TP de Python avancé

Jean-Philippe Attal (Campus de Cergy)
Ecole internationale des sciences du traitement de l'information

13 novembre 2016

Le TP est composé de 5 exercices permettant d'utiliser les API et des libraires spécialisées en Python.

Exercice 1 : Une étude sur les dauphins

A partir du jeu de données que l'on vous aura fourni, il vous est demandé de réaliser en python les choses suivantes (On attend comme résultat la figure 1) :

Question 1
Ecrivez un script permettant la visualisation du graphe, vous pourrez utiliser la librairie *Mathplotlib*.

Question 2
Implémentez la méthode spectrale en utilisant la méthode *linalg* de *numpy*.

Question 3
Implémentez un script permettant la visualisation des éléments du vecteur propre associé à la seconde plus petite valeur propre.

Question 4
Implémentez un script permettant la visualisation des éléments des vecteurs propres associés aux trois plus petits éléments propres (V_2,V_3 et V_4).

Question 5
Implémentez un script permettant la visualisation des deux communautés en utilisant une bissection.

Question 6
Reprendre la question 5 en utilisant un K-means.

Question 7
En utilisant *igraph* ou *networkX*, refaire les question 1,2 et 3 en utilisant la propagation de labels. En observant les résultats de la propagation de labels, qu'en déduisez-vous ? Quels sont les différences avec la méthode spectrale ?

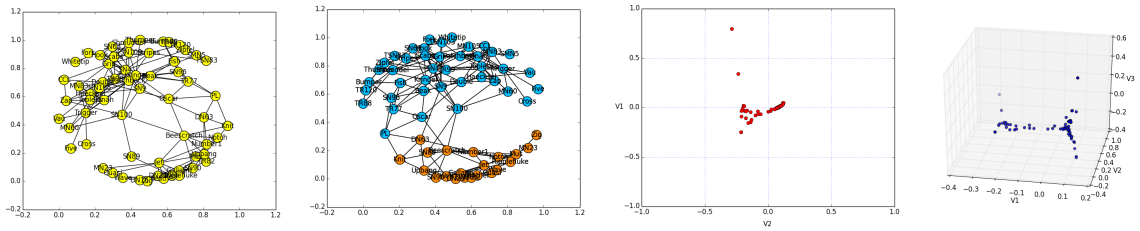

FIGURE 1 – Résultats attendus

Exercice 2 : API, métro de Washington

La mairie de Washington a mis en place une API pour mettre en temps réel toutes les informations sur son réseau de métro. Ainsi, si une perturbation à lieu, l'API retourne la durée (une estimation) et le lieu.

Question 1
Il vous est demandé de vous enregistrer sur le site *http ://developer.wmata.com/*.

Question 2
Ecrire un script permettant de calculer le plus court chemin entre deux stations tout en évitant d'éventuelles perturbations.

Exercice 3 : Clustering sur les IRIS

Il vous est demandé de télécharger le jeu de données concernant les Iris. Dans la suite de l'exercice, vous utiliserez le package scikit learn.

Question 1
En utilisant l'algorithme *Fuzzy C Means*, trouvez les clusters pour les Iris.

Question 2
En utilisant l'algorithme *DBSCAN*, trouvez les clusters pour les Iris.

Question 3
Créez une grille de similarité, figure 2, basée sur l'information mutuelle normalisée pour comparer DBSCAN et Fuzzy C Means. Le code devra être générique et ne prendre que des fichiers de résultats en entrée.

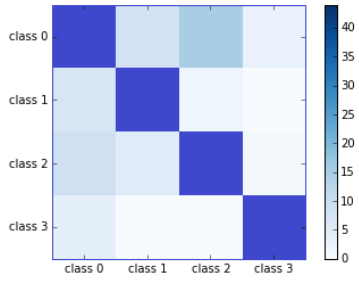

FIGURE 2 – Résultat attendu

Question 4
Il vous est demandé de réaliser la figure ci-dessous, figure 3. Vous pourrez faire de même les sépales.

