Bruce Campbell ST-617 Discussion Group 3

Tue Jul 05 09:27:36 2016

Chapter 4

Problem 4

a)

In 1 dimension we will use $\frac{1}{10}$ of the available data if we classify using observations that are within 10% of the range of the predictor to our test point.

b)

In 2 dimensions we will use $\frac{1}{100}$

c)

In p dimensions we will use $\frac{1}{10^p}$

d)

From above we see that if we have n points in our training set and the dimension is p, then there will be on average $\frac{n}{10^p}$ data points on average in a neighborhood of a test point that only includes obervations within 10% of each predictors range. Forlarge p this will be a small number. We also see that if we'd like k points in this neighborhood on average then we would need a test set that contains $k * 10^p$ points.

e)

For $p = 1$ to capture 10% of the data we need an interval of length $l = \frac{1}{10}$, for $p = 2$ we would need an square of length $l = \sqrt[2]{\frac{1}{10}}$, and for $p = 100$ we would need a hypercube interval with sides of length $l = \sqrt[100]{\frac{1}{10}}$. We see that we need larger and larger proportion of the feature space to capture the required fraction of data. In fact $lim_{p \to +\infty} \sqrt[p]{\frac{1}{10}} = 1$