Examen 1ère session (2h) - 17 mai 2024

Rappels : **Seul document autorisé : feuille A4 manuscrite, recto-verso.** Les calculatrices et autres appareils électroniques doivent être éteints et rangés. Le barème (sur 20) n'est donné qu'à titre indicatif.

Attention : soignez la présentation de votre copie ainsi que votre écriture ! Toute réponse illisible ou incompréhensible sera notée 0.

Exercice 1 *Cours (3pts)*
On considère un ensemble fini $\mathcal{U}$ d'éléments : $\mathcal{U} = \{x_1, x_2, \ldots, x_n\}$.

Q. 1. Donner l'expression de la fonction caractéristique χ_E d'un sous-ensemble E de $\mathcal{U}$.

Q. 2. Donner l'expression de la fonction caractéristique $\chi_\emptyset$ de l'ensemble vide.

Q. 3. Soit E et F deux sous-ensembles de $\mathcal{U}$ définis par leurs fonctions caractéristiques χ_E et χ_F respectivement. Donner l'expression des fonctions caractéristiques des ensembles $E \cup F$, $E \cap F$ et E^c en fonction de χ_E et χ_F.

Q. 4. Soit E un sous-ensemble de $\mathcal{U}$ et on note c_E son centre de gravité. En utilisant la fonction caractéristique χ_E, donner l'expression de l'inertie intra-cluster de E.

Exercice 2 *Arbres de décision (4pts)*
On considère la base d'apprentissage suivante :

Température	Saison	Pluie	Activité
élevée	printemps	non	pas de sortie
élevée	été	oui	sortie
basse	printemps	non	sortie
basse	printemps	oui	pas de sortie

Q. 1. En utilisant l'algorithme de construction d'arbres de décision vu en cours, et en détaillant les étapes et calculs réalisés, construire un arbre de décision permettant de prédire l'activité connaissant les valeurs pour les 3 attributs de description (température, saison et pluie).
Indications : quelques valeurs de logarithme en base 2 : $\log(\frac{1}{2}) = -1$, $\log(\frac{1}{3}) = -1.58$, $\log(\frac{2}{3}) = -0.58$, $\log(\frac{1}{4}) = -2$ et $\log(\frac{3}{4}) = -0.42$.

Exercice 3 *Descente de gradient, fonctions de coût et évaluation (5pts)*
Avec les notations du cours, une donnée décrite vectoriellement est notée $\mathbf{x} \in \mathbb{R}^d$ et associée à une étiquette y. Une base de données entière est notée : $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_N, y_N)\}$. On se limite dans cet exercice à l'étude des modèles linéaires : l'objectif est donc d'apprendre un vecteur $\mathbf{w}$.

Q. 1. Donner la dimension du vecteur $\mathbf{w}$.

Q. 2. L'idée générale d'une fonction de coût est de mesurer les erreurs d'un système d'apprentissage automatique. Par exemple, dans le cas de classification, on pourrait simplement compter le nombre d'erreurs de classification. Cependant, cette dernière proposition n'est pas utilisée en pratique : pourquoi ? Citer deux fonctions de coût classiques (noms des fonctions et expressions mathématiques). Expliquer les points forts et faibles de ces fonctions de coût.

Q. 3. Rappeler le but et le principe général de l'algorithme de descente de gradient et décliner cet algorithme sur l'une des fonctions de coût précédentes. Expliquer le principe d'une descente de gradient stochastique ET le principe de la descente de gradient classique. Expliquer clairement ce qui est fourni en entrée et ce qui est obtenu en sortie de l'algorithme.

Q. 4. Comment évaluer le modèle appris ? Expliquer la procédure et rappeler brièvement le principal piège à éviter.

Q. 5. Discuter la complexité algorithmique et les performances des algorithmes de descente de gradient par rapport à l'algorithme des k-plus proches voisins.

Exercice 4 *Perceptron (4pts)*

On considère un perceptron simple avec deux entrées notées x_1 et x_2 et une sortie y telle que :

$$y = \begin{cases} 1 \text{ si } w_1 x_1 + w_2 x_2 - w_0 > 0 \\ 0 \text{ sinon} \end{cases}$$

On considère que 1 représente la valeur vrai et 0 la valeur faux.

Q. 1. Trouver des poids pour que le perceptron calcule la fonction ET logique.

Q. 2. Même question avec la fonction OU logique.

Q. 3. Essayer de trouver des poids pour la fonction XOR et commenter.

Q. 4. Construire un réseau de neurones qui calcule la fonction XOR.

Exercice 5 *Apprentissage non-supervisé (4pts)*

Les questions de cet exercice sont indépendantes.

Q. 1. Quelles sont les différences entre l'algorithme des k-moyennes et l'algorithme des k-médoïdes ?

Q. 2. On se place dans un espace de dimension 2 où chaque exemple est représenté par un couple de valeurs réelles. On note d_1 la distance entre 2 groupes d'exemples calculée par l'approche *centroid linkage*, et d_2 la distance calculée par l'approche *complete linkage*. Donner, en justifiant votre réponse et en illustrant par une représentation graphique, un exemple de 3 clusters A, B et C contenant chacun exactement 4 exemples et tels que : $d_1(A, B) > d_1(A, C)$ **et** $d_2(A, B) < d_2(A, C)$.

Q. 3. On se place dans un espace vectoriel réel de dimension 2. Après avoir rappelé les indices de Dunn et de Xie-Beni présentés en cours, donner, en justifiant votre réponse, les caractéristiques d'un clustering optimal pour chacun de ces 2 indices.

Q. 4. Soit `A` et `B` deux array numpy contenant des exemples d'apprentissage et possédant le même nombre de colonnes. Donner le code python **efficace** permettant de calculer la distance entre `A` et `B` en utilisant l'approche *simple linkage*.