Examen 2ème session (2h) - 28 juin 2018

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

Exercice 1 *Questions de cours (5 points)*
On considère une base d'apprentissage $\mathcal{X}$ contenant n exemples décrits par p attributs.
Soit $x_1 = (x_{11}, x_{12}, \ldots, x_{1p})$ et $x_2 = (x_{21}, x_{22}, \ldots, x_{2p})$, deux exemples de $\mathcal{X}$.

Q. 1. Donner l'expression de $x_1 \cdot x_2$ le produit scalaire de x_1 et de x_2 en fonction de leurs coordonnées.

Q. 2. Donner la valeur de x_2 telle que $x_1 \cdot x_2$ soit égal à la moyenne des composantes de x_1.

Q. 3. Montrer que la distance de Manhattan est bien une mesure de distance.

Q. 4. Qu'est-ce que la méthode *out of bags* ? Quel est son principe ? Quels sont ses avantages ? En quoi se différencie-t-elle de la méthode par *cross validation* ?

Q. 5. Qu'est-ce que le sur-apprentissage et le sous-apprentissage ? Comment les identifier lors de l'apprentissage d'un modèle ? Comment les éviter ?

Q. 6. Expliquez brièvement ce qui caractérise l'apprentissage par renforcement.

Exercice 2 *(3 points)*
Un collègue vous fournit un ensemble de données correspondant à un problème qu'il voudrait automatiser : il s'agit d'une table de $n = 1000$ exemples qui ressemblent à ceci :

alt	**tai**	**grp1**	**grp2**	**re**
225	0.03	A	1	BON
3800	-0.23	B	3	MAUVAIS
2750	-2.52	A	2	BON
327	1.27	C	1	MAUVAIS
...	...	...	...	...

Q. 1. Est-ce un problème d'apprentissage supervisé ou par renforcement ? Pourquoi ?

Q. 2. Quelle solution proposer pour réaliser la prédiction de la colonne **re** ? (quel modèle ? quel traitement des données ? quelle mesure pour valider l'approche proposée ? etc...)

Exercice 3 *(6 points)*
Un fonctionnaire décide d'augmenter son salaire en jouant régulièrement au PMU (les courses de chevaux). Afin d'optimiser ses chances de gagner, il développe un modèle d'apprentissage statistique afin d'augmenter ses gains.

Le PMU propose différentes manières de jouer (tiercé, quinté, 2 sur 4,...). À l'aide du quotidien "Paris – TURF", il obtient chaque matin, pour chaque course, une information sur les chevaux y participant. Un cheval est décrit par un vecteur de caractéristiques réelles : son âge, sa vitesse, son endurance, sa note moyenne,... (Attention ! deux chevaux peuvent avoir la même description). De même, une course est aussi décrite par un vecteur de caractéristiques.

Le joueur dispose d'un historique de plusieurs centaines d'exemplaires du "Paris – TURF" parus durant ces dernières années et des résultats des courses qu'ils décrivent.

Q. 1. Notre joueur s'intéresse tout d'abord à développer un modèle capable de prédire si un cheval termine ou non une course (*ie.* prédire "oui" ou "non"). On peut remarquer que le fait de terminer une course ne dépend pas des autres participants à la course.

1a) Dans quel cadre d'apprentissage doit on se placer ?

1b) Décrire le(s) modèle(s) utilisé(s) pour réaliser une telle prédiction en détaillant : l'espace de description des données, le type de modèle utilisé, l'espace de sortie du modèle, le nombre de paramètres appris par le modèle, les données d'apprentissage, la méthode d'apprentissage utilisée.

1c) Avant d'utiliser le modèle appris, il est nécessaire d'estimer s'il fournit de bonnes prédictions.

i) Sur quelles données doit-on entraîner et tester ce modèle ?

ii) Quelle mesure d'évaluation utiliser sachant que, si le modèle se trompe, la mise du pari est perdue et si le modèle a raison, la mise est remportée (quitte ou double) ?

1d) On souhaite utiliser le modèle pour savoir si un nouveau cheval terminera ou non une course. Décrire la procédure à suivre.

Q. 2. On s'intéresse maintenant à savoir si un cheval va gagner ou non une course. Pour chaque course, il y a 10 chevaux partants. Dans quel cadre d'apprentissage se place-t-on ?

Q. 3. Décrire le(s) modèle(s) utilisé(s) pour cette prédiction en précisant : l'espace de description des données, le type de modèle utilisé, l'espace de sortie du modèle, le nombre de paramètres appris par le modèle, les données d'apprentissage, la méthode d'apprentissage utilisée.

