CX 4010 / CSE 6010
Assignment 3
Clustering

Due Date: 11:59pm on Thursday, September 10
Submit a single zipfile as described herein to Canvas

In this assignment, you will develop a program to cluster data samples from a data set into k distinct clusters (categories) using the K-means algorithm. K-means is a well-known approach used in machine learning to cluster data. The input to the algorithm is a set of data samples, where each sample is a point in n-dimensional space. In other words, each sample is a tuple of n data values $(x_1, x_2, \ldots, x_n)$, where each dimension represents some attribute of interest (time, temperature, location, etc.). The goal is to partition the data samples into k distinct "clusters" where the samples within each cluster are "similar." For example, one of the earliest applications of clustering arose in the 19th century where the locations of deaths from a cholera outbreak were clustered and used to identify the cause of the outbreak. In this case, deaths were found to be clustered geographically around contaminated water wells.

The name "K-means" is commonly used to refer to both the clustering problem and a specific algorithm used to solve it. The value of k is an input given to the algorithm. K-means is one algorithm in a class of machine-learning techniques known as *unsupervised learning*. Given a set of data points (samples) in an n-dimensional space, the *centroid* μ is the point defined by the arithmetic average of the data points along each dimension. Here, our interest is the set of centroid points, where each centroid is defined by the points within one of the k clusters $(\mu_1, \mu_2, \ldots\ \mu_k)$. We would like the data points within the cluster to be "close" to each other, so a natural metric for a cluster is to consider the distance of each data point within the cluster from the cluster's centroid. The K-means problem is given a set of n data points $(x_1, x_2, \ldots\ x_n)$ and the value k, find $S = (S_1, S_2, \ldots\ S_k)$ that is a partition of the data points into k sets the minimizes the sum of squares distance:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \sqrt{\|x - \mu_i\|^2}$$

The algorithm for solving this problem is commonly called the "K-means algorithm" or Lloyd's algorithm. The algorithm repeatedly performs the following steps:

1. Assign each data sample to the cluster with minimum distance between the data point and that cluster's centroid.
2. After all samples have been assigned to clusters, update the cluster's centroid to the average of the points assigned to that cluster.

For this assignment, assume the first k data points in the data set are the initial centroids. Your program should repeat the above two steps until the root-mean-square falls below some predefined threshold (defined by you) or until the algorithm completes some maximum number of iterations if it is not able to reach this threshold.

For this assignment, write a C program that takes two command-line parameters:

1. The name of the input file containing the data to be clustered
2. The value of *k*

Assume the input data file for the K-means program has the following format:

```
num_items num_attrs

item1_attr1 item1_attr2 ... item1_attrm

item2_attr1 item2_attr2 ... item2_attrm

...

itemn_attr1 itemn_attr2 ... itemn_attrm
```

The first line contains the number of data items `num_items` and the number of attributes `num_attrs`. The following `num_items` lines contain the data items, with each line containing `num_attrs` attribute values.

You are provided with some code to help with reading in command-line parameters and setting up an array to store the input data based on the parameters read in the first line. This code will help you to refer to the array without the need for poinster references. If you are using an array to store the data values, C indexing for an array with `nRows` rows and `nCols` columns works such that `a[i][j]` is equivalent to `*( a + i*nCols + j)`. Note that if you intend to pass the array to a function, you will have to use pointer syntax within that function to access the array elements.

~~In addition to creating the output file,~~ For debugging purposes your program should print to the screen evidence that it is producing correct results. For example, you might print after each iteration the coordinates of the centers of each of the *k* clusters, the number of data items in each cluster, and the maximum and minimum root-mean-square of the distances of the data points to the cluster center across the clusters. Although it is not required, for testing purposes, you may wish to define a file format to hold the results produced by the K-means program and output these results. You should construct some synthetic data sets to verify that your program is working correctly.

Once you are confident your program is working, you should run it using the provided data sets: WineData_2col.txt and WineData_3col.txt. The data, taken from the reference at the end of this assignment, represent attributes of red variants of the Portuguese "Vinho Verde" wine. In particular, WineData_2col.txt includes residual sugar (g/dm^3) and total sulfur dioxide (mg/dm^3), while WineData_3col.txt includes residual sugar (g/dm^3), total sulfur dioxide (mg/dm^3), and alcohol (% vol.). Although you should not need it to complete this assignment, more information about the dataset can be found at https://archive.ics.uci.edu/ml/datasets/wine+quality.

Additional Requirement for Graduate Students (CSE6010) Only:

The datasets provided give values of the attributes (e.g., residual sugar) as raw measurements that include units. However, it is common to normalize the data as a z score (or standard score), which indicates how far from the mean a data point is in terms of the standard deviation. Perform the classification task both with the data as given and after transforming the data to z-scores. For each attribute j, you should calculate the z-score $z_{i,j}$ of each data item $x_{i,j}$ by subtracting the mean $\bar{x}_j$ for that attribute and dividing by the standard deviation σ_j for that attribute:

$$z_{i,j} = \frac{\left(x_{i,j} - \bar{x}_j\right)}{\sigma_j}.$$

You will need to calculate the mean and standard deviation for each attribute; the standard deviation can be calculated as

$$\sigma_j = \sqrt{\frac{\left(x_{i,j} - \bar{x}_j\right)^2}{n}},$$

where n is the number of items.

You will need to provide an additional slide discussing your output as related to this normalization step; see below.

Optional Extra Credit (for all students):

Earlier, you were asked to use the first k data points in the data set as the initial centroids. However, there are other ways to set the initial centroids. One option is to select the centroids randomly from the given data points.

For this extra credit, implement a random selection of centroids from the data points; you will need to learn about how to generate pseudo-random numbers in C. Make sure you do not select the same data point twice regardless of the value of k. You should study the behavior of the program for different choices of centroids and include an additional slide to discuss these points; see below.

Submission information

You should submit to Canvas a single zipfile that is named according to your Georgia Tech login—the part that precedes @gatech.edu in your GT email address. To receive full credit, your code must be well structured and documented so that it is easy to understand. Be sure to include comments that explain your code statements and structure.

The zipfile should include the following files:

(1) your code (all .c and .h files). If you are using linux or Mac OS, we recommend you use a makefile to compile and run your program, and you should include it if so. A sample makefile is provided.

(2) a README text file (not formatted in a word processor, for example) that includes the compiler and operating system you used for compiling and running your code along with instructions on how to compile and run your program.

(3) a series of slides composed in PowerPoint or similar software, saved either in PowerPoint or as a PDF and named slides.pptx or slides.pdf, in the following order:

- 1 slide: your name and a brief explanation of how you developed/structured your program. This should not be a recitation of material included in this assignment document but should focus on the main structural and functional elements of your program (e.g., the purpose of any loops you used, the purpose of any if statements you used to change the flow of the program, the purpose of any functions you created, etc.). In other words, under the assumption that the mathematics behind what you are doing is already known, what were the main things you did to translate those requirements into code?
- 1 slide: a description of your testing procedure and evidence of correct operation (e.g., test using different values of k, including invalid values, etc.). Explain why you think your program is correct from these tests.
- Grad students only: 1 slide: a brief description of why normalization may be desirable and what effect, if any, it has on K-means output for these datasets.
- Optional extra credit: 1 slide: a brief discussion of the reasons for and effects of using random data set members as the initial centroids compared to using the first k data set members.
- 1 slide: a summary and interpretation of your results. What value of k did you find most useful for the two datasets (could be different values) and why? Do you think the algorithm produced meaningful clusters for these two datasets? Why or why not?

Reference:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.