COEN 140 Machine Learning and Data Mining

Lab Assignment #5: K-Means Clustering

Guideline: Submit a pdf report to Camino. Also submit all the source code needed to generate the results as a separate zip file to Camino.

Problem The Iris dataset contains 150 data samples of three Iris categories, labeled by outcome values 0, 1, and 2. Each data sample has four attributes: sepal length, sepal width, petal length, and petal width.

Here is the code snippet to load the dataset. Note: since K-means clustering is unsupervised learning, we don't need to split the data into a training set and a test set.

```
from sklearn import datasets
iris = datasets.load_iris()
print(list(iris.keys()))
print(iris.feature_names)

X = iris.data # each row is a sample
y=iris.target # target labels
```

Implement the K-means clustering algorithm to group the samples into K=3 clusters. Initialize the cluster centers by the first 3 data samples. The objective function to minimize is defined as: $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \|\mathbf{m}_k - \mathbf{x}_n\|_2^2$. Each iteration includes an assignment step and a cluster-center update step.

Calculate the objective function value J **after the assignment step in each iteration**. Exit the iterations if the following criterion is met: $J(\text{Iter} - 1) - J(\text{Iter}) < \varepsilon$, where $\varepsilon = 10^{-5}$, and Iter is the iteration number. Plot the objective function value J versus the iteration number Iter. Comment on the result.

Demo and explain to TA (10%):

Show in a certain iteration

1. How you execute the “assignment” step
2. How you update the cluster centers
3. How you calculate the objective function value
4. How you check the stopping criterion

Grading:
Demo: 10%
Report: 60%
Source Code: 30%