TP pandas

Sergio Peignier

But du TP :
— Vous faire découvrir les avantages potentiels de Pandas.
— Ce TP a vocation n'a pas vocation à présenter Pandas de manière exhaustive.

1 Installation

Je vous conseille d'installer miniconda (ou anaconda) en local afin de pouvoir installer et gérer vos librairies python plus facilement. Pour miniconda aller sur `http://conda.pydata.org/miniconda.html` et télécharger la version qui correspond à votre OS/machine. Pour anaconda aller sur `https://docs.continuum.io/anaconda/install` (ne pas installer anaconda sur les machines de BIM, anaconda bouffe trop de mémoire). Suivre les instructions sur `http://conda.pydata.org/docs/install/quick.html`. Si vous installez Miniconda pour python 2.7, sur un ordinateur 64bit, avec Linux comme OS, vous pouvez taper les instructions du fichier envoyé en pièce-jointe.

2 Définition et structures de base

Lire `http://pandas.pydata.org`. Quelles sont principales caractéristiques de la librairie ?

La librairie repose sur plusieurs structures de données assez intéressantes, en particulier les 'DataFrames' et les 'Series'. Le fonctionnement de ces structures rappelle en quelque sorte celui des 'ndarrays' de Numpy et les dictionnaires. Lire `http://pandas.pydata.org/pandas-docs/version/0.15.2/dsintro.html#dsintro` pour avoir une idée de ce que ces structure de données permettent de faire. Regarder en particulier comment on peut créer des 'Series' et des 'DataFrames', comment on peut accéder aux données que ces structures contiennent. Pour ce faire suivre le tutoriel linéairement jusqu'à la sous-section 'Console display' de la section 'DataFrame'.

3 Plus d'opérations

Lire`http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html` depuis 'Viewing Data' jusqu'à 'Grouping' (inclus). Copier le code de certains exemples (ceux qui vous semblent difficiles) et le modifier afin de mieux le comprendre.

4 Application

— Télécharger le jeux de données 'iris.data' dans le répertoire 'Data Folder' `https://archive.ics.uci.edu/ml/datasets/Iris`.
— Que représentent ces données ?
— Charger le jeu avec pandas. Peut-on directement charger le jeu sans le télécharger depuis son URL ?
— Standardiser le jeux de données (z-score) en une ligne
— Calculer la moyenne et l'écart-type par espèce de chaque descripteur (sepal length, sepal width ...).
— Construire un échantillon de taille S, de fleurs aléatoirement choisies. Calculer la moyenne de sepal length par espèce. Répéter l'opération 30 fois, pour $S = [10, 30, 50, 70]$ afin de remplire un DataFrame qui aura 3 colonnes (une par espèce), et une quatrième colonne qui contiendra la valeur de S (taille de l'échantillon). Chaque ligne correspond à chacune des mesures. Calculer la variance des moyennes de sepal length pour chaque taille d'échantillon.
— Représenter les fleurs projetées dans l'espace 2D définit par sepal length et sepal width (voir matplotlib.pyplot pour les graphiques), chaque espèce aura une couleur différente.