Assignment 4 - COP 4531 (FALL 2015)

Due Date - 10/29/2014 3:25 PM

Project Description

In this assignment you will implement (1) Lloyd's method for the k-means objective function and (2) the complete-linkage algorithm. For both algorithms, implement the versions covered in class (and described in the online notes). For Lloyd's method, run the algorithm 100 times with different initial random centers and select the clustering with lowest k-means cost. You will also implement hamming distance for finding the distance between two clusterings.

The output of each of these two algorithms should be cluster assignments as vectors of integers. For example, if you have two clusters and 10 points then print a vector of size 10 for the cluster assignments:

$$A = [0, 0, 0, 0, 1, 1, 1, 0, 1, 0].$$

Here the i^{th} element of A gives the cluster assignment for the i^{th} point.

Run these two algorithms on the three provided data sets. For each data set, report on the following:

1. The hamming distance between the clusterings found using Lloyd's method (the best one over 100 runs) and the clustering obtained by complete-linkage.

2. The running time of each algorithm (note that for Lloyd's method, it will be the combined running time over the 100 runs). The running time should be measured in actually time (such as 5 seconds and 10 milliseconds), not the asymptotic running time of the algorithms.

Programming languages and formatting

You are free to use any programming language and library available on linprog. However, you should implement the algorithms yourself and NOT use any library implementation of the algorithms. It must be possible to compile and run your code in linprog, and if it is not, it will be considered broken. As a reminder, please do not copy code from online or your classmates. You may use the pseudocode provided in class to design your solutions.

Test cases

The supplied test cases are:

- data1.txt - One dimensional data. Set the number of clusters to 2.
- data2.txt - Two dimensional data. Set the number of clusters to 2.
- data3.txt - Four dimensional data. Set the number of clusters to 3.
- We will also test on additional data sets after submission

The data looks as follows, with the first line giving the number of elements, and the second line indicating the number of clusters, followed by the data points.

```
4
2
(1,2)
(2,2)
(3,2)
(1,7)
```

Submission Guidelines

Your submission must have the following:

- A one page report documenting the hamming distance between the clusterings on each of the three provided data sets and the running times of the algorithms.
- A README file that describes how the code can be compiled and run, **particularly how we can call Lloyd's method and complete linkage on new data**. Also list any external dependencies that need to be satisfied for compiling and running the code.
- A Makefile that can be used for compilation. Please note that the TAs will not write Makefiles for compiling code or write any commands except for single word commands like **make** for compiling.
- Your submission should have a single point of entry for running the code, like a main() method in C or its equivalent in any other language.
- **Please e-mail a single zip or tar file of your submission named using your last name (*e.g.* kiswani.tar.gz) to ssk09c@my.fsu.edu by 3:25 PM on Wednesday October 29**.

Mark Breakdown

- Report comparing Lloyd's method and Complete linkage on the 3 provided data sets in terms of the hamming distance between the two clusterings *and* their running times. List the numeric results and summarize your findings. - [40 Points]
- Design of program - [10 Points]
- Error free compilation - [10 Points]
- Correctness on supplied data **and additional** tests [40 Points]

Bonus (+10)

Implement a visualization in 2D and 3D for displaying the clusters and the cluster assignments using different colors. For Lloyd's method also display the cluster centers.