Programming Project 2

October 22, 2012

Due date: Mon, November 5 2012 at 4pm.

Description: *Summary:* You are requested to provide a clustering of the points in the given dataset.

In this project, you are given a set of 100-dimensional points and you are asked to discover the hidden structure (if any) behind them. In order to do that, you will apply clustering algorithms like the ones we saw in class, and provide a label for each point. That label will be the coordinates of the representative of the cluster to which the point belongs.

Evaluation: The evaluation function that will be used is the Root Mean Square Error as defined below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{1}$$

Here, n is the number of points, y_i is the representative you suggested for point i and $\hat{y}_i$ is the centroid of the cluster to which that point actually belongs. You can see how well your solution scores if you go to the Leaderboard.

We have split the data in two parts. The ranking you see in the leaderboard reflects your performance in 70% of the points. The other 30% gets evaluated in a second leaderboard which is hidden to you. This will ensure that you won't overfit on the results of the leaderboard.

Data: The file test.csv is a Comma Separated Values file. Each line corresponds to a point in a 100-dimensional space and the values of that point in each dimension are separated with commas. This means that the value for the i-th dimension of point j is the i-th number at line j.

What you are asked to do is to suggest a clustering over these points. This way, every point in the input file will be matched to a single cluster and, consequently, can be labeled with the centroid of the cluster it belongs to. You will need to create a new file, which, on the i-th line will have the coordinates for the centroid of the cluster that you believe point i belongs to. An example of such a file is Solution-barycenter.csv

We also provide you with a baseline solution, "Barycenter", to get a better feeling on how good you are performing. In this solution, all the data points have been matched to the barycenter of the dataset.

Writeup: In addition to submitting your solution online, you need to provide us with a 2-page writeup that describes the algorithm you have implemented and the special tricks you implemented in order to make it work. Also, describe your strategy for selecting a particular algorithm and how you did your offline evaluation of the method.

Instructions: As a first step, you need to signup to the platform. Go to `http://inclass.kaggle.com` and use your BU email. After that, search for *BU CS 565 : Clustering* and you will find 3 projects. Each of them has a different dataset with different difficulty level. **Take part and submit solutions for all of these 3 competitions.** Create a team and download the dataset from "Data Files". When you have a proposed solution, upload your *.csv* file at "Make a submission". We suggest that you don't just upload your final solution, but submit all your intermediate ones to check your standing. You can see how well your solution is ranking if you go to the "Leaderboard". Notice that the deadline on the competition page may not be accurate.

If you encounter any problems, send an email at *cmav@bu.edu*

Privacy and confidentiality: If you do not want to reveal your real name when you signup you can use your BUID or any other name as your alias. However, please send us an email with the alias you used. You should do that by Fri, Oct 26, 2012.