Veuillez prendre connaissance des consignes ci-dessous :

- Durée de l'examen: **2h30**
- Une feuille de note recto verso 8.5" X 11" ou A4 et une calculatrice programmable sont permises.
- 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 $\boxed{1{,}95}$.
- Vous devez détailler au moins une itération de chaque méthode.
- Vous pouvez répondre aux questions en français ou en anglais.

Please read carefully the following instructions:

- Exam duration: **2h30**
- A double-sided sheet of size 8.5" X 11" or A4 and a programmable calculator are allowed.
- 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 $\boxed{1.95}$.
- You must detail at least one iteration of each method.
- You are allowed to answer in French or in English.

1 Préparation de données

La popularité des produits d'un magasin comme Amazon suit une distribution telle que celle présentée ci-dessous (Figure 1).

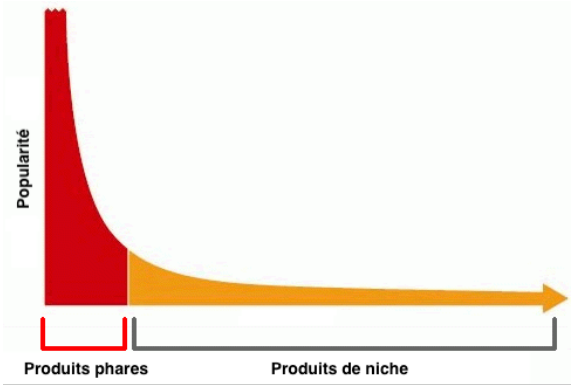


Figure 1:

Votre tâche est de construire un modèle de régression linéaire sur les attributs des produits pour prédire l'index de popularité d'un nouveau produit que l'on veut lancer dans le marché.

(a) Prévoyez-vous traiter les valeurs de la variable cible (c.-à-d. de l'index de popularité) avant de construire un modèle de régression linéaire ? Justifiez votre réponse.

C'est souhaitable de considérer le logarithme des valeurs de la variable cible étant donné qu'elle est régie par une distribution de la loi de puissance.

(b) En raison d'une faute dans l'intégration des données, il y a des produits dont la valeur d'index de popularité est *manquante*. Est-ce une bonne idée d'imputer les valeurs manquantes à partir de la valeur moyenne des index de popularité calculée sur les autres produits ? Justifiez votre réponse.

Non. La moyenne est largement influencée par les grands index de popularité. Une meilleure option serait d'imputer à partir de la valeur médiane.

(c) On souhaite utiliser le pays de fabrication des produits comme un des attributs du modèle de régression. Les valeurs possibles pour cet attribut sont *Chine, Japon, Canada et États-Unis*. Fournissez la transformation nécessaire pour utiliser cet attribut dans le modèle de régression linéaire.

On peut construire quatre attributs binaires, un pour chaque valeur possible de la variable catégorique

médianne: 2.67

2 Transformation des données

Appliquez l'algorithme d'emballage (*sequential forwarding*) pour sélectionner les attributs avec la plus grande performance de classification tels que la somme de leurs corrélations soit strictement inférieure à 0.5. 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.

La matrice de corrélation pour les 4 attributs:

$$\begin{array}{c} \\ X^1 \\ X^2 \\ X^3 \\ X^4 \end{array} \begin{array}{c} \begin{array}{cccc} X^1 & X^2 & X^3 & X^4 \end{array} \\ \begin{bmatrix} 1.0 & 0.3 & 0.2 & 0.4 \\ 0.3 & 1.0 & 0.5 & 0.1 \\ 0.2 & 0.5 & 1.0 & 0.4 \\ 0.4 & 0.1 & 0.4 & 1.0 \end{bmatrix} \end{array}$$

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%

On note c la somme des corrélations à une étape de l'algorithme

$S_1 = \{\}$

$S_2 = \{X_1\}$ car meilleure précision de l'algorithme n'utilisant qu'un attribut Note: ne pas compter $c = Corr(X_1, X_1) = 1$: algorithme s'arrête toujours au premier attribut sinon.

$S_3 = \{X_1, X_3\}, c = 0.2 < 0.5$

$S_4 = \{X_1, X_2, X_3\}, c = Corr(X_1, X_3) + Corr(X_1, X_2) + Corr(X_2, X_3) = 0.2 + 0.3 + 0.5$ ce qui n'est pas inférieur à 0.5

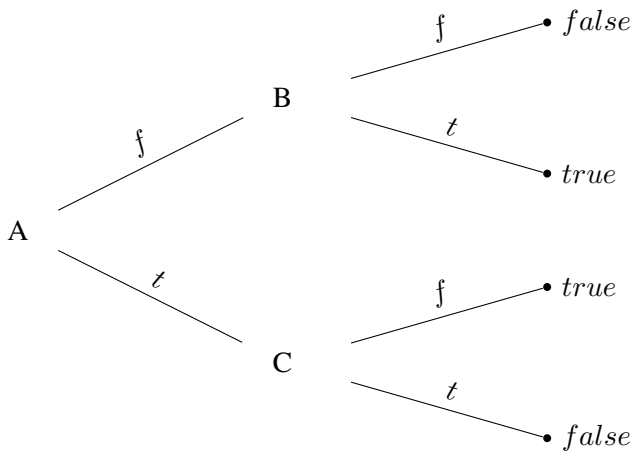
On choisit alors X_1 et X_3

- Sélection optimale ? Non, avec 3 attributs on remarque que sélectionner X_2, X_3, X_4 aurait donné un modèle avec 91% de précision.

médianne: 2

3 Classification

Étant donné l'arbre de décision suivant et les tables pour les données d'entraînement et de test.



