Project 1 write up

Amanuel Tesfaye

COSC347S24

1 Implement Learning

Implemented a classifier function that takes in points in space (collected from a csv file or generated by some distribution D) as well as their corresponding labels to create a d-dimensional container bounded by the minimum and maximum values of our points in each dimension. Used this function to classify points as either "+" or "-" depending on their location with respect to this container. My classifier defaults to "-" if there are no "+" points in the distribution of interest.

2 Make Some Heatmaps

I created three Heatmaps.

2.1 First Heatmap

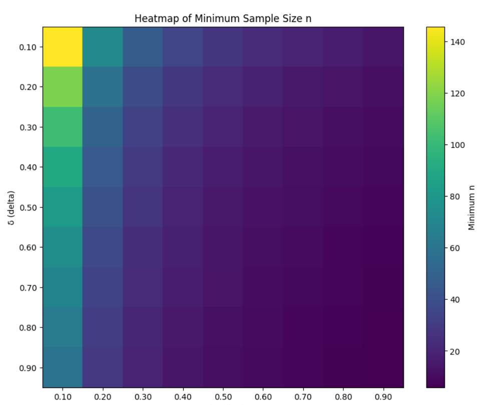

Figure 1: Heatmap using the bound we derived in class

2.2 Second Heatmap

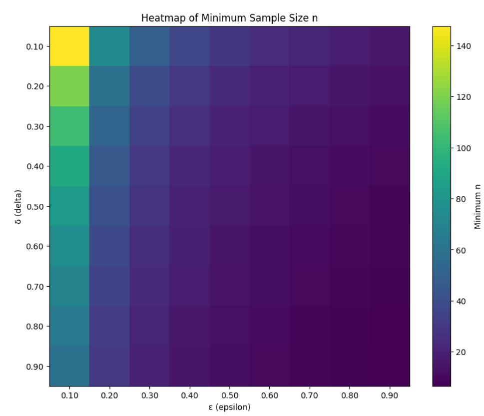

Figure 2: Heatmap using second bound

2.3 Generated a Uniform Distribution in 2D to numerically estimate our bounds

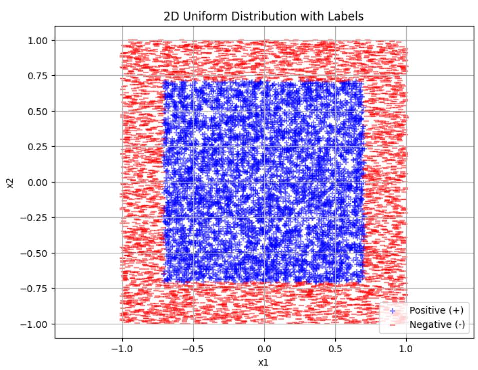

Figure 3: Uniform distribution used with 10,000 points in 2D

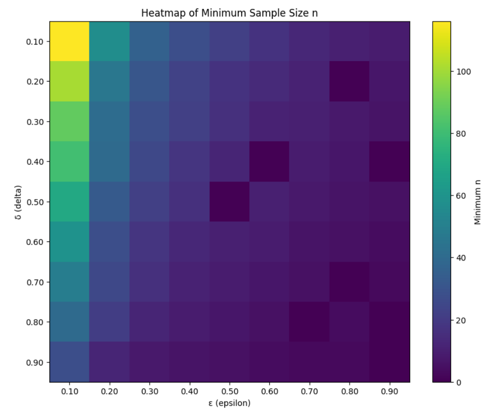

Figure 4: Empirical estimation of our bounds using T=200 for various n

For a given n, I used $(f - volume - h - volume)/f - volume$ to compute the generalization error

3 Muck With d

I generalized my classifier to work with any arbitrary d, given some distribution D.

3.1 More Heatmaps

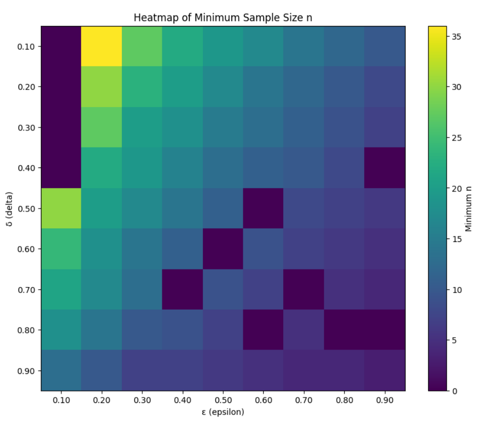

Figure 5: Heatmap with n=3

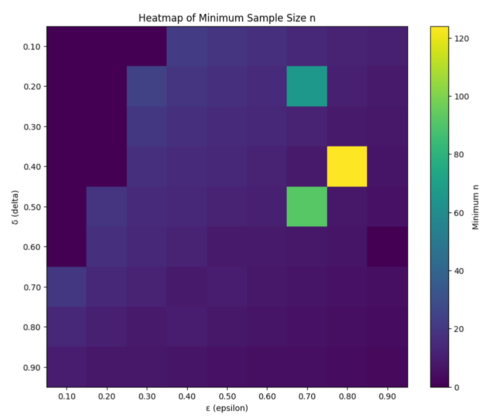

Figure 6: Heatmap with n=4

3.2 Plotting Distances

I used the Manhattan distance (L1 norm) to calculate the difference between observed and expected values. I used this mainly because I was mostly seeing a one-sided shift (towards lower/darker n values) on my Heatmaps. Thus, I wouldn't lose any information by taking the L1 norm of my arrays.

