



Série de TD N°3

Apprentissage & Classification Supervisés

M. Taffar

Dpt Informatique. Université de Jijel

Cours pour Master 1 –IA, nov.-21

Exercice 1

1. Dites sur quel **principe** se base l'**apprentissage supervisé** ?
2. Que représente en termes d'apprentissage les notions : **généralisation** et **intelligibilité** pour l'intelligence d'un système.
3. En faisant les déclarations convenables, écrire un **pseudo-algorithme** qui **permet de construire un arbre de décision** à partir d'un ensemble d'exemples d'apprentissage.

Exercice 2

- 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 connaît le **nombre n_i d'individus à ce nœud**.

Exercice 3

- Ecrire un programme (procédure ou fonction) qui **calcule le gain 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 4

- Considérant un espace de descriptions comprenant 4 attributs : la **Forme**, la **Couleur**, la **Texture** et la **Densité**, prenant respectivement les valeurs suivantes :
 - **Forme** = {conique, circulaire, ovale}.
 - **Couleur** = {rouge, verte, bleue}.
 - **Texture** = {dégradée, mosaïque, simple}, et
 - **Densité** = {oui, non}.
 - **Nature** : est l'attribut cible (la classe d'appartenance), binaire de valeurs : **oui** (pour valeur authentique) et **non** (pour la valeur non authentique).

Id. Objet	Forme (F)	Couleur (C)	Texture (T)	Densité (D)	Nature
1	ovale	verte	simple	non	Non-Authentique
2	circulaire	bleue	dégradée	oui	Non-Authentique
3	conique	bleue	dégradée	oui	Non-Authentique
4	circulaire	rouge	dégradée	oui	Authentique
5	ovale	verte	mosaïque	oui	Authentique
6	ovale	bleue	mosaïque	non	Non-Authentique
7	conique	verte	mosaïque	non	Non-Authentique
8	circulaire	verte	simple	oui	Authentique

Exercice 4 (suite)

- De l'ensemble d'apprentissage de 8 exemples, *construisez un arbre de décision* qui permettra par la suite de reconnaître la nature d'un nouvel objet à partir de ses caractéristiques.
- En s'inspirant de l'ensemble des 8 exemples d'apprentissage, *écrivez un algorithme qui permet de construire un AD*, comprenant les parties (procédures ou fonctions) suivantes :
 - *Initialisation()* ;
 - *Calcul_Mélange()* ;
 - *Calcul_Gain* (de chaque attribut) ;
 - *Déterminer_si_Nœud_Terminal* (Classe) ;
 - *Construction_Arbre* : Nœud et les différentes branches (liste chaînée) ;
 - *Mise_à_Jour_Table* (table de données) ;
 - *Programme Principal*.
- Dérouler votre algorithme sur l'exemple 5 (de la table précédent).
- Donnez la classe (reconnaissez la nature) du nouvel objet O_x ayant le vecteur descripteur suivant :

$$O_x = \{\text{conique, rouge, simple, oui}\}$$

Exercice 5 –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 5 (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 6

- Soit l'ensemble d'apprentissage

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

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 7 (1/5)

PARTIE 1

- En utilisant les listes chaînées, écrire un programme Python qui permet de construire un arbre de n (ex., $n=64$) niveaux (la racine : est le niveau 0). Ce programme doit comprendre :
- Une fonction qui doit permettre **d'insérer ou d'ajouter** (selon une condition) un nouveau élément (nœud) à l'arbre à une position bien définie : **(1)** En tant que nœud feuille ; ou **(2)** En tant que nœud descendant ayant b branches à définir (valeur maximale de $b=10$),
- Une fonction qui permet de parcourir l'arbre à la recherche d'un élément (sur une valeur de nœud), puis soit :
 - d'afficher ou de modifier (mettre à jour) sa valeur et son niveau dans l'arbre, soit
 - de supprimer cet élément (nœud) de l'arbre et ses branches menant aux nœuds descendants, et de mettre la structure de l'arbre à jour.

Exercice 7 (suite 2/5)

PARTIE 2

- Soit un ensemble d'apprentissage $X = \{x_1, \dots, x_m\}$.
- Chaque échantillon ou exemple $x_i = (a_1, a_2, \dots, a_p, c_j)$ est un descripteur d'objet à p paramètres auquel s'ajoute sa classe d'appartenance $c_j, j = 1 \dots k$, pour un problème à k classes.
- Chaque variable $a_t, t = 1 \dots p$, ne peut prendre que 3 modalités (valeurs) au maximum (modalités à définir dans la déclaration du programme).
- Utiliser le programme précédent (de la partie 1) pour construire un arbre de décision pour l'ensemble d'apprentissage X . Dans ce cas, chaque nœud (élément) de l'arbre (liste) doit comprendre les champs suivants :

Exercice 7 (suite 3/5)

PARTIE 2

- Niveau dans l'arbre (code),
- Son intitulé (nœud feuille –Classe ou nœud test –variable),
- Sa valeur de **mélange initiale** de la population d'apprentissage qui en résulte à ce nœud, de la partition précédente,
- Les valeurs de mélanges de la variable test pour chaque modalité (si la variable peut prendre v valeurs, alors v valeurs de mélanges sont calculer ; avec une possible construction, pour cette variable, d'un nœud à v branches dans l'arbre, si son gain est majoré),
- Pointeurs vers les nœuds suivants (si nœud n'est pas une classe). Chaque pointeur sera étiqueté par la valeur (modalité) de la variable test au nœud supérieur.

Exercice 7 (suite 4/5)

PARTIE 2

- En plus, des fonctions précédentes de manipulation de la structure de l'arbre, écrire une fonction pour :
- Calculer le mélange initial d'une population p_i à un nœud de l'arbre.
- Calculer le mélange d'une variable test a_t pour une valeur a_{ti} d'une branche de l'arbre.
- Calculer le gain au niveau d'un nœud pour une variable a_t , $t = 1 \dots p$, qui peut prendre $2 \leq b \leq 5$ valeurs a_{t1} , a_{t2} , a_{t3} , a_{t4} et a_{t5} (la valeur de b dépend de la variable test et elle est variable d'un nœud à un autre).
- Comparer les gains des variables pour choisir la variable test à un nœud de l'arbre.

Exercice 7 (suite 5/5)

PARTIE 2

- Enfin, tester votre programme sur un jeu de données de votre choix (par exemple, celui vu au cours) ou que vous synthétisez, en prenant :
 - $m = 12$ (taille de l'ensemble d'apprentissage),
 - $p = 4$ (le nombre de modalités \mathbf{b} dépend de la variable test et peut varier d'une variable à un autre) et
 - $k = 2$ (nombre de classes).