EEE-361 HOMEWORK 3

April 2, 2021

The purpose of this homework assignment is the practical validation of applying The Johnson-Lindenstrauss Lemma. The lemma formally says that, given vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^m$, for some $n, m \in \mathbb{N}$, and given $\epsilon > 0$ and $k \geq (8 \log n)/\epsilon^2$, then there exists a projection $\boldsymbol{J}$ from $\mathbb{R}^m$ to $\mathbb{R}^k$ that approximately preserves the distances, that is,

$$(1-\epsilon)\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 \leq \left\|\boldsymbol{J}\boldsymbol{x}_i - \boldsymbol{J}\boldsymbol{x}_j\right\|_2^2 \leq (1+\epsilon)\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 , \tag{1}$$

for all $1 \leq i, j \leq n$. As stated in page 155 of the book, a key point given in the proof of the lemma is that a random projection will do for such a purpose, and therefore, there must be many such $k \times m$ projections.

Given the values of m, n and ϵ, you will first determine the value of k to be used and then generate 10 different realizations of the random projection $\boldsymbol{J}$ to be used, as indicated in page 153 of the book. Formally, every column of $\boldsymbol{J}$ will be independently generated from a 0 mean normal distribution with $\sigma^2 = 1/k$. For each realization of $\boldsymbol{J}$, you will write a code that determines whether or not condition (1) holds, for all pairs of independent vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $1 \leq i, j \leq n$. Note that the cases when $i = j$ should not be checked. Here, you should first determine the number of all possible such pairs (let us denote it by N) and then the number of cases when (1) holds (let us denote it by N_{true}). Denoting by v_i the percentage of cases that condition (1) holds for the ith realization of $\boldsymbol{J}$, you will have

$$v_i = \frac{N_{true}}{N} \times 100.$$

Then, v_{av} will denote the average of these percentages over all realizations of $\boldsymbol{J}$, that is,

$$v_{av} = \frac{1}{10} \sum_{i=1}^{10} v_i.$$

So, given the values for m, n, ϵ, and the vectors $\boldsymbol{x}_i$, $1 \leq i \leq n$, you now have a number v_{av}. What we want to see is the accuracy of the preservation of distances over different values of m, n and ϵ. Therefore, you will find the values of v_{av} for the cases when $m = 10^4, 5 \times 10^4, 10^5$ and $\epsilon = 0.1, 0.3, 0.7, 0.9$. The value of n is dependent on m and you will have three such values for it, namely, $n = m, m/10, m/100$. The tasks you should complete are listed below. First, note that for each n and m, you will first generate the vectors $\boldsymbol{x}_i \in \mathbb{R}^m$, for $1 \leq i \leq m$. You will do this by independently generating n realizations of length m vectors with i.i.d. zero-mean unit-variance normal random variables.

1. For each $m \in \{10^4, 5 \times 10^4, 10^5\}$, you will plot the values of v_{av} with respect to the values of $\epsilon \in \{0.1, 0.3, 0.7, 0.9\}$. There will be three lines in the plot, each corresponding to the three different values of n as given above. Note that you will end up with a total of three figures, each containing three lines.
2. For each $\epsilon \in \{0.1, 0.3, 0.7, 0.9\}$, you will plot the values of v_{av} with respect to the values of $m \in \{10^4, 5 \times 10^4, 10^5\}$. There will be three lines in the plot, each corresponding to the three values of n as given above. Note that you will end up with a total of four figures, each containing three lines.

Report your plots (note that there will be a total of 7 such plots, each of which contains three different lines corresponding to the values of n) and comment on the results. What trend do you see and why? How does the dimension affect the accuracy? How does the sample size and ϵ affect it?