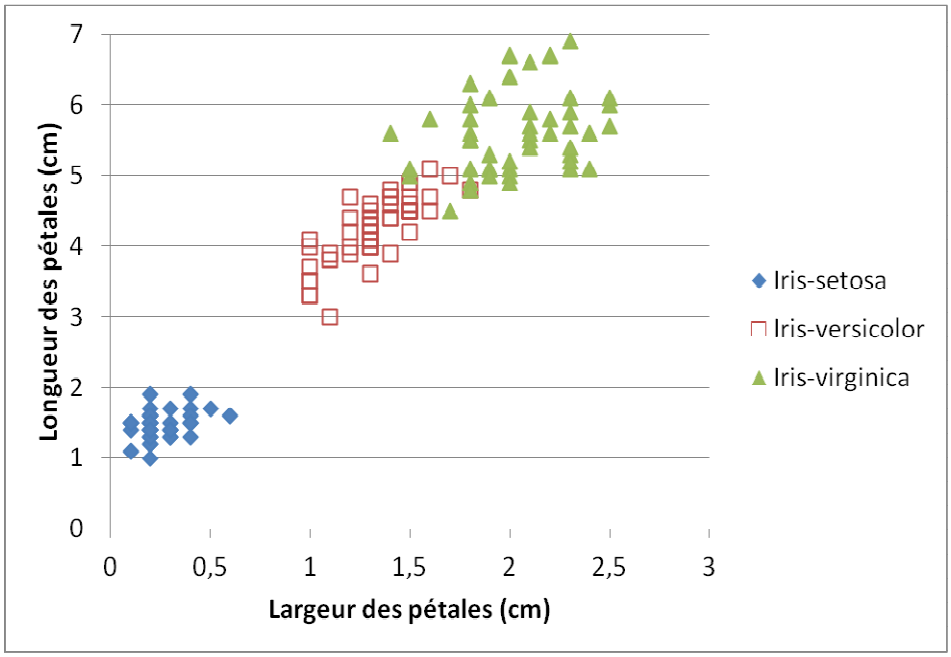
CY TECH	**ING2-GI : MACHINE LEARNING** **ARBRE DE DECISION ET RANDOM FOREST**
Rédigé par : Astrid Jourdan	A l'intention de : Elèves d'ING2-GI
Durée : 6h	Dernière modification : 21/09/2020

Exercice 1

Considérons le fameux jeu de données « Iris » (http://archive.ics.uci.edu/ml/datasets/Iris) représenté ci-dessous. Il s'agit de trois types d'Iris caractérisés entre autre par la longueur et la largeur des pétales.



1) Représentez sur le graphique la règle : *Si la longueur des pétales est inférieure à 2 alors l'iris est de type Setosa*.
2) Dans le cas où la longueur des pétales est supérieure à 2, ajoutez sur le graphique la règle : *Si la largeur des pétales est supérieure à 1,7 alors l'iris est de type Virginia, sinon l'iris est de type Versicolor*.
3) Représenter l'arbre correspondant à ces règles.
4) Refaire la même chose en inversant l'ordre de longueur et largeur.

Exercice 2

Considérons l'exemple très simple sur le football.

X_1=Match à domicile ?	X_2=Balance positive ?	X_3=Mauvaises conditions climatiques ?	X_4=Match précédent gagné ?	Y=Match gagné
V	V	F	F	V
F	F	V	V	V
V	V	V	F	V
V	V	F	V	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F
V	F	V	F	F

Quelle variable sera choisie à la racine de l'arbre si l'on souhaite maximiser le gain à l'aide de l'entropie ?

Exercice 2

On considère des images en noir et blanc codées sur 4 pixels. Chaque image est donc codée par un élément $(x_1,x_2,x_3,x_4)\in\{0,1\}^4$, où les pixels noirs sont notés 1 et les pixels blancs sont notés 0 et sont numérotés dans l'ordre

x_1	x_2
x_3	x_4

On dispose de la base d'apprentissage ci-dessous pour laquelle les images ont été réparties selon deux classes.

Classe 1 ***Classe 2***

On souhaite construire un arbre de décision sur cet échantillon avec comme variables explicatives (attributs), x_1, x_2, x_3 et x_4.

1) Ecrire la base d'apprentissage sous la forme d'un tableau
2) Quelle variable sera choisie à la racine de l'arbre si l'on souhaite maximiser le gain à l'aide de l'indice de Gini ?
3) Représenter ensuite l'arbre T avec un partage des nœuds suivant les variables dans l'ordre suivant : X_2, X_3, X_1.
4) Proposer un arbre élagué. Calculer l'erreur d'ajustement de chacun des arbres ?
5) On considère l'ensemble de test suivant

Classe 1 ***Classe 2***

Calculer l'erreur de prévision pour les deux arbres.

Exercice 3

L'objectif du tutoriel http://apiacoa.org/blog/2014/02/initiation-a-rpart.fr.html est de dérouler un arbre de décision avec R.

1) A la suite de ce tutoriel, construire un arbre de décision avec R sur le jeu de données « Iris ».
   - Construire une base d'apprentissage et une base test
   - Trouver le bon paramétrage
   - Calculer l'erreur d'apprentissage et l'erreur de prévision

2) Soient les deux jeux de données simulées : Test_Classif_dpt.txt et Test_Classif_Correl.txt
   - Représenter le nuage de points avec la couleur des classes pour les deux jeux. Un arbre de décision peut-il séparer ces classes ?
   - Comparer la complexité des arbres construits dans chacun des cas.

Exercice 4

1) Utiliser le package R randomForest pour construire une forêt sur le jeu de données « iris ».

Ci-dessous un aide-mémoire pour la fonction randomforest :

http://www.duclert.org/r-apprentissage/random-forest-R.php

et un tutoriel

http://perso.ens-lyon.fr/lise.vaudor/classification-par-forets-aleatoires/

- Quelles sont les variables les plus importantes pour classifier les iris ?
- Combien y-a-t-il de paramètres à fixer avant de lancer l'algorithme ?

2) Utiliser la méthode des random forest pour faire de la reconnaissance de caractères avec le jeu de données : http://archive.ics.uci.edu/ml/datasets/Letter+Recognition