ASSIGNMENT 2

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 curse of dimensionality non-parametric approaches often performs 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 observation's 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?

Ans. For p = 1, with the feature X. Since we are given to use only 10% of the range of the uniform distribution at range between 0 and 1. Hence, the fraction of available observations that we will make prediction will be

For X = 0.6 X € [0.55,0.65] = (0.65-0.55) * 100 / (1-0) = 10%

(b) Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1, X2) are uniformly distributed on [0, 1] × [0, 1]. We wish to predict a test observation's response using only observations that are within 10 % of the range of X1 and within 10 % of the range of X2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0.6 and X2 = 0.35, we will use observations in the range [0.55, 0.65] for X1 and in the range [0.3, 0.4] for X2. On average, what fraction of the available observations will we use to make the prediction?

ANS. The given set of observations with P = 2, with features X1 and X2 that are also uniform distribution over the same range such that (X1, X2) € [0,1] X [0, 1]

So, the fraction of available observations that we will use to make prediction is given by :

= (10%) * (10%) = 1% .

(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 observation's response using observations within the 10 % of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

ANS. Here we are given with the set of observations with P = 100 features and all of which are uniform distributed over the range between 0 and 1.

So, the fraction of available observations that we will use to make prediction is given by

$= (0.1)^{100} * 100 = (10)^{-98\ \%}$

(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.

ANS. This is curse of dimensionality effect, from the above discussion, as the number of features increases then the percentage of observations that are used to predict KNN very small. Hence, there are fewer neighbors.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube 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.

ANS. According to the question

For P=1 => length = (0.1)

For p =2 -> length is $(0.1)^{½}$

=0.316

For p = 100 -> length = $(0.1)^{1/100} = 0.977$

If we have large number of features, it will be better to include all the features.

6. Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, βˆ0 = −6, βˆ1 = 0.05, βˆ2 = 1.

Logistic function is given by

$P(x)= e^{\beta 0 + \beta 1 X1 + \beta 2 X2} / 1 + e^{\beta 0 + \beta 1 X1 + \beta 2 X2}$

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

ANS. Here X1 = 40hrs, X2 = 3.5

P(X) = $e^{-6 + 0.05*40 + 1*3.5}$ /1+ $e^{-6+0.05*40+1*3.5}$ = $e^{-0.5}$ 1+ $e^{-0..5}$ = 37.75 %

(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?

ANS. Here P(x) = 0.5 , X2 = 3.5

0.5 = $e^{-6 + 0.05*X1 + 1*3.5}$ / (1 + $e^{-6+0.05*X1+1*3.5}$)

$e^{-6 + 0.05*X1 + 1*3.5}$

0.5 * $e^{0.05*X1-2.5}$ = 0.5

$e^{0.05}*X1-$ 2.5 = 1

0.05*X1 – 2.5 = $\log_e(1)$

0.05 * X1 = 2.5

X1 = 50 hrs.

7. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine many companies and discover that the mean value of X for companies that issued a dividend was X¯ = 10, while the mean for those that didn't was X¯ = 0. In addition, the variance of X for these two sets of companies was ˆσ2 = 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.

Hint: Recall that the density function for a normal random variable is f(x) = √ 1 2πσ2 e−(x−µ)2/2σ2 . You will need to use Bayes' theorem.

ANS. According to Bayes theorem -

Pk(x) = $\frac{\pi_k f_k(x)}{\Sigma_{l=1}^{k} \pi_l f_l(x)}$

$f_k(x)$ = $\frac{\mathbf{1}}{(\sqrt{2}\boldsymbol{\pi}\boldsymbol{\sigma}_k\mathbf{2})}$ $e^{\frac{-\mathbf{1}}{2\boldsymbol{\sigma}^2}(x-\mu_k)^2}$

Here π_{yes} = 0.8, π_{no} = 0.2, μ_{yes}= 10, μ_{no}= 0, x=4, σ^2 =36

$P_{yes}(x)$=

$$0.8 * \frac{1}{\sqrt{2}\pi(36)} e^{\left(-1 \frac{1}{2*(36)}(4-10)^2\right)}$$

$$0.8 * \frac{1}{\left(\sqrt{2}\pi 36\right)} e^{\left(\frac{-1}{2*(36)}(4-10)^2\right)+0.2*\frac{1}{\sqrt{2}\pi(36)}e\left(-\frac{1}{2*(36)}(4-0)^2\right)}$$

$P_{yes}(x)$=

$$e^{\frac{-1}{2*36}(36)}$$

$$0.8 * e^{\left(\frac{-1}{2*(36)}(36)+0.2*e^{\frac{-1}{2*36}(16)}\right)}$$

$P_{yes}(x)$=0.752

In other words, there is a 75.2% chance that a company will issue a dividend this year given that its percentage profit was X=4 last year.

9. **This problem has to do with odds.**

(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

Odds = $\frac{P(x)}{1-P(x)}$

0.37 = $\frac{P(x)}{1-P(x)}$

P(x) + 0.37 * P(x) = 0.37

P(x) = 0.27

P(x) = 27%.

With a 0.37 chance of missing a credit card payment, approximately 27% of people will default.

(b) Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?

