Tutoriels de référence :

http://tutoriels-data-mining.blogspot.fr/2015/08/python-statistiques-avec-scipy.html

http://tutoriels-data-mining.blogspot.fr/2015/08/python-les-matrices-avec-numpy.html

Vous serez également à réaliser des graphiques à l'aide « Matplotlib » :

http://www.labri.fr/perso/nrougier/teaching/matplotlib/

Préambule :

- Sauf mention explicite contraire, tous les tests sont à 5% durant cet exercice.
- Concernant les indications fournies pour les exercices ci-dessous, nous utilisons les alias suivants :
    - import numpy as np
    - import scipy.stats as stat
    - import scipy.cluster as cluster

Questions :

On souhaite traiter le fichier « **Iris.txt** ». Il décrit 150 fleurs à l'aide de 4 caractéristiques physiques (longueur et largeur des sépales et des pétales). Elles sont réparties en 3 espèces d'iris spécifiées à l'aide de la 5ème variable « species ». Visualisez le dans un éditeur de texte (ex. NotePad) pour identifier sa structure (organisation des colonnes, les deux premières ont été mises en commentaires).

Importation des données

1. Charger les données dans une matrice NumPy (np.loadtxt). Affichez le nombre de lignes et de colonnes (shape).
2. Scindez les données en deux parties : **X** correspond à la matrice des descriptions (colonnes 0 à 3), **y** au vecteur des espèces (colonne 4).

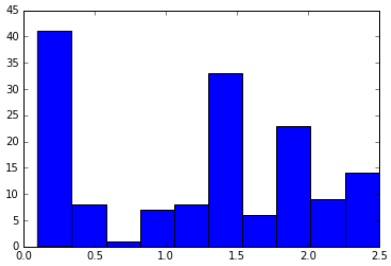
Description statistique

3. Affichez les statistiques descriptives sur X (stat.describe).
4. Affichez exclusivement les moyennes des variables. Deux solutions possibles, soit passer par les fonctions de la librairie Numpy (np.mean), soit exploiter la sortie de describe() ci-dessus.
5. Affichez les 1er et 3ème quartiles des variables de X (stat.scoreatpercentile).
6. Calculez les intervalles interquartiles pour chacune des variables.

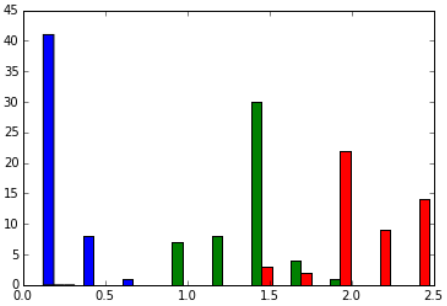
Adéquation à la loi normale

7. Tester la normalité de chacune des variables de X en utilisant le test d'Agostino (https://en.wikipedia.org/wiki/D'Agostino's_K-squared_test) (stat.normaltest). Conclusion ?

8. Affichez l'histogramme de fréquence de « x3 », la 4ème variable de X. Utilisez les commandes graphiques de « matplotlib.pyplot » (hist). La distribution pouvait être compatible avec la loi normale ?



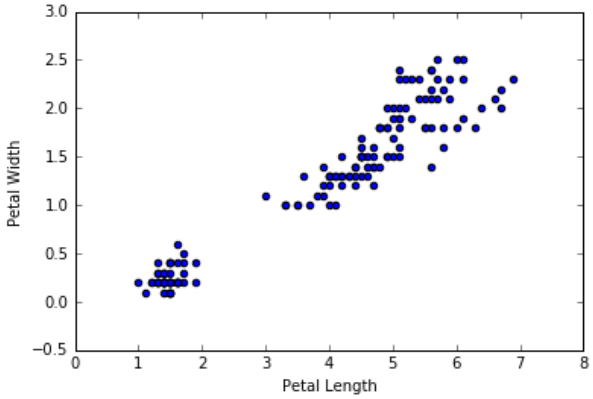
9. Refaites le même graphique mais après avoir scindé les individus selon leur groupe (espèce). Que constatez-vous alors ? (une piste simple consisterait à scinder x3 en 3 vecteurs x31, x32 et x33 pour chaque espèce).



