IS 777- Fall 19
HOME WORK 2- WRITTEN

Tejaswini Srinivas
fg43775@umbc.edu

Collaborated with –
Aishwarya Bhonde
Harish Ramamoorthy

1 Regression

a) Explain the differences between the KNN classifier and KNN regression methods.

KNN classifier and KNN regression are closely related mathematically. When the value of K and the prediction point (x0) are given, KNN regression identifies the K training observations closet to the x0, which is represented by N0. It then estimates f(x0) using the average value of the training observations values and it is represented as N0. This can be written as

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

The differences between them are the outcome of the observations-
KNN classifier is to resolve the classification problems for qualitative response(Y) as the majority class of the k-nearest neighbor.
KNN regression is to resolve regression problem to predict the quantitative value for f(X) with the average value of training observation Y's in the k-nearest neighbor of X.

b)

a) Least Square Regression Line equation is given

ŷ = β0+β1X1+ β2X2+ βX3+ β4X4+ β5X5

Applying the values to the above equation

ŷ=50+10GPA+1IQ+30GENDER+0.01GPA*IQ-10GPA*GENDER

I. For a fixed value of IQ and GPA, males earn more on average than females.

ŷ = 50+10GPA+1IQ+30*0+0.01GPA*IQ-10GPA*0
ŷ =50+10GPA+1IQ+0.01GPA*IQ

II. For a fixed value of IQ and GPA, females earn more on average than males.

ŷ = 50+10GPA+1IQ+30*1+0.01GPA*IQ-10GPA*1
ŷ =80+10GPA+1IQ+0.01GPA*IQ-10GPA
ŷ =80+1IQ+0.01GPA*IQ

By equation the result of i and ii

50+10GPA+1IQ+0.01GPA*IQ = 80+1IQ+0.01GPA*IQ

50+10GPA=80

GPA=3

Assuming GPA=3

Substitute this in the results of i and ii

ŷ =50+10GPA+1IQ+0.01GPA*IQ

ŷ =50+10*3+1IQ+0.01*3*IQ -----result (I) when GPA=3

ŷ =80+1IQ+0.01GPA*IQ

ŷ =80+1IQ+0.01*3*IQ-----result (II) when GPA=3

From the above results Male and Female earn same values so I and II are incorrect.

III. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough

ŷ =50+10GPA+1IQ+0.01GPA*IQ

IV.For a fixed value of IQ and GPA, females earn more on average than females provided that the GPA is high enough.

ŷ =80+1IQ+0.01GPA*IQ

Considering GPA(female) =3.6 <GPA(male)=4 and IQ is fixed

ŷ =80+1IQ+0.01GPA*IQ

ŷ =80+1IQ+0.01*3.6*IQ

ŷ =80+1IQ+0.036*IQ ---------------Female

ŷ =50+10GPA+1IQ+0.01GPA*IQ

ŷ =50+10*4+1IQ+0.01*4*IQ

ŷ =90+1IQ+0.04*IQ ----------------Male

For fixed value of GPA and IQ , provided the GPA high for males, male earn more than female , which means that statement III is correct .

Now consider GPA(female) =4> GPA(male) =3.6 and IQ is fixed

ŷ =80+1IQ+0.01*4*IQ

ŷ =80+1IQ+0.04*IQ ---------------Female

ŷ =50+10GPA+1IQ+0.01GPA*IQ

ŷ =50+10*3.6+1IQ+0.01*3.6*IQ

ŷ =86+1IQ+0.036*IQ -------------------Male

For fixed value of GPA and IQ , provided the GPA high for female, female earn slightly less than man , which means that statement IV is incorrect.

From this we can say that Male earn more than Female if the GPA is high, so **III is correct** as proved in the equation

80+1IQ+0.036*IQ =90+1IQ+0.04*IQ

- *Predict the salary of a female with IQ of 110 and a GPA of 4.0.*

$\hat{y}$ = 80+1IQ+0.01GPA*IQ

$\hat{y}$ = 80+1*110+0.01*4*110

$\hat{y}$ = 194.4

The starting salary of female is **$194400**

- *True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.*

False. As 0.01 is the coefficient for GPA/IQ which is very small, to test the model we have to assume $\beta 4=0$ and look at the p-value to determine the statistical conclusion.

