Examen 1ère session (2h) - 13 mai 2022

Rappels : **Aucun document n'est autorisé.** Les calculatrices et autres appareils électroniques doivent être éteints et rangés. Le barème (sur 20) n'est donné qu'à titre indicatif.

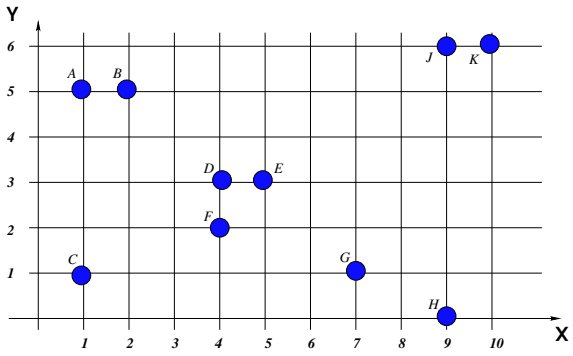
Exercice 1 *Clustering (5 points)*

Q. 1. Soit un ensemble d'exemples $\mathbf{X} \in \mathbb{R}^{n \times d}$ et soit $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ avec pour tout $k = 1, \ldots, K$, $\mathcal{C}_k$ un sous-ensemble de $\mathbf{X}$. Donner les 2 propriétés que doivent vérifier les $\mathcal{C}_k$ pour que $\mathcal{P}$ soit une bonne partition de $\mathbf{X}$ en K clusters.

Q. 2. Après avoir énoncé par une phrase ce que cette mesure représente, donner l'expression de J_k la mesure de l'inertie du cluster $\mathcal{C}_k$ de $\mathbf{X}$.

Q. 3. Soit d une mesure de distance, et soit $A = \{a_1, a_2, \ldots, a_{|A|}\}$ et $B = \{b_1, b_2, \ldots, b_{|B|}\}$ deux groupes d'exemples. Donner l'expression de $dist(A, B)$ la distance entre A et B en utilisant l'approche *simple linkage*. Pour illustrer, donner un schéma permettant d'illustrer la distance ainsi calculée.

Q. 4. On considère l'ensemble $\mathbf{X}$ dont les données sont affichées dans la figure ci-contre. Pour simplifier les calculs, on considère que ces données n'ont pas besoin d'être normalisées.



a) En appliquant l'algorithme de clustering hiérarchique utilisant l'approche *simple linkage*, construire le dendrogramme correspondant à ces données. Vous donnerez les groupes qui sont construits à chaque étape de l'algorithme.

b) Combien de groupes (et lesquels) obtient-on pour une distance inférieure à 3 ?

Exercice 2 *Apprentissage supervisé (5 points)*

Dans cet exercice, on va étudier l'utilisation d'une mesure différente de l'entropie de Shannon pour sélectionner les attributs dans l'algorithme de construction d'arbres de décision. La mesure que l'on souhaite utiliser est l'index de Gini qui, pour un ensemble d'éléments X se calcule par :

$$H_G(X) = 1 - \sum_{x \in X} p^2(x).$$

Q. 1. Tracer la courbe de de $H_G(X)$ sur $[0, 1]$, dans le cas où X ne comporte que 2 éléments et comparer avec celle de l'entropie de Shannon. Est-ce que l'index de Gini peut être utilisé pour remplacer l'entropie de Shannon pour la construction d'un arbre de décision ?

Q. 2. On considère une urne contenant des boules et des cubes qui sont de 2 couleurs (noir ou blanc) et dont la répartition est donnée dans la table ci-contre.

	Blanc	Noir
Boule	4	2
Cube	1	3

La forme de l'objet (boule ou cube) est l'attribut de description de l'objet, sa couleur (blanc ou noir) est sa classe Y. Dans cette question, on utilise l'index de Gini comme mesure de désordre.

a) Calculer $H_G(Y)$ le désordre de la classe dans l'urne.

b) Calculer $H_G(Y|Boule)$ et $H_G(Y|Cube)$.

c) Calculer le gain d'information $I_G(Forme, Y)$ obtenu en utilisant la forme pour prédire la classe.

Q. 3. On considère une urne contenant 20 objets (boule et cube) : 10 objets noirs et 10 objets blanc.

a) Donner une répartition des objets en boule et cube de telle façon que le gain d'information apporté par la connaissance de la forme soit nul (remarque : il faut qu'il y ait obligatoirement les 2 types d'objets).

b) Avec la répartition que vous avez donné en a), est-ce que le résultat précédent (gain d'information nul) dépend du choix de la mesure (entropie de Shannon ou index de Gini) utilisée ? Expliquer.

