CS328: Homework 1

Your submission should be a single Jupyter notebook containing all the answers and the code. Use markdown or LateX for the answers to the theoretical questions. You can, of course, work on Colab and submit the resulting notebook after downloading it.

Copying code is not allowed, from others or any sources. Discussion with others is okay, but everything, both code and answers, has be be developed individually. Also give names of all collaborators.

1. Suppose you define a clustering objective in the following manner – give a partitioning $\mathbb{C} = \{C_1, \dots C_k\}$, define

   $$cost(\mathbb{C}) = \sum_i \frac{1}{|C_i|} \sum_{x,y \in C_i} \|x - y\|_2^2$$

   i.e. cost of a cluster is the sum of all pairwise squared distances. Give an algorithm for this.

2. For the k-means problem, show that there is at most a factor of four ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers.

3. Create a random variable X for which Markov's inequality is tight. Give proof for your answer. If it is tight, then why are we using other inequalities e.g. Chebyshev and Chernoff?

4. Download the MNIST dataset from http://yann.lecun.com/exdb/mnist/. We will use the test dataset and test labels only.

   (a) Cluster them first using k-means clustering, $k = 10$, with $kmeans++$ initialization (implement the complete Lloyd's algorithm yourself). Check the Rand-index of the clustering against the true labels. Use the sklearn module for rand-index.

   (b) Do the same for k-center clustering, $k = 10$. Implement the greedy algorithm discussed in class. Report the Rand-index here too.

   (c) Run the single linkage agglomeration till there are $k = 10$ clusters. Report Rand-index here too.

   (d) Run the same algorithms (k-means and k-center) but on a rank-k approximation of the training data matrix. Note that if A is the training data matrix (images × pixels), then you can just use $U_k \Sigma_k$ for the clustering, no need to use V_k. Evaluate for $k = 2, 5, 10$ and report the rand-index values.

5. Suppose you have a population of 1 million people, out of which at least 1% are coffee drinkers. You want to get the estimate of this fraction by using sampling. Give the algorithm and the estimate. What kind of error bounds can you give with probability 99%?