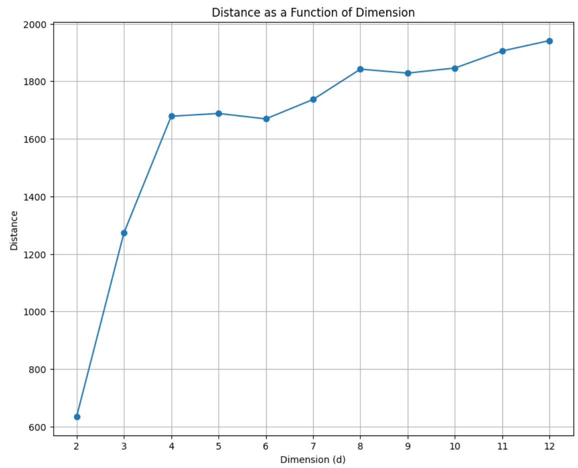

Figure 7: Ploting L1 norm for several d values

4 Muck With D

I played around with different distributions and estimated the generalization errors this time around.

4.1 Triangular Distribution

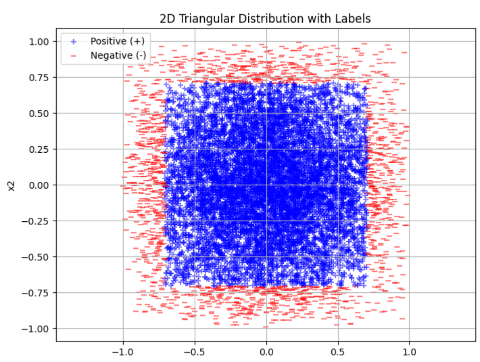

Figure 8: Triangular distribution used with 10,000 points in 2D

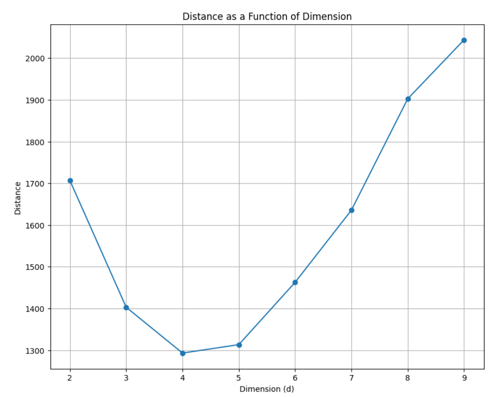

Figure 9: Plotting L1 norm for several d values

4.2 Laplace Distribution (scale = 0.2)

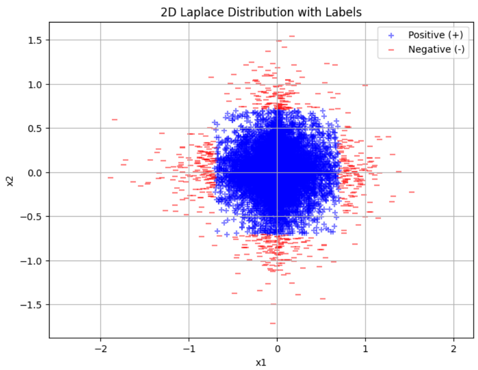

Figure 10: Laplace distribution in 2D (scaled by 0.2)

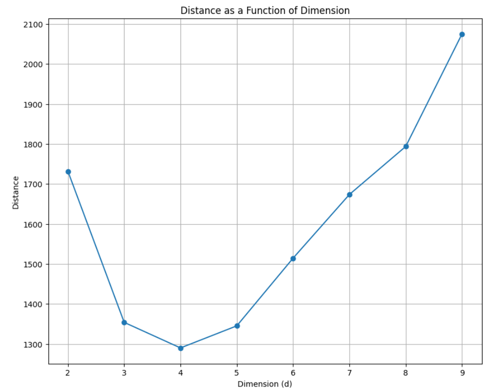

Figure 11: Plotting L1 norm for several d values

Additionally, the observed values of n tend to be generally lower than the theoretical ones (more Heatmaps on google colab).

5 Go For Gold (Twice)

Among the distributions considered, the uniform distribution exhibited the greatest deviation from the theoretical bounds. This distribution doesn't allow for high variability in the values of n. Additionally, using a Laplacian distribution with a scale of 0.1 minimized the difference between the bounds. It has a high density of "+"ses around the center of the distribution with sparsely populated "-"ses spread in the periphery.