CSC311: Assignment 1 Writeup

Kiarash Sotoudeh

Question 1

(a)

Suppose we have a classification dataset where each data point has one feature. The feature takes on a real value between $[0, 1]$. What is the minimum number of data points we need to guarantee that any new test point is within ($\leq$) 0.01 of an old point? (equivalently: What is the smallest set of points S such that every point in $[0, 1]$ is within 0.01 of a point in S?)

Answer: 50 points.

We want every point in $[0, 1]$ to be within max 0.01 of a point in S. We can divide the interval $[0, 1]$ into intervals of length 0.02. This is because in worst case scenario, the new test point is in the middle of two existing points and in that case the distance to the nearest point is 0.01. The example set of points is $S = \{0.01, 0.03, 0.05, \ldots, 0.99\}$. Counting these points we see that only 50 points are needed.

(b)

Explain why such a guarantee is more difficult to maintain when we are working on a problem with 10 features.

Answer: When we have 10 features we have to make sure the new test point is within 0.01 of an old point in all 10 dimensions. In that case we have to check the distance for each component of the vector. According to the curse of dimensionality discussed during class, the total volume of $[0, 1]^d$ is 1 and so $O((\frac{1}{\epsilon})^d)$ points are needed to cover the space. This means for $d = 10$ with $\epsilon = 0.01$ we need $O((\frac{1}{0.01})^{10}) = O(10^{20})$ points to cover the space!

(c)

For each choice of dimension $d \in [2^0, 2^1, 2^2, \ldots, 2^{10}]$, sample 100 points from the unit cube, and record the following average distances between all pairs of points, as well as the standard deviation of the distances.

- (i) Squared Euclidean or ℓ_2 distance $= \|\mathbf{x} - \mathbf{y}\|_2^2 = \sum_j (x_j - y_j)^2$
- (ii) ℓ_1 distance $= \|\mathbf{x} - \mathbf{y}\|_1 = \sum_j |x_j - y_j|$

Plot both the average and standard deviation as a function of d.

Answer:

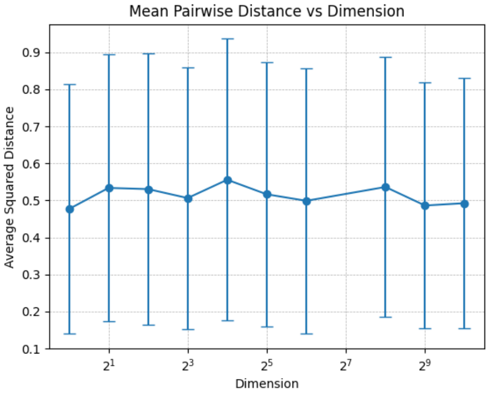

Figure 1: Average distance between points in the unit cube.

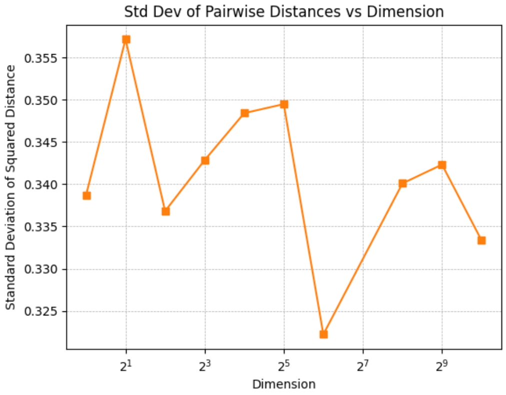

Figure 2: Standard deviation of distance between points in the unit cube.

(d)

Answer: I dont even know bro