CS 484 Data Mining

Fall 2015
Professor Jessica Lin

HW 4 – Due 11/17

In this assignment you will implement the k-means algorithm. You can use any programming language, provided that the program can be run on zeus. Provide a README file with instructions on how to run your program.

Input: filename, k

The input file should contain *n* rows and *m+1* columns, where *n* is the number of records, and *m* is the number of attributes in the dataset. The last column is the class attribute (note that this is for evaluation purpose only. The k-means algorithm itself does not require a class attribute). For simplicity, you may assume that the first *m* columns contain numeric data.

Output: ~~an *n-by-k* matrix showing cluster membership. Each matrix cell contains a number from 1 to k.~~ an *n-by-k* matrix showing cluster membership (each matrix cell contains either 0 or 1). Since I messed up here, you can also have a n-by-1 matrix, where each matrix cell contains a number from 1 to k

Step 1: Your implementation should initialize cluster centroids at random. Your implementation should iterate until the algorithm converges (i.e. no cluster assignments change from one iteration to the next). Submit the documented source code of your implementation.

Step 2: Cluster the Iris dataset without the class attribute (https://archive.ics.uci.edu/ml/datasets/Iris). Run your algorithm and compute the accuracy of your clustering. Unlike classification, the cluster IDs of your k-means clusters will in general not correspond with the class labels. You must first compute the optimal mapping of cluster IDs to class labels (for example, associate class 1 with cluster 3, class 2 with cluster 1, and class 3 with cluster 2). You should also compute the confusion matrix for your clustering results. Submit results for one sample run giving the initial clustering, final clustering, and confusion matrix.

Step 3: Run your algorithm 50 times over the Iris dataset, each time using a different initialization. For each run, compute both the accuracy of clustering, as well as the k-means objective function value of the final clustering. Plot the distribution of these accuracies and also plot the distribution of the objective function values. Finally, give a scatter plot showing the correlation (if any) between the clustering accuracy and the clustering objective function value. Write a paragraph to interpret these graphs - what can you conclude?

Submission: Please submit your source code, README and report electronically on Blackboard, and bring a hard copy of the report to class.