Henry Bowman Hill

Assignment 1

Problem 2a

Question: WHY does the accuracy here (monotonically) increase as the number of dimensions D increases? This seems unintuitive and against the “curse of dimensionality” discussed in class. Write your answer into a Word/Overleaf and save it as problem2a.pdf

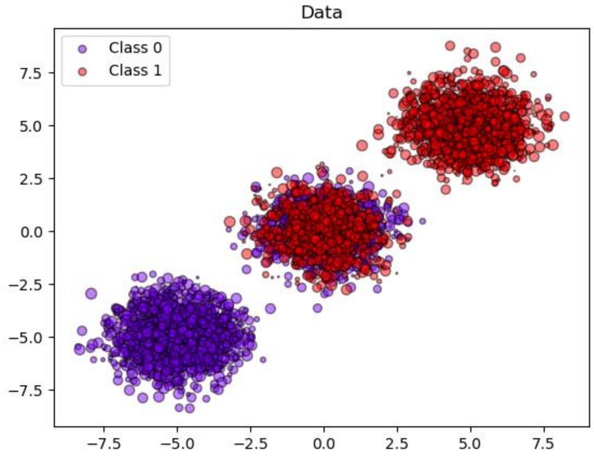

*Figure 1*

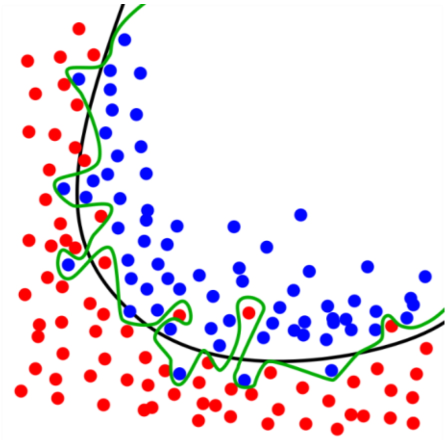

*Figure 1*

If you examine the data set from sk-learn, it gives you Figure 1 visualizing how the data is formed. In the file the function is generated with 6 centers and a massive standard deviation of 5, as you look at the data it becomes apparent that accuracy is increasing directly with dimensions because of an overfit line. Figure 2 is an example of how a model can overfit data in an attempt to improve the model accuracy where the green line represents an overfit line, and the black line is the true regression. The accuracy may be increasing but only because the model is losing the ability to generalize to new data and some variance is unknowingly being used to represent the model structure. This is most likely due to noisy values in the data set, causing the KNN to memorize this particular data set. The model will have limited applicability to new or unseen data as it cannot find patterns and produce correct results. As the number of dimensions increase the model is forced to estimate using a fixed sample size, which is why we see accuracy monotonically increase and irrationally stop at perfect accuracy past 20 dimensions.