CSCI 381/780 Machine Learning
Assignment #1
Clustering
Due before class, Wednesday, 5 October 2016

Introduction

This assignment asks you to implement the K-means clustering algorithm described in class and test on the "seeds" dataset.

The algorithm

The K-means algorithm is a method for clustering a set of data points into several similar groups. Given an input K value, the algorithm starts by randomly selecting K centroids. Then it iterates between **assigning cluster memberships** and **recalculating centroids**. A data point is assigned as a member to the closest cluster. Use the Euclidean distance in this assignment. New centroids are calculated by taking average values of each feature. The two steps are repeated until there is no more change in membership or centroids.

The Seeds dataset

Measurements of geometrical properties of kernels belonging to three different varieties of wheat. A soft X-ray technique and GRAINS package were used to construct all seven, real-valued attributes.

Attribute Information:
1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

All of these parameters were real-valued continuous.

Evaluation and analysis

Perform the following analysis with your implemented algorithm:

1. Set K=3.
2. Randomly generate 3 starting centroids.
3. Test your algorithm on the centroids until convergence.
4. Calculate the Intercluster Variability (IV) and Extracluster Variability (EV) of the final clusters.
5. Repeat steps 2-4 for 5 times.
6. Also try steps 2-4 on a set of 3 starting centroids that you **manually** pick by examining/visualizing the dataset in whatever way you desire.
7. Write a report with **at least** the following elements
    - Explain how you randomly generate the starting centroids in step 2

- Explain how you manually pick the starting centroids in step 6
- Describe any details of your implementation that may be important
- Report all sets of starting centroids, final centroids, IV, EV, and IV/EV in a table.
- Discuss your results! Which set of starting centroids generates the best results? Any observation why this set of centroids is good?
- Report any other observations you may have on your experimental results.

Deliverables

Submit a single .zip file named LastName.FirstInitial.HW# containing the following to the instructor (changhe.yuan@qc.cuny.edu). In addition, submit a hardcopy of your reports in class on the due date.

1) Well commented code; any programming language is allowed.
2) A readme file explaining how to compile and run your program;
3) A report explaining the K-means algorithm briefly, documenting any interesting issues you encountered when implementing the method, and reporting your experimental results.