2 Classification

a) Prove that equation 1 & 2 are equivalent

p(X) = eω0 + ω1X / 1 + eω0 + ω1X (1)

p(X) / 1- p(X)= eω0 + ω1X (2)

Substituting p(X) value in equation (2)

$p(X) / 1- p(X)= [e^{\omega 0 + \omega 1X} / 1 + e^{\omega 0 + \omega 1X}] / [1- (e^{\omega 0 + \omega 1X} / 1 + e^{\omega 0 + \omega 1X})]$

$= [e^{\omega 0 + \omega 1X} / 1 + e^{\omega 0 + \omega 1X}] /[(1 + e^{\omega 0 + \omega 1X} /1 + e^{\omega 0 + \omega 1X})- (e^{\omega 0 + \omega 1X} / 1 + e^{\omega 0 + \omega 1X})]$

$= [e^{\omega 0 + \omega 1X} / 1 + e^{\omega 0 + \omega 1X}] / 1 + e^{\omega 0 + \omega 1X}$

$= e^{\omega 0 + \omega 1X}$

$p(X) / 1- p(X)= e^{\omega 0 + \omega 1X}$

From the above derivation, we can say that equation (1) & (2) are equivalent

b) *In this question, examine the difference between LDA and QDA.*

- *If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?*

If the Bayes decision boundary is linear, we would expect QDA to perform better on training data as it is more flexible. On the other hand, we would expect LDA to perform better on test data as QDA suffer over fitting on data due to high variance without a decrease in bias.

- *If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?*

If the Bayes decision boundary is non-linear, QDA would be accurate as we expect a decrease in bias**.**

- *In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?*

  The test prediction accuracy of QDA is to be expected to improve, this is because QDA is more flexible, as the sample data increases the variance is considered less.

- *True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.*

  False. With less sample data points and many predictors, the variance from utilizing a progressively adaptable strategy, for example, QDA, would prompt over fit, yielding a higher test rate than LDA.

c) *Given –*
x1 = hours studied, x2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, β0=−6,β1=0.1,β2=2.

- *Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.8 gets an A in the class.*

P(x) = exp [β0+ β1x1+ β2x2] / [1+ exp(β0+ β1x1+ β2x2)]
Substituting the values, β0=−6,β1=0.1,β2=2.
P(x) = exp [-6+0.1(hours studied)+2(UG GPA)] / [1+ exp(-6+0.1(hours studied)+2(UG GPA))]
Given, hours studied(x1) = 40 and GPA(x2) = 3.8
P(x) = exp [-6+0.1*40+2*3.8] / [1+ exp [-6+0.1*40+2*3.8]]
P(x) = exp (5.6) / [1+ exp (5.6)]
P(x) = 270.426 / 271.426
P(x) = 0.99631
The probability is **99.631%**

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

x1= hours studied? , x2=3.8 GPA, P(x) = 0.5
Substituting the above values in below equation
P(x) = exp (β0+ β1x1+ β2x2) / (1+ exp(β0+ β1x1+ β2x2))
0.5 = exp(-6+(0.1*x1)+(2*3.8)) / (1+ exp (-6+(0.1*x1)+(2*3.8)))
0.5 = exp (-6+7.6+(0.1*x1)) / (1+ exp(-6+7.6+(0.1*x1)))
0.5 = exp(1.6 + (0.1*x1)) / (1 + exp (1.6 + (0.1*x1))))
0.5(1 + exp (1.6 + (0.1*x1)))) = exp (1.6 + (0.1*x1))
0.5 + 0.5* exp (1.6 + (0.1*x1))) = exp (1.6 + (0.1*x1))
0.5 = exp (1.6 + (0.1*x1)) - 0.5* exp (1.6 + (0.1*x1)))
0.5 = exp (1.6 + (0.1*x1)) *(1-0.5)
0.5 = 0.5* exp (1.6 + (0.1*x1))
1 = exp (1.6 + (0.1*x1))
Log on both sides
log(1) = 1.6 + (0.1*x1)
0 = 1.6 + (0.1*x1)
-1.6 = 0.1*x1

X1 = **-16 hours**
The value is in Negative number, this means that the study needs to study hard to get an A in the class.