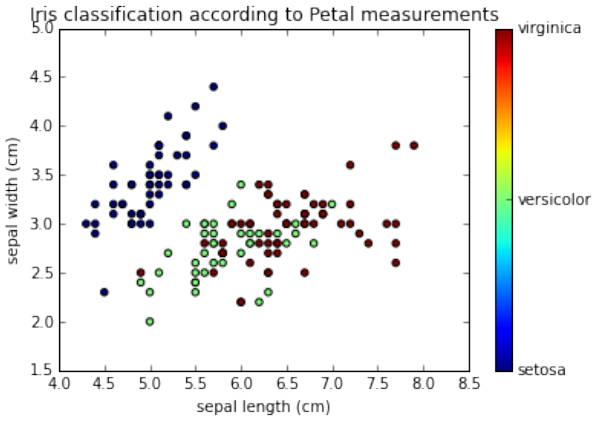

FIGURE 3 – Résultat attendu

Exercice 4 : Réseaux de neurones

En allant sur l'adresse suivante $http://scikit-learn.org/stable/auto\_examples/classification/plot\_digits\_classification.html$, vous trouverez un modèle permettant la reconnaissance de forme, notamment sur les digits.

Question 1
Il vous est demandé d'écrire un script permettant, pour une image montrant un nombre issu d'internet, de retourner sa valeur. Vous utiliserez pour cela un réseau de neurones disponible sur scikit learn avec comme fonction logistique :

- la tangente hyperbolique
- la Sigmoïdale logistique
- le Softmax
- la Gaussienne
- l'inverse de l'identité

Question 2
Effectuez une grille de score sur des exemples que vous aurez vous même définis, avec comme sortie un booléen.

Exercice 4 : Recommandation

On considère un ensemble de pages web. On vous demande d'utiliser la librairie *Scikit Learn.*
Question 1
Il vous est demandé de concevoir un modèle de recommandation tel que pour chaque document, les k plus proches seront proposés à l'utilisateur. Vous implémenterez le TD-IDF, le TF, la pondération des termes, et

enfin la similarité entre les documents.

Question 2
Un utilisateur propose un vecteur de mots. Créer un système permettant de retourner les documents les plus similaires et pertinents pour cette requête.

Exercice 5 : Visualisation en python

Il vous est demandé d'aller sur le site suivant *https* : *//www.kaggle.com/c/titanic/data* et de télécharger les données.

Question 1
Vérifiez s'il y a des individus ayant leurs modalités à NaN. Quelle variable a le plus de modalités à NaN ?

Question 2
En utilisant la librairie *cPickle*, créez un tableau qui ne contient que les variables *Survived*, *Pclass*, *Sex*, *Age* et *Embarked*. Ce fichier pourra être chargé à n'importe quel moment (*serialisation*).

Question 3
Ecrire une fonction permettant de tracer l'histogramme de la répartition de l'âge des passagers.

Question 4
Discrétisez la variable *age* par pas de 10 ans. Utiliser la fonction précédente pour tracer le nouvel histogramme.

Question 5
Nous étudions la variable *Survived* et sa corrélation avec les autres variables.

a) Calculez la proportion de passagers ayant survécu.
b) Calculez le nombre et la proportion de femmes et puis d'hommes qui ont survécu.
c) Représentez ces valeurs par un diagramme en bâtons. Obtenez les figures 5 et 6.
d) Pensez-vous d'après les observations, que les variables *Survived* et *Sex* puissent être indépendantes ?
e) Ecrire une fonction générique prenant en argument une variable "*variable*" du tableau de données anciennement créé. Cette fonction devra tracer le diagramme en bâtons représentant le nombre de survivants/morts par groupe donné par la variable "*variable*" (Ainsi, si "*variable*" est la variable *Sex*, on obtient le même diagramme qu'à la question précédente). Vous pouvez également obtenir la figure 7.
f) Appliquez cette fonction aux variables *Pclass*, *Tranche d'age* et *Embarked* et interpréter les graphiques.
g) Calculez les statisiques du χ^2 et la *p-value* correspondante pour tester l'indépendance des variables *Survived* et *Sex*. Que pouvez-vous dire des résultats ?

Question 6
Ecrire une fontion qui prend en argument une variable *variable* du tableau de données et calcule la statistique du χ^2 et la *p-value* correspondante pour tester l'indépendance des variables *Survived* et "*variable*".

Question 7
Appliquez cette fonction aux variables *Pclass*, *Tranche d'age* et *Embarked* et interpréter les résultats.

Question 8
Faites une représentation en *square binning* de l'âge en fonction de la classe. Vous devez obtenir le résultat de la figure 4. La couleur de chaque carré représente le nombre d'individus. Effectuez une interprétation.

Question 9

Reprenez toutes les questions précédentes en utilisant la notion de *dataframe* en utilisant la librairie *pandas*. Qu'appelle-t-on dataframe et quel est l'avantage de son utilisation ?

