FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

TP-Projet DataScience – Partie 1

La cueillette aux champignons

Licence 3 SdN
Matière d'ouverture

Data Science
2020-2021

Ce TP est constitué de 2 séances. Il existe plus de 35 000 sortes de champignons en France. L'objectif est de créer un programme qui détecte si un champignon est comestible ou non, à l'aide de l'algorithme K-NN.

Pour les L3 SdN : vous êtes libres de l'implémenter en Python, en Java ou en C.

Pour les étudiants en matière d'ouverture : il n'est pas nécessaire de coder, vous utiliserez le logiciel Orange (https://orange.biolab.si/). Note : le cours d'initiation à la programmation suivi par certains L3 n'est pas suffisant pour se lancer dans le codage du K-NN car nous n'avons pas vu les listes, ni la gestion des fichiers, ni les structures de données.

La question 1 du projet est différente pour les L3 SdN et les étudiants en matière d'ouverture. Les questions 2 et 3 sont communes.

Les données

Vous avez à votre disposition un jeu de données « artisanal » **tp_mushrooms_dataset_150.csv** qui contient une cueillette de 150 champignons français dont les caractéristiques ont été renseignées par un biologiste. Un pharmacien mycologue s'est ensuite chargé d'indiquer pour chaque champignon s'il est comestible (« e ») ou vénéneux (« p »)[1].

Vous disposez également d'un second jeu de données **tp_mushrooms_dataset_5000.csv** de 5000 champignons, qui est une extraction du guide des champignons Nord-Américains (The Audubon Society Field Guide to North American Mushrooms (1981)).

Question 1 : Implémentation (pour les L3 SdN)

Cahier des charges

- Le programme prend en argument un nom de fichier csv, la désignation d'un individu (champignon, fruit, patient,…) et l'index de la colonne avec la classe à prédire. Exemple :

  knn.py tp_mushrooms_dataset_150.csv champignon 0

- Le programme doit pouvoir fonctionner sur différents types de données : fruits, champignons, patients,… Ces fichiers seront toujours des CSV, avec la première ligne qui contient les noms des colonnes séparés par des virgules et les lignes suivantes représentent les individus.
- Au démarrage du programme, il est possible de préciser la valeur de K
- Une fois paramétré, le programme demande en boucle de saisir un individu (je vous conseille de le copier/coller) puis calcule ses plus proches voisins et suggère une classe.

Par défaut, vous utiliserez une distance euclidienne et un vote majoritaire pour déterminer la classe.

Facultatif : implémentation du vote pondéré par la distance. Plus un individu est proche, plus il compte pour la pondération.

[1] Il s'agit en fait d'une extraction du jeu de données issu de l'UCI Machine Learning repository (The Audubon Society Field Guide to North American Mushrooms (1981)), mais pour des raisons pédagogiques on « va faire comme si » c'était des champignons français ☺

Implémentation

a) Réalisez un parseur pour charger le fichier tp_mushrooms_dataset_150.csv en mémoire. Créez une fonction afficherStats qui affiche le nombre de d'individus chargés, le nombre d'attributs, et le nombre de prédictions

```
Chargement du fichier tp_mushrooms_dataset_150.csv contenant des individus de type champignon
8114 individus de type champignon
23 attributs
prediction : edible(p: 3910, e: 4204)
```

b) Implémentez le calcul de distance entre un nouvel individu et le $i^{ème}$ individu chargé en mémoire.
c) Créez un petit menu pour charger un individu et lancer 1-NN dessus. Vous pouvez déjà compléter le tableau de la question 2 pour K=1.
d) Créez un petit menu pour préciser la valeur de K
e) Implémentez K-NN

```
Entrez un individu à évaluer (valeurs séparées par des ',', comme dans le fichier chargé)p,x,f,g,f,c,f,c,n,n,e,b,s,s,w,w,p,w,o,p,k,v,d
Voisin n° 599 distance :1.0 classe: p
Voisin n° 6698 distance :1.4142135623730951 classe: p
Voisin n° 6148 distance :1.4142135623730951 classe: p
Voisin n° 5992 distance :1.4142135623730951 classe: p
Voisin n° 5113 distance :1.4142135623730951 classe: p
total : p: 5 /5
total pondéré : p: 3.82842712474619
```

Question 1 : K-NN avec Orange (pour les étudiants en matière d'ouverture)

Vous réaliserez le « TP K-NN Orange » pour vous familiariser avec le logiciel Orange. Pour ceux qui ne sont pas familiers avec les TP d'informatique, le fonctionnement est le suivant :

- Suivez la trame du TP, n'hésitez pas à contacter l'enseignant lorsque vous êtes bloqués (après avoir cherché par vous-même 5 min ☺)
- Contrairement à un TD, il n'y a pas de correction. En cas de doute sur une question, vous pouvez faire valider vos réponses par l'enseignant

Question 2 : Evaluation des performances

Vous allez analyser les performances sur le jeu de données de 150 champignons, et sur le jeu de données de 5000 champignons.

Vous préciserez votre protocole expérimental :

- Combien de champignons ont été utilisés pour l'apprentissage
- Combien de champignons vous avez utilisés pour le test

Je recommande 10 champignons si vous évaluez manuellement. Vous pouvez augmenter si vous évaluez automatiquement.

a) Complétez le tableau suivant pour le jeu de données de 150 champignons

Paramètres	K=1	K=5, majorité	K=10, majorité	K=5 majorité pondérée	K=10 majorité pondérée
Moyenne Accuracy					
Moyenne Sensibilité					
Moyenne Confiance					

b) Quel paramétrage faut-il utiliser ? Pourquoi ?

c) Complétez le tableau suivant pour le jeu de données de 5000 champignons

Paramètres	K=1	K=5, majorité	K=10, majorité	K=5 majorité pondérée	K=10 majorité pondérée
Moyenne Accuracy					
Moyenne Sensibilité					
Moyenne Confiance					

d) Quel paramétrage vaut-il mieux utiliser ? Pourquoi ?

e) Que peut-on conclure des résultats sur les deux jeux de données ?

Question 3 : La fricassée de champignons

Le fichier **tp_mushrooms_eval_etu.csv** contient 10 champignons. Proposez une technique pour les classifier. Indiquez lesquels sont comestibles et lesquels sont vénéneux. Peut-on faire confiance à ces prédictions ?

Champignons	Comestible ?
C1	
C2	
C3	
C4	
C5	
C6	
C7	
C8	
C9	
C10	