Stats 141C High Performance Statistical Computing Spring 2018

Homework 3

Lecturer: Cho-Jui Hsieh Date Due: June 7, 11:59pm, 2018

Keywords: *Clustering, Sparse data*

For this homework, we will use the data downloaded from `http://www.stat.ucdavis.edu/~chohsieh/teaching/STA141C_Spring2018/hw3_data.zip`. In this folder, we provide two datasets: "data_dense.pl" (this is the same with the training data we used in homework 2) and "data_sparse_E2006.pl" (this is a sparse TF-IDF document data). Use the following line to load data:

```
X = pickle.load(open(''data_dense.pl'', ``rb''))
```

If you use python3, try the following line:

```
X = pickle.load(open(''data_dense.pl'', ``rb''), encoding='latin1')
```

Problem 1. K-means clustering [65 pt]

Implement the "k-means" algorithm to cluster the dense data ("data_dense.pl") into $K = 10$ clusters. The k-means algorithm can be found in lecture 11, page 15 or 16 (they are equivalent). Initialize cluster centers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_{10}$ using 10 randomly sampled data points. Print out the k-means objective

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_n \in C_k} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2$$

at each iteration, where C_k is the data points belong to k-th cluster, and $\boldsymbol{m}_k$ is the cluster center of the k-th cluster. Run the program for 40 iterations and report the objective function and running time at iteration 10, 20, 30, 40. Discuss your findings.

Problem 2. K-means for sparse data [35 pt]

Apply the same k-means algorithm to sparse data ("data_sparse_E2006.pl"). Note that in this pickle file X is stored in Compressed Sparse Row (CSR) format, and you will need to modify your code accordingly to use sparse matrix (turn the data into dense matrix will be out-of-memory). Run the program for 40 iterations and report the objective function and running time at iteration 10, 20, 30, 40. Discuss your findings.