	1,9	1,9	2,1	1,9

moyenne de la ligne X_i

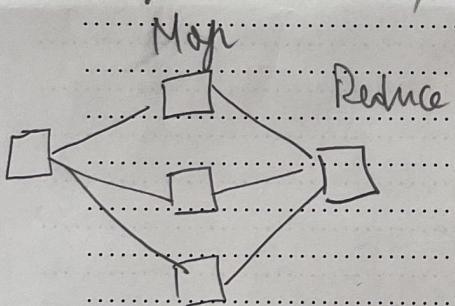
	X_1	X_2	X_3	X_4	Poids
X_1	0,526	0,158	0,0952	0,211	0,24755
X_2	0,158	0,526	0,738	0,0526	0,24365
X_3	0,105	0,263	0,476	0,211	0,26375
X_4	0,211	0,053	0,190	0,526	0,245
Σ	1	1	1	1	

3. (2 points) Expliquez dans vos propres mots le paradigme diviser pour régner. Vous pouvez illustrer votre explication avec un schéma. Dans quel cas cette approche est-elle adaptée ?

1-5

Explain in your own words the divide and conquer paradigm. You may illustrate your explanation with a figure. When is this approach relevant?

Réduction de données → quand la taille des données est plus petite, il est plus facile et plus performant d'appliquer des algorithmes sophistiqués et complexes. Il est également plus facile d'interpréter les résultats ainsi que de bien les visualiser. La réduction de données peut se faire en terme du nb de lignes (enregistrements) et/ou en terme de nombre de colonnes (dimensions). On peut utiliser cette approche pour faire un échantillonnage de données, c.-à-d que les enregistrements sont échantillonés pour créer une BD plus petite. Ainsi, une fraction de données est sélectionnée et retenue pour l'analyse. De plus, on peut aussi faire une réduction de dimension.



Map Reduce est un exemple de cette stratégie

Avec Map Reduce, on peut stocker et traiter le big data. C'est un framework permettant de distribuer un problème en divisant les calculs en sous-problèmes. C'est un modèle de programmation parallèle, mis à l'échelle de millions d'ordinateurs simples. Sert à la tolérance aux panne, grâce à la redondance.

Map est une fonction exécutée par chaque ordinateur pour résoudre un sous-problème. Reduce est une fonction qui combine les solutions de chaque sous-problème en la solution finale.

Hadoop → Map Reduce permet le traitement de données massives de manière parallèle / distribuée (fault-tolerant). Donc, le paradigme de diviser pour régner est possible avec Map Reduce, car on vent

trouver le Big Data qui est grand en distribuant et divisant les calculs en sous-problèmes et les réduire, c-à-d de combiner les solutions de chaque sous-problème en une solution finale. Cela est plus facile pour la gestion de données.

$$R_1(A=1, B=2) \quad R_2(A=1, B=2)$$

$$(A=3, B=4) \quad R_2(A=5, B=6)$$

Matricule : 1955913

INF8111 - Fouille de données

4. (2 points) Soit deux relations $R_1(A: \text{entier}, B: \text{entier})$ et $R_2(A: \text{entier}, B: \text{entier})$ sans doublons. Implémentez en pseudocode les fonctions Map et Reduce afin de calculer la différence entre ces deux relations. Pour rappel, la différence, notée $R_1 - R_2$, est une opération portant sur deux relations R_1 et R_2 ayant le même schéma et construisant une troisième relation dont les occurrences sont constituées de celles ne se trouvant que dans la relation R_1 . Utilisez des noms de variable explicites ou commentez votre code.

Consider two tables $R1(A: \text{int}, B: \text{int})$ and $R2(A: \text{int}, B: \text{int})$ without duplicates. Implement in pseudocode the Map and Reduce functions to compute the difference between the two tables. As a reminder, the difference noted $R1 - R2$ is an operation on two tables $R1$ and $R2$ with the same schema that returns a third table whose tuples belong solely to the table $R1$. Use explicit variable names or comment your code.

3,5

Non
Ici, je suppose que je fais une intersection.
Pour la Map, pour chaque ligne, on retourne (r, r). Pour Reduce, chaque clé est associée à deux valeurs, la fonction retourne (r, r), sinon rien.
Impossible : List<tuple<int>>
ou tuple<int>

Map (String file_id, List<Int> numbers):
while (numbers.hasNext()):
 for Number_R1 in file_id.R1:
 for Number_R2 in file_id.R2:
 if (number_R1 == number_R2):
 emit (number, 1)
 end if
 end for
 end for
end while

Ne fait pas l'opération demandée.

D'après
algorithme
on prend en
entrée toutes
les tables:
Problème
potentiel de
mémoire

Reduce (int key, Iterator<Int> values)

for each number on values:

emit (key, number)

end for

lo clé lo Valen
..... a ermit

5. (2 points) Donnez la solution analytique de la régression ridge. Pour rappel, la régression ridge est une régression linéaire avec un terme de pénalité correspondant au carré de la norme euclidienne du vecteur de poids w , ce qui revient à minimiser la fonction de coût suivante : $J(w) = \sum_{i=1}^n (f(X_i) - y_i)^2 + \lambda \sum_{j=1}^d (w_j)^2$.

Give the closed form solution to the ridge regression. As a reminder, the ridge regression is a linear regression with a penalty term corresponding to the square of the Euclidean norm of the weight vector w , which is equivalent to minimizing the following cost function : $J(w) = \sum_{i=1}^n (f(X_i) - y_i)^2 + \lambda \sum_{j=1}^d (w_j)^2$.

$$\textcircled{03} \quad 0 = \sum_{i=1}^n (w^T X_i - y_i)^2$$

On soit que :

$$w = d \cdot l \quad a^T = a$$

$$X = n \cdot d \quad [y_1, y_2, \dots, y_n] \cdot [x_1, x_2, \dots, x_n]^T = 0$$

$$Y = n \cdot l \quad [x_1, x_2, \dots, x_n] \cdot [y_1, y_2, \dots, y_n]^T = 0$$

$$y^T e = 0 \quad x^T e = 0$$

$$x^T w = y$$

D'abord, la régression ridge est une pénalisation des carrés,

c.-à-d. une mise à zéro des coefficients.

Recompense la mise à zéro des coefficients.

→ On peut arriver à un résultat final où $w = (x^T x)^{-1} x^T y$

$$L = \|xw - y + \lambda e\|^2 \rightarrow \text{ici je suppose que le } x \text{ est un coefficient (comme ex: b)}$$

$$= (xw - y + \lambda e)^T (xw - y + \lambda e)$$

$$= w^T x^T x w - w^T x^T y + \lambda w^T x^T e - y^T x w + y^T y - \lambda y^T e$$

$$+ \lambda^2 e^T e$$

$$= w^T x^T x w - 2 w^T x^T y + 2 \lambda w^T x^T e - 2 \lambda e^T y + \|y\|^2 + \lambda^2 \|e\|^2$$

$$+ \lambda^2 \|e\|^2$$

$$L = w^T x^T x w - 2 w^T x^T y + \|y\|^2 + \lambda^2 \|e\|^2$$

$$w^T x^T y = y^T x w$$

$$x^T w = \lambda e^T x w$$

$$\frac{\partial L}{\partial w} = -2 x^T x w + 2 x^T y$$

$$x^T x w = x^T y$$

$$x^T x = \|x\|^2 = \sum x_i^2$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = (x^T x)^{-1} x^T y \quad \text{As invérifiable}$$

$$\frac{\partial}{\partial x} x^T x = 2x$$

Note : le gradient de L par rapport à w est égal à $= 2x^T (xw - y)$