ANS. Plugging in 0.16 for the probability of default into the odds formula, we get an odds of 0.16/(1−0.16)=.1905*0.16/(1−0.16)=.1905* that the individual will default.

Chapter 5

2. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

Given number of observations = n.

Since bootstrap allows sampling with replacement,

->(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

ANS, Every observation in original sample is independent and has equal probability to appear in each bootstrap observation, and probability that j^{th} observation is is from original sample = $\frac{1}{n}$

Probability that the first bootstrap is not the j^{th} observation from the original sample is 1 - $\frac{1}{n}$

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

ANS.The probability that the second observation is not the j^{th} observation from the original sample is again 1 - $\frac{1}{n}$

(c) Argue that the probability that the jth observation is not in the bootstrap sample is (1 − 1/n)

Ans. Since we are sampling with replacement and the total observations always remain the same I.e, ***n,*** then the probability that h^{th} observation is not in bootstrap sample is $(1 - \frac{1}{n})^n$.

Since there have been n – no.of replacements and the individual events are independent.

(d) When n = 5, what is the probability that the jth observation is in the bootstrap sample?

ANS. From the above we have derived that the formula is $(1 - \frac{1}{n})^n$

SInce the n = 5, then (1 - $\frac{1}{5}$) ^{5} = $\frac{2101}{3125}$ = 0.672

(e) When n = 100, what is the probability that the jth observation is in the bootstrap sample?

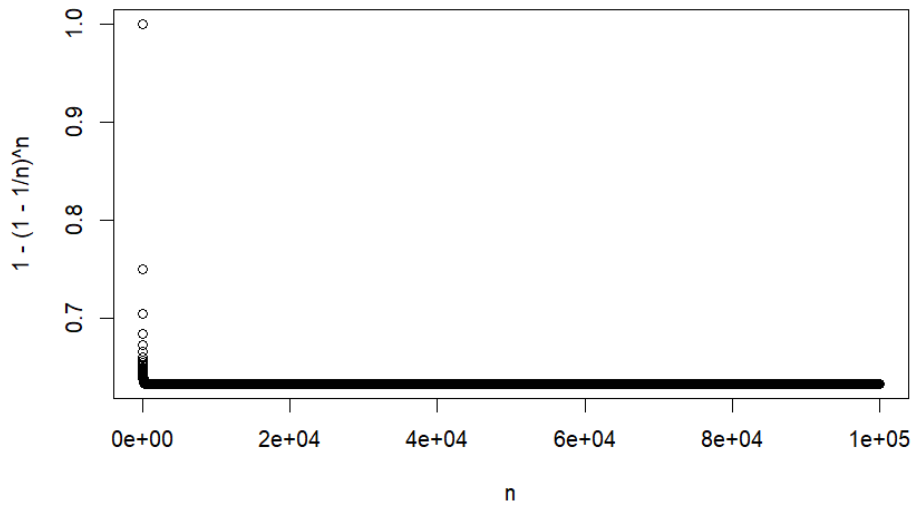
ANS. When $n=100$ the probability that the jth observation is in the bootstrap sample is $1-(1-1/100)^{100}\approx 0.634$

(f) When n = 10, 000, what is the probability that the jth observation is in the bootstrap sample?

ANS. When $n=1000$, the probability that the jth observation is in the bootstrap sample is $1-(1-1/1000)^{1000}\approx 0.632$.

(g) Create a plot that displays, for each integer value of n from 1 to 100, 000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

```
> n <- seq(1,100000)
> plot(n,1-(1-1/n)^n)
```



ANS.

(h) We will now investigate numerically the probability that a bootstrap sample of size n = 100 contains the jth observation. Here j = 4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store <- rep(NA, 10000)
> for(i in 1:10000){
    store[i] <- sum(sample(1:100, rep=TRUE) == 4) > 0
}
> mean(store)
```

Comment on the results obtained.

data <- rep(NA, 10000)

```
> for (i in 1:10000)
+ {
+   data[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
+ }
> mean(data)
[1] 0.6347
```

The resulting fraction of 10,000 bootstrap samples that have the 4th observation is close to our predicted probability of 1-(1-1/100)^100 = 63.4%

3. We now review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.

Ans. If there are n observations, we can do K-fold cross validation by dividing the data into k equal groups of k n. (approximately). The algorithm is now fit to the remaining (k - 1) groups, and the mean squared error, MSE1, is calculated, using the first group as a validation set. This process is repeated k times, with a new group being taken into account for the validation set each time. In this method, each group will be used as a training set (k-1) times and just once as a validation set. By averaging the values, one can calculate the k estimates of test error that are produced by this method, which is given by:

$CV(k) = 1/k \sum^{k}_{i=1} MSE_i$

(b) What are the advantages and disadvantages of k-fold cross validation relative to i. The validation set approach? ii. LOOCV?

Validity set Approach:

Benefits: This strategy is easy to execute and has a straightforward conceptual foundation.

Disadvantages:

• Depending on which observations are included in the training set and which observations are included in the validation set, the validation estimate of test error rate can vary greatly.

• In this method, the model is only fitted to observations that are part of the training set, therefore the model may perform worse if fewer observations are utilized in training. This can potentially cause the test error rate to be overestimated.