Q. 4. Décrire la procédure à suivre afin de pouvoir utiliser ce modèle pour déterminer si un nouveau cheval terminera ou non une course.

Q. 5. Proposer un autre modèle d'apprentissage équivalent à celui de la question précédente.

Exercice 4 *(6 points)*

On considère la base d'apprentissage représentée dans la figure donnée en Annexe. Cette base contient 20 exemples, dont la description est le couple représenté par leurs coordonnées (x, y), et la classe est soit *rond* (notée R) soit *carré* (notée C).

Dans ce qui suit : utiliser l'Annexe (à rendre) pour les réponses graphiques demandées.

Q. 1. En utilisant l'algorithme des k plus proches voisins, avec $k = 3$ et la distance euclidienne, représenter graphiquement la frontière de séparation des classes et donner, en justifiant, la classe des 5 points suivants : le point L de coordonnées $(3, 3)$, M de coordonnées $(10, 3)$, N de coordonnées $(7, 4)$, P de coordonnées $(9, 5)$, et Q de coordonnées $(4, 2)$. En cas d'égalité de distances, les points de classe R seront considérés en priorité.

Q. 2. On considère les 4 seuils de coupure suivants :

— sur X : coupure c_1 : $x \leq 3.5$

— sur X, coupure c_2 : $x \leq 8.5$

— sur Y, coupure c_3 : $y \leq 4.5$

— sur Y, coupure c_4 : $y \leq 3.5$

En utilisant uniquement ces 4 seuils, construire un arbre de décision binaire en donnant le détail des calculs d'entropie réalisés. Donner une représentation graphique de cet arbre et représenter graphiquement la frontière de séparation entre les classes correspondante.

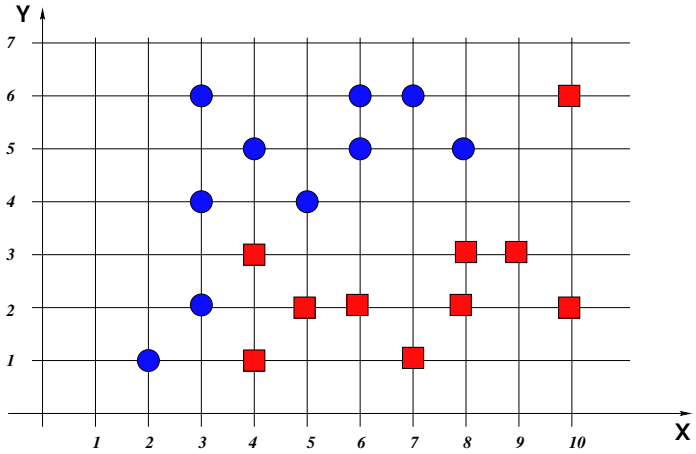
Q. 3. On considère que la classe R correspond à la valeur $+1$ et la classe C correspond à la valeur -1 et on décide d'utiliser l'algorithme du perceptron. Sans dérouler l'algorithme, mais en justifiant votre réponse, tracer la frontière de décision obtenue. Quelle est la particularité de cette frontière ?

Q. 4. En fait, les points M et P sont de la classe *carré* et L, N et Q sont de la classe *rond*. Donner la matrice de confusion pour chacun des modèles appris dans les questions précédentes (k-ppv, arbre de décision, et perceptron). Quel est le taux d'erreur de chacun de ces modèles ? Lequel est préférable ?

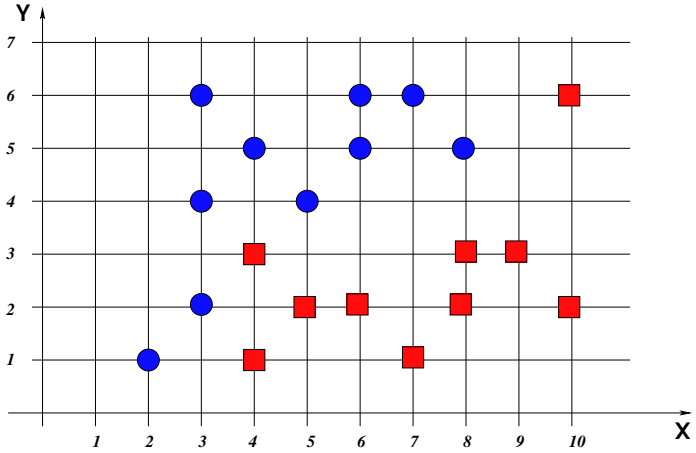
NUMERO D'ANONYMAT :

À rendre avec votre copie

Exercice 2 - Question 1 : (k-ppv)



Exercice 2 - Question 2 : (arbre)



Exercice 2 - Question 3 : (perceptron)