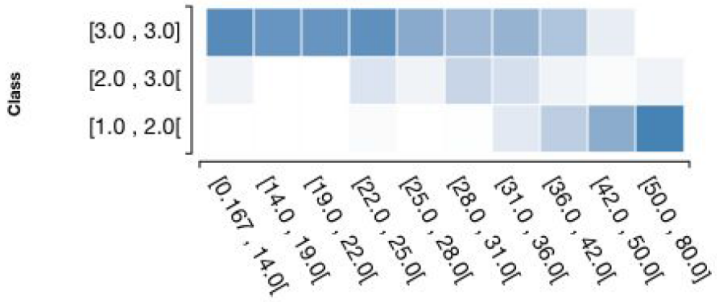

FIGURE 4 – Résultat attendu

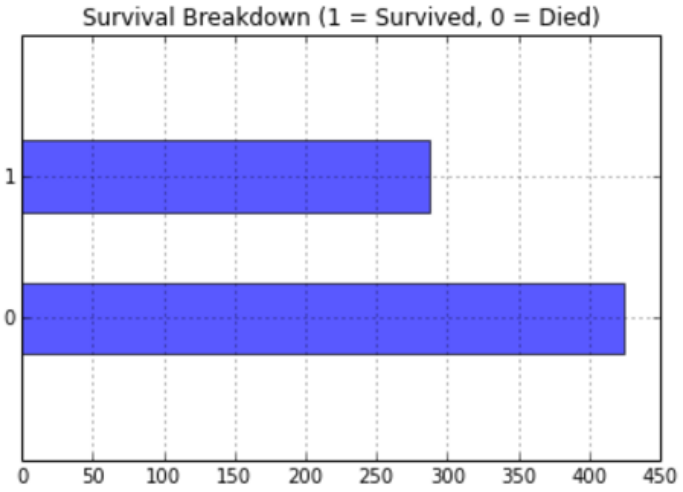

FIGURE 5 – Résultat attendu

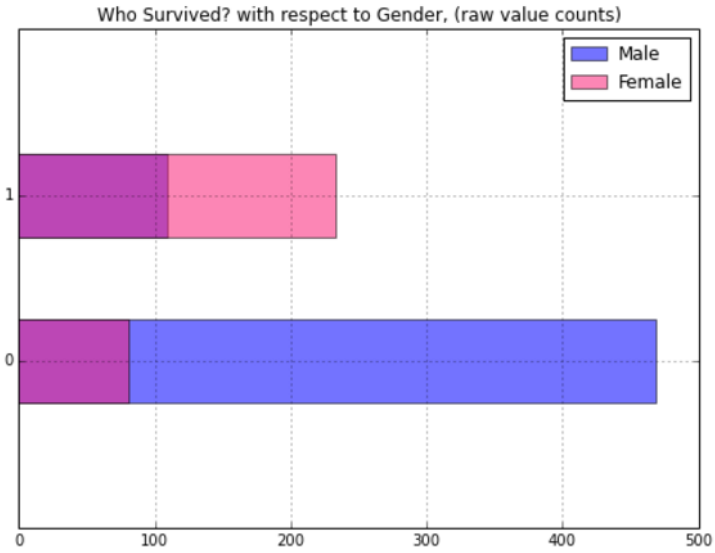

FIGURE 6 – Résultat attendu

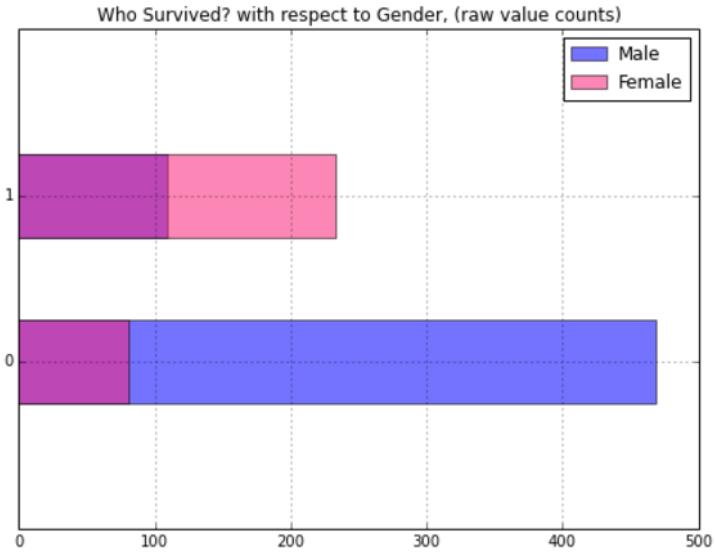

FIGURE 7 – Résultat attendu