TD

Classification des Iris

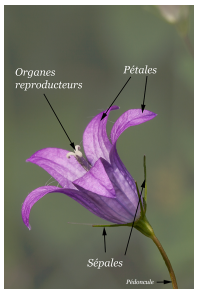
https://makina-corpus.com/blog/metier/2017/initiation-au-machine-learning-avec-python-pratique

Savoirs et compétences :

- ❒ *A voir*

1 Introduction

La base des Iris utilisée est celle réalisée par un botaniste, Ronald Fisher, en 1936 à l'aide d'une clef d'identification des plantes (type de pétales, sépale, type des feuilles, forme des feuilles, ...). Puis, pour chaque fleur classée il a mesuré les longueurs et largeurs des sépales et pétales.



L'idée qui nous vient alors, consiste à demander à l'ordinateur de déterminer automatiquement l'espèce d'une nouvelle plante en fonction de la mesure des dimensions de ses sépales et pétales que nous aurions réalisée sur le terrain. Pour cela nous lui demanderons de construire sa décision à partir de la connaissance extraite des mesures réalisées par M. Fisher.

Autrement dit, nous allons donner à l'ordinateur un jeu de données déjà classées et lui demander de classer de nouvelles données à partir de celui-ci. C'est un cas d'apprentissage supervisé (mais nous le transformerons aussi en non supervisé). Une fois alimentés avec les observations connues, nos prédicteurs vont chercher à identifier des groupes parmi les plantes déjà connues et détermineront quel est le groupe duquel se rapproche le plus notre observation.

2 Visualisation des données

Dans ce TD pour s'affranchir de l'acquisition et de la gestion des données, nous utiliserons un set de données directement disponible dans la bibliothèque sklearn.

```
from sklearn import datasets
iris = datasets.load_iris()
```

L'objet iris contient plusieurs attributs que l'on peut lister avec l'instruction

```
>>> print(dir(iris))
        ['DESCR', 'data', 'feature_names', 'target', 'target_names']
```

Ainsi, on a :

- `DESCR` (str) contient les caractéristiques du set de données ;
- `data` (numpy.ndarray) contient un tableau de 150 lignes et 4 colonnes correspondant au `feature_names` ;
- `feature_names` (list de str) contient le nom des caractéristiques (sepal length (cm), sepal width (cm), petal length (cm), petal width (cm)) ;
- `target` (numpy.ndarray) contient un tableau de 150 lignes (1 colonne) correspondant à l'espèce de chaque iris contenu dans `data` ;

- `target_names` (numpy.ndarray de str) contient un tableau de 3 lignes contenant les noms des espèces d'iris (setosa, versicolor, virginica).

2.1 Observation des données

Une des premières étapes peut être de visualiser les données et d'avoir un aperçu des éventuelles corrélations entre les données. Pour cela, il est intéressant de visualiser sur des graphes comment sont réparties les espèces en fonction des différents critères. Pour cela, il serait possible de réaliser une « matrice de graphiques » avec matplotlib. Cependant les bibliothèques panda (manipulation de données) et seaborn (visualisation de données) permettent de gérer plus efficacement les set de données.

```
import seaborn as sns
import pandas as pd
sns.set()
df = pd.DataFrame(data, columns=iris['feature_names'] )
df['target'] = target
df['label'] = df.apply(lambda x: iris['target_names'][int(x.target)], axis=1)
df.head()
```