Datamining : TP5

M2 Informatique, Université Paris 7
Anne-Claire Haury

Exercice 1

1. Dans un terminal, ouvrir ipython et lancer le code suivant, ligne par ligne :

```
from sklearn.datasets import load_digits
import pylab as plt

digits = load_digits()

X      = digits['data']
images = digits['images']
labels = digits['target']

plt.imshow(images[0], cmap = 'gray')
plt.show()
print labels[0]


plt.imshow(images[100], cmap = 'gray')
plt.show()
print labels[100]

print 'Taille de la matrice : ', X.shape
```

   Il s'agit des données de chiffres manuscrits. L'objectif de cet exercice est de les clusteriser.

2. Coder l'algorithme des KMeans sous forme de classe.
3. Appliquer cet algorithme à X avec différents nombres de clusters.
4. Comment évaluer cet algorithme à l'aide des vrais labels ? Coder une fonction qui renvoie une valeur entre 0 et 1 (1 si l'algorithme est parfait, 0 s'il a tout faux)

   Pour eviter toute confusion a venir, l'algorithme des K Means est bien non-supervise, c'est a dire qu'il n'est pas base sur les vrais labels. Ici, nous essayons seulement de l'evaluer puisque nous connaissons ces valeurs.

Exercice 2

- Regroupez-vous par groupe de projets
- Créer un document qui se nomme `projet_v0_<nom de votre groupe>.txt`.
- Dans ce document, répondez en groupe aux questions suivantes :
  1. Quelle forme pourrait prendre une démo de votre projet ? (Entrées, sorties, interface, etc.)
  2. Où et comment allez-vous récupérer vos données ?
  3. De combien de données avez-vous besoin au départ ?
  4. Sous quelle forme allez-vous les stocker ?