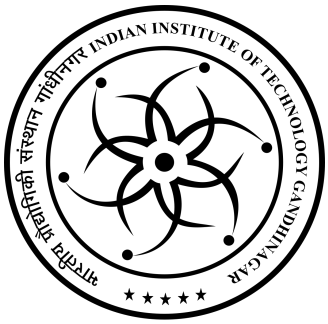

INDIAN INSTITUTE OF TECHNOLOGY GANDHINAGAR

Introduction to Data Science

Assignment 1

Submitted by:

Harshil Jain 17110060

Q.1. a) For i = 1, ..., 1000000, let X_i be an indicator random variable whether the ith ballot is misrecorded (i.e. $X_i = 1$ if the ballot is misrecorded and $X_i = 0$ otherwise). We can compute

$$\mu = E[X] = \sum_{i=1}^{1000000} E[X_i] = \sum_{i=1}^{1000000} 0.02 = 20000$$

Note that 4% of the votes would be $0.04 \times 1000000 = 40000$, so since the X_i are independent, we can apply Chernoff bound

$$Pr(X - \mu > \delta\mu) \leq e^{\frac{\delta^2\mu}{3}}$$

with $\delta = 1$ to get

$$Pr(X > 2\mu) = Pr(X > 40000) \leq e^{\frac{-20000}{3}}$$

Thus, the the probability that more than 4% of the votes are misrecorded in an election of 1,000,000 ballots is less than $e^{\frac{-20000}{3}}$.

Q.1. b) Candidate A received 510,000 votes and candidate B received 490,000 votes. Each vote is independently misrecorded with probability $p = 0.02$.

Let X_i be a random variable that the ith vote received by candidate A is misrecorded. Hence, it is an indicator random variable, i.e. $X_i = 1$ if the ballot is misrecorded and $X_i = 0$ otherwise.

$$\mu_1 = E[X] = \sum_{i=1}^{510000} E[X_i] = \sum_{i=1}^{510000} (0.02 \times 1 + 0.98 \times 0) = 10200$$

Let Y_i be a random variable that the ith vote received by candidate B is misrecorded. Hence, it is an indicator random variable, i.e. $Y_i = 1$ if the ballot is misrecorded and $Y_i = 0$ otherwise.

$$\mu_2 = E[Y] = \sum_{i=1}^{490000} E[Y_i] = \sum_{i=1}^{490000} (0.02 \times 1 + 0.98 \times 0) = 98000$$

Let Z be a random variable defined as $Z = X - Y$.

We want $Pr(Z > 10000)$ for B to win the election.

Hence, by linearity of expectation,

$$E[Z] = E[X - Y]$$

$$\therefore E[Z] = E[X] - E[Y]$$

$$\therefore E[Z] = \mu_1 - \mu_2$$

$$\therefore E[Z] = 10200 - 9800$$

Thus, $\mu' = E[Z] = 400$

Since each Z_i are independent, we can apply Chernoff bound.

$$Pr(Z - \mu' > \delta\mu') \leq e^{\frac{\delta^2\mu'}{3}}$$

$$\therefore Pr(Z - 400 > 400\delta) \leq e^{\frac{400\delta^2}{3}}$$

$$\therefore Pr(Z > 400(1+\delta)) \leq e^{\frac{400\delta^2}{3}}$$

Equating $400(1+\delta) = 10000$

Thus, $\delta = 24$

Hence,

$$Pr(Z > 10000) \leq e^{\frac{400(24)^2}{3}}$$

$$Pr(Z > 10000) \leq e^{-76800}$$

Thus, the probability that B wins the election is less than e^{-76800}.

Q.2.a The greedy and optimal algorithms have been implemented as discussed in class. The optimal algorithm for k = 2 and k = 3 has been run only on the first 100 data points of the dataset.

Q.2.b Explanation of algorithm for bonus question: The greedy algorithm chooses a random point and then looks for the unhappiest point from the selected set of points and keeps on iteratively repeating the process till k points have been chosen. However, the objective function primarily depends on the choice of the random point at the beginning of the algorithm, if it turns out that the point chosen at random in some corner case leads to a non-optimal choice of further centres. Hence, a way to improve the greedy algorithm, is to repeat this process of the greedy algorithm many times, i.e. during each iteration choosing a random point and then following the greedy algorithm. After a certain number of iterations, the set of centres having the least objective function is chosen. This will in most cases lead to a better objective function than just the traditional single iteration of the greedy algorithm. Certainly, it will be lead to a greater objective function than the optimal set of k centres.

The code for this question can be found on this link.

Q.3. Given a set X of points, a distance function on X is a map $d : X \times X \rightarrow R_+$ that is symmetric, and satisfies $d(i, i) = 0$ for all $i \in X$. The distance is said to be a metric if the triangle inequality holds, i.e.,

$$d(i, j) \leq d(i, k) + d(k, j) \forall i, j, k \in X$$

It is often required that metrics should also satisfy the additional constraint that $d(i, j) = 0 \iff i = j$

However $d(x, y) = \min_i |x_i - y_i|$ **does not** satisfy the property: $d(i, j) = 0 \iff i = j$

For example, consider two vectors x and y each n dimensional having different values at each of the n values except one. Hence by the definition of this metric, $d(x, y) = 0$, however $x \neq y$.

Hence $d(x, y) = \min_i |x_i - y_i|$ is not a distance function.

Q.4. The attributes chosen for the purpose of plotting are population growth and the gini co-efficient (measure of inequality) for the years 1989 to 2017.

The main problem faced was to clean the csv files to remove entries where thee data was missing. After this process, the information of countries common to both csv files were filtered out. Also, the gini indices had values from years 1860 to 2017 and the population growth data was only from 1989 to 2017. Hence only the years 1989 to 2017 were used for the purpose of plotting.

The code for this question can be found on this link.
