TD2
Apprentissage Supervisé

Exercice 1

1. Dites sur quel principe se base l'apprentissage supervisé ?

2. Quelles sont les différentes méthodes utilisées en apprentissage automatique pour défaire la *complexité de la non-linéarité*.

3. Pourquoi la notion de ressemblance basée sur la proximité utilisant une métrique est subjective. Expliquer par un exemple sur les classes de visages : homme et femme.

4. Interpréter la formule de la fonction *Gain()* (sa valeur max et min) et du mélange (Gini ou Entropie).

5. Écrire un pseudo-algorithme qui permet de construire un arbre de décision à partir d'un ensemble d'exemples d'apprentissage.

Exercice 2

1. Ecrire un algorithme (procédure ou fonction) qui calcule le mélange (selon Gini) au niveau d'un nœud d'une population P ayant n individus et $m = 3$ classes d'appartenances $C = \{c_1, c_2, c_3\}$.

Pour chaque classe c_i, on connait le nombre n_i d'individus à ce nœud.

2. Ecrire un programme (procédure ou fonction) qui calcule le g*ain G* d'une variable test v au niveau d'un nœud d'un AD pour une population $P = 8$ individus à ce nœud.

   – Le mélange initial de P (selon Gini) à ce nœud est : $m_P = \frac{15}{32}$. La variable v comporte n modalités (ou valeurs possibles), prenant le cas où $n = 2$: ''*oui*'' et ''*non*''.

   – Pour $i = 1$ (les ''*oui*'' de la population), $P_{i=1} = 5$ et son mélange $m_{P_{i=1}} = \frac{12}{25}$.

   – Pour $i = 2$ (les ''*non*'' de la population), $P_{i=2} = 3$, et où le mélange est $m_{P_{i=2}} = 0$.

- Réponse du programme : $G = 0.169$

Exercice 3 –ALGORITHME ID3

- L'idée est de pouvoir déterminer les moyens de transports d'une personne, selon que la personne est Homme ou Femme, possédant ou non un (ou des) véhicule(s), son niveau de revenu et si elle peut se payer un voyage de luxe ou pas.

- Le problème consiste à construire un classifieur (AD) de type ID3 en se basant sur l'entropie comme fonction de calcul de quantité d'informations dans un nœud (population).

- L'espace de descriptions comprend 4 attributs : le sexe, le nombre de véhicule en possession, les frais de voyage et le niveau de revenu. Le mode de transport sera l'attribut cible (la classe d'appartenance) pouvant prendre 3 des valeurs : Bus, Train ou Voiture.

Exercice 3 (suite)

Id. Personne	Sexe (S)	Nombre Véhicule Possédé (V)	Frais de Voyage (F)	Niveau de revenu (R)	Mode de Transport (T)
1	H	0	Moins cher	Faible	Bus
2	H	1	Moins cher	Moyen	Bus
3	F	1	Moins cher	Moyen	Train
4	F	0	Moins cher	Faible	Bus
5	H	1	Moins cher	Moyen	Bus
6	H	0	Moyen	Moyen	Train
7	F	1	Moyen	Moyen	Train
8	F	1	Cher	Elevé	Voiture
9	H	2	Cher	Moyen	Voiture
10	F	2	Cher	Elevé	Voiture

- A partir de l'ensemble d'apprentissage des 10 exemples, **construisez un arbre de décision ID3** qui permettra par la suite de classer et reconnaître le moyen de déplacement d'une personne.

Exercice 4

- Soit l'ensemble d'apprentissage

$$E_A = \left\{ \begin{array}{c} e_1 = [(0,1),R]; e_2 = [(2,5),R]; e_3 = [(7,15),R]; e_4 = [(8,17),R]; \\ e_5 = [(5,19),V]; e_6 = [(0,-1),V]; e_7 = [(3,11),V]; e_8 = [(9,35),V]; e_9 = [(1,-1),B]; \\ e_{10} = [(3,7),B]; e_{11} = [(4,14),B]; e_{12} = [(10,98),B] \end{array} \right\}$$

formé des points du plan image 2D.

Chaque échantillon e_i est donné par ses coordonnées (x_i, y_i) et sa classe d'appartenance. Trois classes : Rouge (R), Verte (V) et Bleue (B) sont définies.