Données d'entraînement :

A	B	C	Class
t	t	f	t
t	f	f	t
t	t	t	f
f	f	t	f
f	f	f	f
f	t	t	t

Données de teste :

A	B	C	Class
t	t	f	t
t	f	f	t
f	t	f	f
f	f	t	f

Répondez aux itens suivants:

(a) Donnez l'*accuracy* de l'arbre pour les données d'entraînement et de test.
100% pour les données d'entraînement et 75% pour les données de teste

(b) Donnez une formule logique propositionnelle équivalente à l'arbre de décision donné.
$(A \wedge \neg C) \vee (\neg A \wedge B)$

(c) Faites un élagage au premier niveau de l'arbre donné en fonction des données d'entraînement. Dessinez cet arbre et calculez son *accuracy* pour les données d'entraînement et de test.
66.6% pour les données d'entraînement et 100% pour les données de teste

(d) Est-ce que vous observez du surapprentissage de l'arbre original? Justifiez votre réponse.
Oui. L'accuracy est de 100% pour les données d'entraînement avec l'arbre original. Cependant, un arbre plus simple performe mieux pour l'ensemble de teste.

médianne: 3

4 MapReduce

Vous disposez d'un grand ensemble de données contenant une liste de produits achetés par des clients. Chaque enregistrement dans l'ensemble de données représente un seul achat et comprend les champs suivants : "Identifiant du client", "Identifiant du produit", "Quantité".

À l'aide du paradigme MapReduce, écrivez un programme pour identifier les produits uniques achetés par chaque client.

Fournissez le pseudocode ou l'outline des fonctions `Map` et `Reduce`.

(Note : Vous pouvez supposer que l'ensemble de données est déjà divisé en morceaux gérables et distribué sur plusieurs nœuds dans un environnement de calcul distribué.)

Pseudocode pour la fonction Map :

```
Map(Clef, Valeur) :
    identifiantClient = extraireIdentifiantClient(Valeur)
    identifiantProduit = extraireIdentifiantProduit(Valeur)
    emit((identifiantClient, identifiantProduit),-1)
```

Note : extraireIdentifiantClient() et extraireIdentifiantProduit() sont des fonctions qui extraient les champs respectifs de l'enregistrement d'entrée.

Pseudocode pour la fonction Reduce :

```
Reduce(Clef, Liste<Valeur>) :
    emit(Clef, Liste[0])
```

médianne: 4

5 Big Data

a) Vrai ou Faux ? Justifiez vos réponses lorsque vous avez identifié une phrase comme fausse.

i MapReduce est largement utilisé dans le domaine du Big Data pour traiter des volumes massifs de données de manière efficace.

ii MapReduce garantit la tolérance aux pannes en répliquant automatiquement les données sur différents nœuds du cluster.

iii La programmation en MapReduce est réservée aux développeurs spécialisés en infrastructures distribuées.

Faux. Bien qu'une compréhension des infrastructures distribuées soit utile, la programmation en MapReduce peut être réalisée par des développeurs avec une connaissance adéquate de ce modèle de programmation.

iv Chaque réducteur (*reducer* en anglais) doit produire autant de paires $(clé, valeur)$ en sortie que le nombre de paires qu'il a reçu en entrée.

Faux. Les *reducers* peuvent générer n'importe quel nombre de paires clé/valeur (y compris zéro).

v Le type de sortie des paires $(clé, valeur)$ des mappeurs (*mapper* en anglais) doit être du même type que leur entrée.

Faux. Le mapper peut produire des paires clé/valeur de n'importe quel type.

vi Malgré la demande croissante de données, la consommation d'électricité des technologies de l'information et de la communication est contrebalancée par plus d'efficacité dans l'utilisation de ressources de calcul - y compris la fermeture des installations plus anciennes au profit de centres de calcul ultra-efficaces.

vii Les scientifiques et les ingénieurs essayent aujourd'hui de rationaliser les processus informatiques, en utilisant des sources d'énergie renouvelables et en recherchant de meilleurs moyens de refroidir les centres de données et de recycler leur chaleur résiduelle.

viii Au fur et à mesure que les modèles ML augmentent en échelle, une tendance générale est qu'ils deviennent plus précis et plus performants. Cependant, des modèles plus grands se traduisent par des demandes de calcul plus importantes et, par extension, des demandes d'énergie plus importantes.

b) Nommez trois actions concrètes qui peuvent être entreprises pour réduire l'empreinte carbone des modèles d'apprentissage machine.

Choisissez judicieusement vos fournisseurs de calcul infonuagique; sélectionnez l'emplacement du centre de données; réduisez les ressources gaspillées; choisissez du hardware plus efficace; utilisez une calculatrice d'émissions pour des modèles en ML; divulgez les émissions associées aux résultats de ML publiés.

médianne: 3.6

médianne de l'examen: 14.82