CS 328 HW 1

Due: 31/01/2020

Discussion is allowed, but you have to write down the solution yourself. Please also write down names of collaborators. For code, you have to write it yourself entirely. Please follow the honor code.

Submission should be via PDF (preferably generated via Latex / Word) uploaded to Canvas. For code, please share a colab link that we can run and evaluate.

1. (a) In an election with two candidates using paper ballots, each vote is independently misrecorded with probability p = 0.02. Use a Chernoff bound to give an upper bound on the probability that more than 4% of the votes are misrecorded in an election of 1,000,000 ballots.

   (b) Assume that a misrecorded ballot always counts as a vote for the other candidate. Suppose that candidate A received 510,000 votes and that candidate B received 490,000 votes. Use Chernoff bounds to upper bound the probability that candidate B wins the election owing to misrecorded ballots. Specifically, let X be the number of votes for candidate A that are misrecorded and let Y be the number of votes for candidate B that are misrecorded. Bound $Pr((X > k) \cup (Y < l))$ for suitable choices of k and l.

2. (a) 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 $k = 2, 4, 8, 20$. For small values of k, say $k = 2, 3$, find the optimal (when the centers are restricted to be input points) and report the approximation factor obtained by the greedy algorithm.

   (b) (Bonus 5 pt): For $k = 20$, give a solution that beats the greedy algorithm in terms of the objective function value and runs in less than an hour (we will believe you on the time).

3. Is the following a distance function? $d(x, y) = \min_i |x_i - y_i|$. Why or why not?

4. Go through the video at https://www.youtube.com/watch?v=hVimVzgtD6w. There are number of libraries to create such visualization: one example is https://python-graph-gallery.com/341-python-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.

   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.