1. En utilisant le classifieur *k-ppv* (k=3) et la distance Euclidienne, déterminer la classe des exemples suivants :

$e_j = [(8,62),B]$; $e_k = [(7,26),V]$; $e_l = [(6,13),R]$; $e_m = [(5,11),R]$; $e_n = [(4,15),V]$; $e_o = [(9,79),B]$;

2. Donner un échantillon $(e_j = ?)$ qui appartient à la fois à la classe R et V.
3. Ecrire un programme qui permet de déterminer la classe de l'échantillon $e_x = (6,15)$ (ou $e_y = [(7,26),V]$) selon le classifieur *k-nn* (pour k=1), en utilisant la distance de Manhattan.
4. Déterminer la ou les fonctions (équations) qui décrivent ce modèle d'apprentissage à 3 classes dans le plan image 2D.

Exercice 5 (1/2)

SVM

Soit l'ensemble de données d'apprentissage suivant :

$$E_A = \left\{ \begin{array}{c} \boldsymbol{e_1} = \left[\boldsymbol{x_1} = \binom{\mathbf{3}}{\mathbf{1}}, \boldsymbol{y_1} = \boldsymbol{O}\right]; e_2 = \left[\binom{3}{-1}, O\right]; e_3 = \left[\binom{6}{1}, O\right]; e_4 = \left[\binom{6}{-1}, O\right]; \\ e_5 = \left[\binom{1}{0}, NO\right];\ e_6 = \left[\binom{0}{1}, NO\right]; e_7 = \left[\binom{0}{-1}, NO\right]; e_8 = \left[\binom{-1}{0}, NO\right] \end{array} \right\}$$

formé des points du plan image 2D. Chaque échantillon e_i est donné par ses coordonnées $\binom{a_i}{b_i}$ et sa classe d'appartenance y_i : $Objet\ (O)$ ou $NonObjet\ (NO)$.

Pour des besoins de programmation, on prend les 2 classes {+1, -1}, tel que à un vecteur d'entrée x_i correspond une sortie $\boldsymbol{f_{w,b}}(x_i) = y_i \in \{+1, -1\}$. Ce classifieur est donné par :

$$\boldsymbol{Classifieur}(x_i) = \boldsymbol{\varphi}(x_i) = \boldsymbol{signe}(\boldsymbol{f_{w,b}}\ (\boldsymbol{x_i}) = \boldsymbol{y_i} = \boldsymbol{w}\ .\ x_i + \boldsymbol{b})$$

- w et b : coefficients ou paramètres du classifieur SVM (w est le vecteur de poids, et b le biais)
- $\boldsymbol{f_{w,b}}(\boldsymbol{x}) = \boldsymbol{w}.\ \boldsymbol{x} + \boldsymbol{b}$ est notre fonction discriminante linéaire à programmer telle que :

$$\boldsymbol{f_{w,b}}(\boldsymbol{x}) \geq \mathbf{0} \text{ si x } \in +1 \text{ (classe O) et } \boldsymbol{f_{w,b}}(\boldsymbol{x}) < \mathbf{0} \text{ si x } \in -1 \text{ (classe } NO)$$

Exercice 5 (2/2)

SVM

- Calculer les vecteurs supports (à partir de leur distance par rapport à la droite de l'hyperplan), puis
- Donner une fonction qui détermine les vecteurs de supports et leur nombre
- Donner une procédure qui détermine l'équation de l'hyperplan SVM (calcule w et b),
- Donner l'algorithme :
    1. En utilisant la bibliothèque sklearn;
    2. Sans l'utilisation de sklearn.

Exercice 6

Le Neurone formel (Perceptron) et problème du XOR

- Pourquoi le perception ne peut modéliser l'opérateur XOR ?
- Donner un modèle de neurone ou de réseau de neurones qui résout ce problème.
- Schématiser le réseau de neurones (nombres: d'entrées, de neurones, de couches, et de sorties) qui résout la fonction XOR, puis,
- Donner l'algorithme de ce réseau :
    1. En utilisant la bibliothèque sklearn;
    2. Sans l'utilisation de sklearn.