CS328: Homework 1

While usage of Python/C/C++/Java etc are all fine, submitting a Jupyter notebook is preferable for each of the coding questions. 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 collaborators.

1. Is the following function $d(x, y) = \min_i |x_i - y_i|$ a metric? Either prove it or give counter-examples.

2. Suppose you define a clustering objecive in the following manner – give a partitioning $\mathbb{C} = \{C_1, \ldots C_k\}$, 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.

3. The k-median problem is defined in a similar way to the k-means problem, except that we do not take the squares of the distances when summing up. For the k-median problem, show that there is at most a factor of two ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers. Based on this (or otherwise) propose a variant of the Lloyd's algorithm for Euclidean k-median. Can you say that the clustering cost is always decreasing?

4. Download the dataset in https://www.kaggle.com/arjunbhasin2013/ccdata. As a good practice, normalize each feature such that the values are all in the range [0, 1]. Treat the CUST ID column as the identity of the point, not a feature. Use the L2 metric as distance. Implement the greedy k-center algorithm for this data and report the k-center objective value for k = 2, 4, 10. For small values of k, say k = 2, 4 , find the optimal (when the centers are restricted to be input points) and report the approximation factor obtained by the greedy algorithm.

5. For the following question you need to submit a link to a recorded video, YouTube link is preferable (can be unlisted). We intend to link to these videos from our public course webpage.

   Go through the video at https://www.youtube.com/watch?v=hVimVzgtD6w. There are number of libraries to create such visualization: one example is GapMinder animation, another is Plotly. Choose any dataset from any of the following websites:

   - https://www.gapminder.org/data/
   - http://www.healthdata.org/data-visualization/gbd-compare or http://ghdx.healthdata.org/gbd-2017 (in Select Articles there are folder with data).
   - https://niti.gov.in/state-statistics.

   Take any two parameters, and either a number of Indian states, or a number of countries including India. Then create such a visualization. We rely on you to choose two parameters that make a somewhat interesting story as Hans Rosling does. If you want to use datasets about pandemic that is also fine — either come up with suggestions, or reach out to us.

   Note that you have to be sometimes careful about missing data, data formatting etc these are all part of the problem. Document what problems you faced and what you did to handle these.