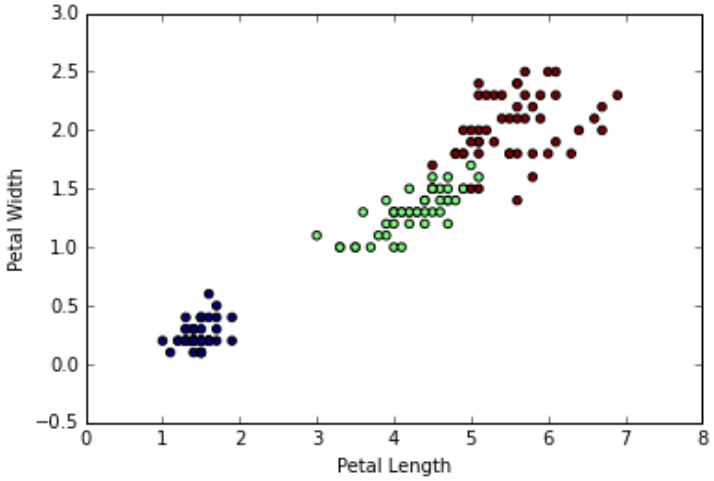
10. Au regard de ces éléments, pouvez-vous refaire les tests de normalité d'Agostino pour x3 dans les 3 sous-populations définies par les groupes (3 tests à faire donc) ? Que nous disent les résultats ?
11. Comptez le nombre d'observations pour chaque modalité de y (np.unique avec l'option return_counts).

Graphiques

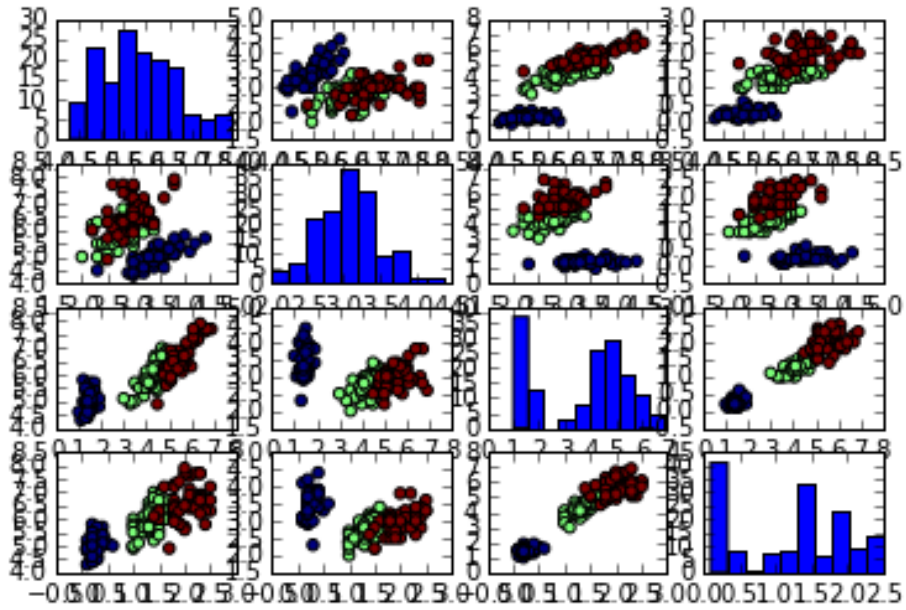
12. Créer un graphique nuage de points où vous croiserez les 3ème et 4ème variables de X (scatter). Que constatez-vous ? (cf. http ://www.labri.fr/perso/nrougier/teaching/matplotlib/)



13. Refaites le même graphique mais en coloriant les points selon leur groupe d'appartenance (espèce) (cf. les options de scatter dans l'aide)



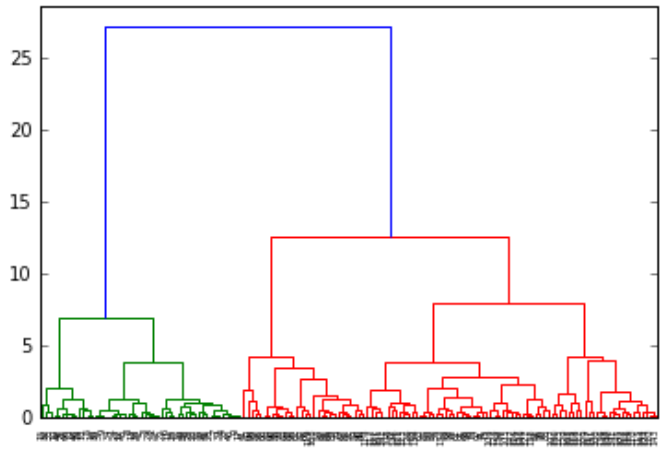
14. On souhaite avoir le même type de graphique mais en croisant les variables deux à deux. Comment pourrait-on faire cela ? (cf. http ://matplotlib.org/users/gridspec.html). Voici un exemple du résultat souhaité (cf. GridSpec et subplot)



