Statistical Learning - Homework 3 (Conceptual)

James Hahn

Chapter 4 - Exercise 4

a) If $x \in [0.05, 0.95]$, then the neighboring observations are in the range $[x - 0.05, x + 0.05]$, representing 10% of the data. If $x < 0.05$, then the neighboring observations are in the range $[0, x + 0.05]$, representing $(100x + 5)\%$ of the data. Finally, if $x > 0.95$, then the neighboring observations are in the range $[x - 0.05, 100]$, representing $(105 - 100x)\%$ of the data. Therefore, if we calculate the expected value of these three probabilities, we get $\int_{0.05}^{0.95} 0.10dx + \int_{0}^{0.05}(100x + 5)dx + \int_{0.95}^{1}(105 - 100x)dx = 9 + 0.375 + 0.375 = 9.75$. Therefore, on average, the fraction of neighboring observations used to make predictions is 9.75%.

b) If X_1 and X_2 are independent, then the expected fraction of available observations used for predictions is $9.75\% \times 9.75\% = 0.9506\%$.

c) If we follow from the pattern in part b), the expected proportion of available observations used for predictions is $9.75\%^{100} \approx 0\%$.

d) If we use an argument with the pattern we discovered in part c), we'll notice $\lim_{x\to\infty}(9.75\%)^p = 0\%$. So, as p, or the number of features, increases, we can see there are fewer and fewer observations nearby to make an estimate for test observations.

e) As seen in the question, when p = 1, the hypercube is a line segment, so the length is 0.1. When p = 2, the hypercube is a square with lengths $0.1^{1/2}$. If we follow this pattern, we can see for p = 100, the hypercube will have dimensions of length $0.1^{1/100}$. To further extend the question, as p increases, we approach hypercube dimensions of length 1: $\lim_{x\to\infty}(0.1)^{1/x} = 1$.

Homework 3 - Question 2 - parts b) through d)

b) In Chapter 4 Exercise 4, we were observing KNN. To argue against the statement "non-parametric approaches often perform poorly when p is large", we can easily claim if a hypercube's dimensions increase as p increases, the classifier should be able to predict samples efficiently, even if p is large. In general, for non-parametric approaches, if we have a lot of features, but few samples, it will be hard to extrapolate predictions based on nearby samples. One thing to ask about before making the generalization in the above quote is the size of the dataset, or the number of samples. If we have a lot of samples, a test observation should be able to extrapolate from nearby previously observed samples, which was not the case when we did not have a lot of samples. As such, we have shown the generalization is not necessarily always true.

c) In high dimensional space, you are often forced to **overfit**.

d) There are two good answers to this question. In general, more data truly is not a bad thing. This follows the general idea of distributions and the central limit theorem that you learn in statistics and probability courses; the more data you have, the better you can observe the underlying distribution of the data. As such, you can create a model that fits your data even better. However, on the other side, the more data, the higher the probability of random noise being injected into your dataset. This in turn screws up models, hurting inference procedures and accuracies. My general rule of thumb is the first approach, that more data is never a bad thing, but as you can see, sometimes it can have a negative effect on your modeling.

Chapter 4 - Exercise 5

a) With a linear Bayes decision boundary, QDA is expected to be better than LDA for the training set. QDA will be better than LDA because it is more flexible, so it will have a closer fit than LDA. On the test set, LDA is better than QDA since we have the risk of overfitting the linearity on the decision boundary.

b) Since this study has a non-linear decision boundary, QDA is expected to be better than LDA on both the training and testing sets since we have a lower risk of overfitting to data with the new assumption.

c) As the training set increases, we expect test accuracy of QDA (which is more flexible and has higher variance than LDA) to improve relative to LDA since variance of data is less of a concern, so overfitting is in turn less of a concern.

d) False because when we have a small sample size, QDA may lead to overfitting (higher variance of data), leading to a lower test accuracy.

Chapter 4 - Exercise 8

In KNN with K=1, the training error rate is 0%. Since the average of training error and testing error is 18%, $(0\% + x\%)/2 = 18\%$, where x is the testing error percentage. With basic algebra, we find the test error to be 36%, which is higher than the logistic regression testing error. As such, we can explain this KNN situation as overfitting. Thus, we want to use the logistic regression.