Hans Trautlein
Due: March 3, 2016

Problem Set 4
Andrew Bray

Chapter 4

4. *When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.*

   (a) *Suppose that we have a set of observations, each with measurements on* $p = 1$ *feature,* X*. We assume that* X *is uniformly (evenly) distributed on* $[0, 1]$*. Associated with each observation is a response value. Suppose that we wish to predict a test observations response using only observations that are within 10% of the range of* X *closest to that test observation. For instance, in order to predict the response for a test observation with* $X = 0.6$ *we will use observations in the range* $[0.55, 0.65]$*. On average, what fraction of the available observations will we use to make the prediction?*

   This kind of makes me think of one of the exam questions, with the "curse of dimensionality." Here you would only use about one tenth ($\frac{1}{10}$) of all predictions to make this prediction.

   (b) *Now suppose that we have a set of observations, each with measurements on* $p = 2$ *features,* X_1 *and* X_2*. We assume that* (X_1, X_2) *are uniformly distributed on* $[0, 1] \times [0, 1]$*. We wish to predict a test observations response using only observations that are within 10% of the range of* X_1 *and within 10% of the range of* X_2 *closest to that test observation. For instance, in order to predict the response for a test observation with* $X_1 = 0.6$ *and* $X_2 = 0.35$*, we will use observations in the range* $[0.55, 0.65]$ *for* X_1 *and in the range* $[0.3, 0.4]$ *for* X_2*. On average, what fraction of the available observations will we use to make the prediction?*

   Not always, but *on average* you would see use on 1% of the obvservations ot make this prediction.

   (c) *Now suppose that we have a set of observations on* $p = 100$ *features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observations response using observations within the 10% of each features range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?*

   This is a very teeny tiny number, not one that you'd normally deal with on a day to day basis. It is $0.10^{100} * 100\% = 10^{-98}\%$.

   (d) *Using your answers to parts (a)(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.*

   This is a linear vs. exponential problem. p will increase in a linear fashion whereas the "near" observations themselves will decrease at an exponential rate. They also will not be that "near" as you choose a larger and larger p.

(e) *Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube*[1] *centered around the test observation that contains, on average, 10% of the training observations. For p =1, 2, and 100, what is the length of each side of the hypercube? Comment on your answer.*

$$p = 1,\ l = 0.10$$
$$p = 2,\ l = \sqrt{0.10} \approx 0.32$$
$$p = 3,\ l = 0.10^{1/3} \approx 0.46$$
$$\dots$$
$$p = n,\ l = 0.10^{1/n}$$

This will slowly approach one but never completely reaches it, it seems like one is the asymptote for the function $l = 0.10^{1/n}$ or $\lim_{n\to\infty} 0.10^{1/n} = 1$.

6. *Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = 6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.*

   (a) *Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class.*

$$X = [40h, 3.5GPA]$$
$$p(X) = \frac{e(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + e(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$
$$p(X) = \frac{e(-6 + 0.05 X_1 + X_2)}{1 + e(-6 + 0.05 X_1 + X_2)}$$
$$p(X) = \frac{e(-6 + 0.0540 + 3.5)}{1 + e(-6 + 0.0540 + 3.5)}$$
$$p(X) = \frac{e(-0.5)}{1 + e(-0.5)} = \mathbf{37.8}\%$$

   (b) *How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?*

   I have skipped some of the explicit steps here as it can be solved by rearranging and simplifying the equation a few times.

[1] *Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.*

$$
\begin{aligned}
X &= [X_1 h, 3.5GPA] \\
p(X) &= \frac{e(-6 + 0.05X_1 + X_2)}{1 + e(-6 + 0.05X_1 + X_2)} \\
0.50 &= \frac{e(-6 + 0.05X_1 + 3.5)}{1 + e(-6 + 0.05X_1 + 3.5)} \\
0.50(1 + e(-2.5 + 0.05X_1)) &= e(-2.5 + 0.05X_1) \\
0.50 + 0.50e(-2.5 + 0.05X_1)) &= e(-2.5 + 0.05X_1) \\
0.50 &= 0.50e(-2.5 + 0.05X_1) \\
\log(1) &= -2.5 + 0.05X_1 \\
X_1 &= 2.5/0.05 = \textbf{50 hours}
\end{aligned}
$$

7. *Suppose that we wish to predict whether a given stock will issue a dividend this year (Yes or No) based on X, last years percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didnt was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.*[2]

   I have to use that enormous equation!

$$p_{yes}(x) = \frac{\pi_{yes}\frac{1}{\sqrt{2\pi}\sigma}e(-\frac{1}{2\sigma^2}(x - \mu_{yes})^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma}e(-\frac{1}{2\sigma^2}(x - \mu_l)^2)}$$

   Then I have to plug in the values above for $pi_yes = 0.80$, $\mu_{yes} = 10$, $\sigma^2 = 36$, $\pi_{no} = 0.20$, $\mu_{no} = 0$, and $X = 4$. With that I end up getting:

$$p_{yes}(4) = \frac{0.80e(-\frac{1}{2*36}(4 - 10)^2)}{0.80e(-\frac{1}{2*36}(4 - 10)^2) + 0.20e(-\frac{1}{2*36}(4)^2)} \approx \textbf{75.2\%}$$

   So given that its percentage profit was $X = 4$ last year the chance that this company will issue a dividend is about 75.2%.

2. ***Challenge Problem**: It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes classifier assigns an observation to the class for which the discriminant function is maximized.*

   I honestly don't think I have the mathematic ability to do this correctly but I wanted to keep the question in the problem set for my records!

[2] *Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$. You will need to use Bayes' theorem.*