Statistical Data Mining I

Homework 3

Student Number:71

Problem 1:

Logistic Regression:

- We use ‘crim’ and ‘crim01’ variables for generalized logistic regression. The Test Error obtained after fitting this generalized logistic regression model on the training data set is: 0.1818182 i.e. 18.1818182%
- We now use ‘crim’, ‘crim01’, ‘chas’ and ‘nox’ variables for generalized logistic regression. The Test Error obtained after fitting this generalized logistic regression model on the training data set is: 0.1581028 i.e. 15.181028%

LDA:

- The Test Error obtained after fitting LDA model using ‘crim’ and ‘crim01’ variables is 0.1343874 i.e. 13.4387352%.
- The Test Error obtained after fitting LDA model using ‘crim’, ‘crim01’, ‘chas’ and ‘nox’ variables is 0.1501976 i.e. 15.0197628%.

KNN:

- The Test Error obtained after performing KNN on the training data set with k=1 is 0.458498 i.e. 45.8498024%.
- The Test Error obtained after performing KNN on the training data set with k=10 is 0.1185771 i.e. 11.85771075%.
- The Test Error obtained after performing KNN on the training data set with k=100 is 0.4901186 i.e. 49.0118577%.

Findings:

1. There’s not much difference between the test errors.
2. The kNN model where k=10 performs better than the other models and gives us the best error rate.

Problem 4:

a)

1. $Y=\beta_0+\beta_1X+\varepsilon$

   The LOOCV error obtained after fitting this model is 7.288162.

2. $Y=\beta_0+\beta_1X+\beta_2X_2+\varepsilon$

   The LOOCV error obtained after fitting this model is 0.9374236.

3. $Y=\beta_0+\beta_1X+\beta_2X_2+\beta_3X_3+\varepsilon$

   The LOOCV error obtained after fitting this model is 0.9566218.

4. $Y=\beta_0+\beta_1X+\beta_2X_2+\beta_3X_3+\beta_4X_4+\varepsilon$

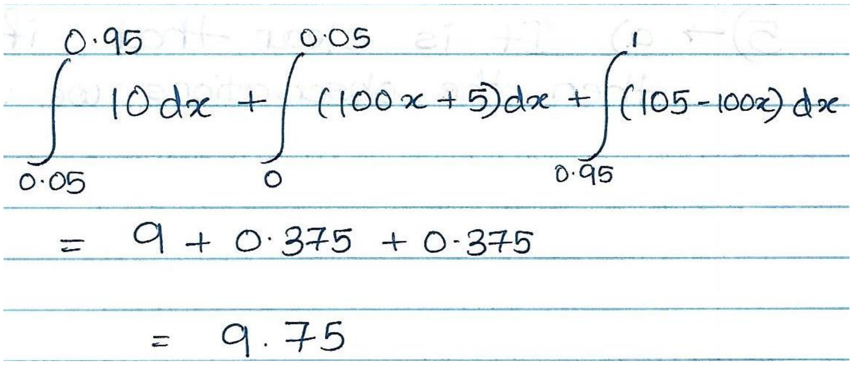
   The LOOCV error obtained after fitting this model is 0.9539049.

b) We can see the LOOCV estimate for the test MSE is minimum for “fit.glm.2”, as the relation between x and y is quadratic this is not surprising.

c) The p-values show that the linear and quadratic terms are statistically significant and that the cubic and 4th degree terms are not statistically significant. This agrees strongly with cross-validation results which were minimum for the quadratic model. All the models agree, on the 5% confidence level; The squared term is seen as significant in all the equation it is present, and the reaming terms are not seem as significant at this confidence level in any model.

Problem 5:

a) If $x \in [0.05, 0.95]$ then the observations are in the interval $[x-0.05, x+0.05]$ and a length of 0.1 is consequently represented which represents a fraction of 10%. We will use observations in the interval $[0, x + 0.05]$ if $x < 0.05$ which represents a fraction of $(100x+5)$ %; by a similar argument we conclude that if $x>0.95$, then the fraction of observations we will use is $(105-100x)$ %. To make the prediction we have to evaluate the following expression to compute the average fraction



$$\int_{0.05}^{0.95} 10\,dx + \int_{0}^{0.05} (100x+5)\,dx + \int_{0.95}^{1} (105-100x)\,dx$$

$$= 9 + 0.375 + 0.375$$

$$= 9.75$$

So we may conclude that, on average, the fraction of available observations we will use to make the prediction is 9.75%.

b) If we assume X_1 and X_2 to be independent, the fraction of available observations we will use to make, the prediction is 9.75% * 9.75% = 0.950625%.

c) We may conclude that the prediction is $9.75\%^{100} \simeq 0\%$ using the fraction of available observations.

d) As we saw in (a)-(c), the fraction of available observations we will use to make the prediction is $(9.75\%)^p$ with p the number of features. So when $p \to \infty$, we have

$$\lim_{p \to \infty} (9.75\%)^{p} = 0$$

3) → a) We will prove this using induction

Let k=1 i.e, there is only class in the dataset.

Let n = total no. of rows in the dataset

if k = 1 → probability of that class being chosen $= \frac{n}{n} = 1$.

Let k=2, i.e, there are two classes in dataset

The complete probability of any of the class being chosen is

Let k be the no. of observations from class 1, i.e (n-k) are the observations from class 2.

Sum of posterior probabilities $= \frac{k}{n} + \frac{(n-k)}{n}$

$$= \frac{k+n-k}{n} = 1$$

Let k=(z-1), i.e there are (z-1) classes

Sum of posterior probabilities $= \frac{1}{n} \sum_{i=1}^{z-1} k_i + (n-k_i) = 1$

Hence by induction, sum of posterior probabilities of classes is equal to one.

3) → b)

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

is equivalent to

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 x)$$

→ Let $Z = e^{\beta_0 + \beta_1 X}$

∴ Equation becomes,

$$p(X) = \frac{Z}{1+Z}$$

$$\frac{1}{p(X)} = \frac{1+Z}{Z} = 1 + \frac{1}{Z}$$

$$\therefore Z = \frac{1}{\frac{1}{p(X)} - 1} = \frac{1}{\frac{1-p(X)}{p(X)}}$$

$$\boxed{Z = \frac{p(X)}{1-p(X)}}$$