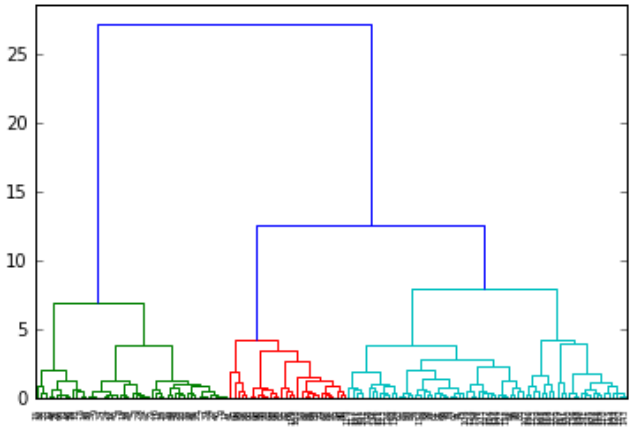
(*on doit pouvoir réduire la taille du texte sur les axes mais bon…*)

Classification automatique (clustering) - CAH

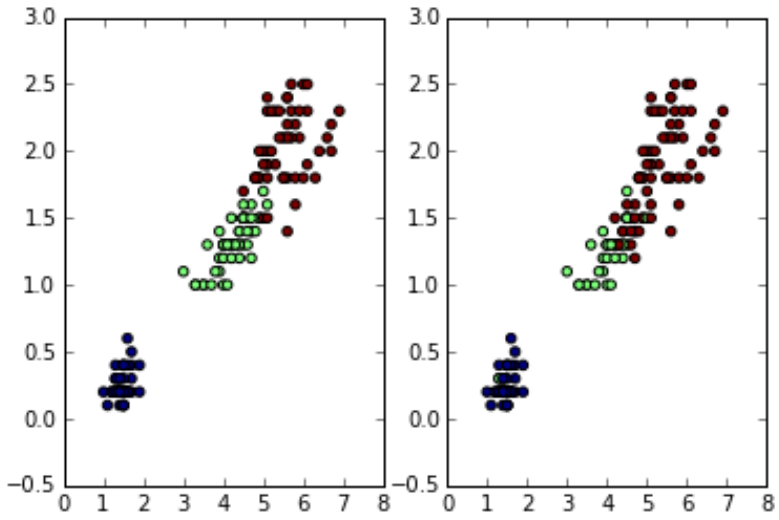
15. On souhaite réaliser une classification automatique sur X. Construire la matrice Z qui correspond aux valeurs centrées et réduites X (stat.zscore). Vérifiez que les moyennes par variables sont bien égales à 0, et les écarts-type à 1.
16. Lancez la CAH (classification ascendante hiérarchique) (cluster.hierarchy.ward), utilisez la méthode de Ward. Affichez le dendrogramme (cluster.hierarchy.dendrogram) (voir aussi : https://docs.scipy.org/doc/scipy/reference/cluster.html). Vous proposeriez un découpage en combien de classes ?



17. Mettons qu'on part sur un découpage en 3 classes. Affichez de nouveau le dendrogramme en faisant apparaître explicitement les 3 groupes (voir les options de cluster.hierarchy.dendrogram**).**



18. Appliquez le découpage en 3 classes afin d'obtenir afin d'obtenir le groupe d'appartenance de chaque individu (cluster.hierarchy.fcluster).
19. Construisez deux graphiques nuages de points croisant les deux dernières variables de X mettant en évidence d'une part les groupes originels définis par y (à gauche), et d'autre part les groupes affectés par la classification automatique (à droite). Que peut-on dire ? (On devrait obtenir un graphique comme ceci)



20. Construisez un tableau croisant les classes d'appartenances observées (species) avec celles issues du clustering. Que constatez-vous ? Est-ce cohérent avec ce que vous observez dans le graphique ci-dessus ? (une piste simple consiste à intégrer les deux vecteurs dans une structure pandas.DataFrame [ex. https://stackoverflow.com/questions/41873198/pandas-create-a-dataframe-from-2d-numpy-arrays-preserving-their-sequential-orde ; la réponse du 06 novembre 2017] ; puis exploiter la commande CrossTab de pandas [ex. http://tutoriels-data-mining.blogspot.fr/2017/02/python-manipulations-des-donnees-avec.html ; instruction n°40]).

21. Calculez le carré du rapport de corrélations pour chaque variable (les 4 variables) en fonction de l'appartenance aux groupes induite par la dernière classification réalisée (https://fr.wikipedia.org/wiki/Rapport_de_corrélation). Quelle est la variable qui influe le plus dans la constitution des groupes ?