Exercice 3 *Ethique et apprentissage (2 points)*

L'essor des algorithmes d'IA dans la vie quotidienne pose des problèmes éthiques (*fairness* en anglais). Prenons le cas –aujourd'hui très réaliste– d'un algorithme évaluant une attribution de crédit à un particulier par une banque. Cet algorithme émet un avis sur toutes les demandes de crédit et pèse très lourd dans l'acceptation ou le refus d'un dossier.

Q. 1. Quelles sont les entrées et sorties du système considéré ? De quel type de problématique de machine learning s'agit-il ?

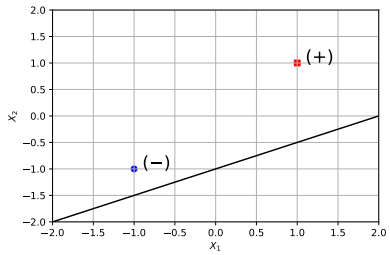
Q. 2. Donner quelques exemples de caractéristiques descriptives (en entrée du système) qui vous semblent –intuitivement– pertinente pour juger un dossier de demande de crédit.

Q. 3. En imaginant que les banquiers présentent beaucoup de caractéristiques à l'algorithme, quel problème éthique peut se poser ?

Q. 4. Le fait d'interdire la manipulation de certain type de données n'est parfois pas suffisant : pouvez-vous imaginer un cas de figure illustrant une telle situation ?

Exercice 4 *Perceptron (2 points)*

Soit le jeu de données minimaliste ci-contre et le perceptron à l'instant t courant étant défini par ses poids : $\mathbf{w} = [0.5, -1]$ et son biais $b = -2$.



Q. 1. Justifier rapidement que la frontière de décision correspond bien aux poids ci-dessus.

Q. 2. Quel point est bien classé, quel point est mal classé ? Donner les scores.

Q. 3. En prenant un pas de mise à jour de $\epsilon = 0.75$, quelle seront les paramètres du modèle après la première epoch ? Attention, ces valeurs dépendantes de l'ordre dans lequel sont vus les points : indiquer les deux résultats possibles et expliquer à quels cas de figure respectifs ils correspondent.

Exercice 5 *Rétro-ingénierie (2 points)*

Soit les notations usuelles suivantes dans la classe des modèles de classification binaires linéaires :

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{pmatrix}, \forall i\ \mathbf{x}_i \in \mathbb{R}^d \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \forall i\ y_i \in \{-1, 1\}, \qquad f(\mathbf{x}_i) = \mathbf{w}\mathbf{x}_i$$

Q. 1. En étudiant le code suivant (qui est fonctionnel, avec tous les imports classiques déjà effectués)

— Indiquer quel est l'algorithme générique implémenté (et nommer ses principales étapes).

— Retrouver la formulation du problème d'apprentissage associé et expliciter le critère de convergence.

```
def train_mystere(X, Y, epsilon):
  n,d = X.shape
  w = np.zeros((d,1))
  while np.where( X@w * Y > 0, 0, 1).sum() >0 :
    index = np.random.permutation(n)
    for i in index:
        xi = X[i]
        yi = Y[i]
```

```
        if xi@w - yi > 0:
            w = w - epsilon * xi
        else:
            w = w + epsilon * xi
    return w
```

Exercice 6 *Ingénierie (3 points)*

Soit l'exercice symétrique du précédent et reprenant les mêmes notations. Nous voulons optimiser la fonction fr coût suivante, correspondant à une sorte de *perceptron carré* :

$$\mathcal{C} = \sum_{i=1}^{n} [(-\mathbf{x}_i \mathbf{w} y_i)_+]^2, \qquad \text{avec la défintion : } (a)_+ = \begin{cases} a & \text{si : } a \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Q. 1. Donner le code de la méthode

Q. 2. Quelles sont les erreurs qui sont plus pénalisées que dans l'algorithme du perceptron, quelles sont celles qui le sont moins ?

Exercice 7 *Evaluation black box (1 points)*

Soit un problème de classification binaire équilibré réputé difficile reprenant encore une fois les notations précédentes. Un expert métier vous indique qu'une performance de 90% est souhaitée mais que les résultats plafonne actuellement à 60%. Un des ingénieurs de votre équipe a eu accès à un modèle très performant (`blackbox_xgb`) et vous propose le code (fonctionnel) suivant :

```
// chargement des donnees X, Y etc...
mod  = blackbox_xgb.train(X,Y)
pred = blackbox_xgb.predict(X)
perf = np.where(pred == Y, 1, 0).mean()

if perf >= 0.9:
  print("Bravo, la solution est prete pour la commercialisation")
elif perf >= 0.6:
  print("OK, vous avez atteint la performance de base")
else :
  print("Echec: votre approche n'atteint pas les performances de reference")
```

Q. 1. Que pensez-vous du travail de votre ingénieur ? Pourquoi ?

Q. 2. (bonus culturel) Quelle est cette approche mystère selon vous ?