3 Decision Trees

a)

Weather	**Traffic**	**Accident Rate**	**No. of Observations**
Sunny	Heavy	High	23
Sunny	Light	Low	5
Rainy	Heavy	High	50
Rainy	Light	Low	22

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Entropy for class [73,27] = -[73/100 $\log_2$(73/100) + 27/100 $\log_2$ (27/100)]
= 0.8414bits

Weather

Entropy for Sunny [23, 5] = [23/28 $\log_2$(23/28) + 5/28 $\log_2$(5/28)]
= 0.6768bits

Entropy for Rainy [50, 22] = [50/72 $\log_2$(50/72) + 22/72 $\log_2$(22/72)]
= 0.8879bits

Traffic

Entropy for Heavy [73, 0] = 0bits
Entropy for Light [0, 27] = 0bits

The average weighted entropy for weather
= (28/100) * (0.6768) + (72/100)*(0.8879)
= 0.8287 bits

Information Gain for Weather = 0.8414 – 0.8287
= 0.0127 bits

Information Gain for Traffic = 0.8414 – 0
= 0 bits

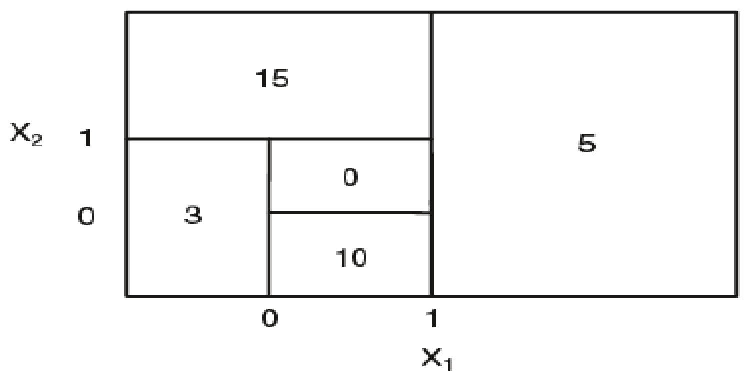
Here, we can see that "Traffic" has less entropy and it provides the best prediction of "Accident rate" so the split is done using Traffic.

b) The second student (T2) constructs decision tree by first performing normalization on the values and then constructs the tree which will result in a graph and will not have any effect on the output class values or the probability of fetching the features as the decision trees are partitioned based on the observation along each axis.

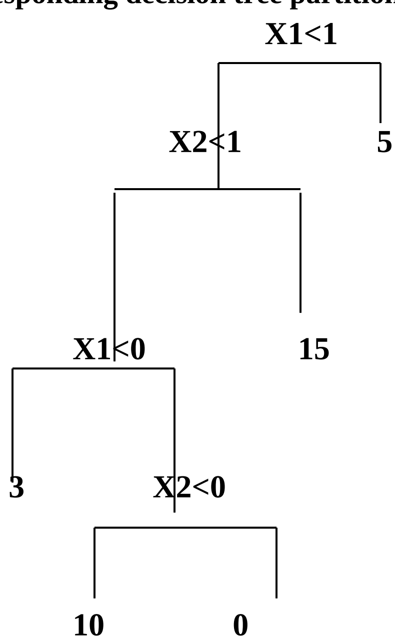
The feature values of the first student (T1) can be altered by considering every decision boundary and then subtracting the mean and by dividing the variance for the equivalent feature value. By doing this we get the same values as that of second student with no difference in probability.
We can say that, there is **no difference** between both the decision tree structures as the probability value will remain the same in T1 & T2.

c)

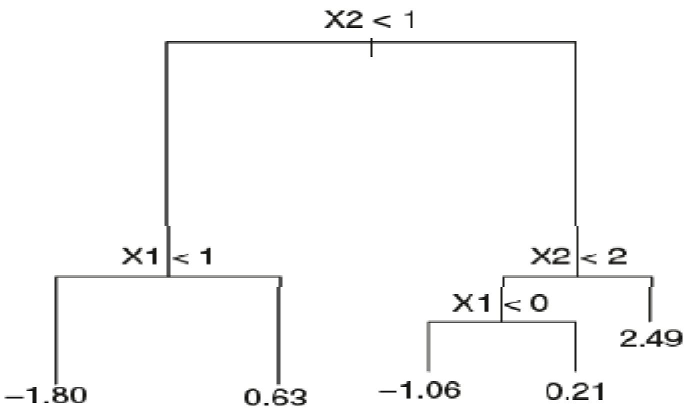
-



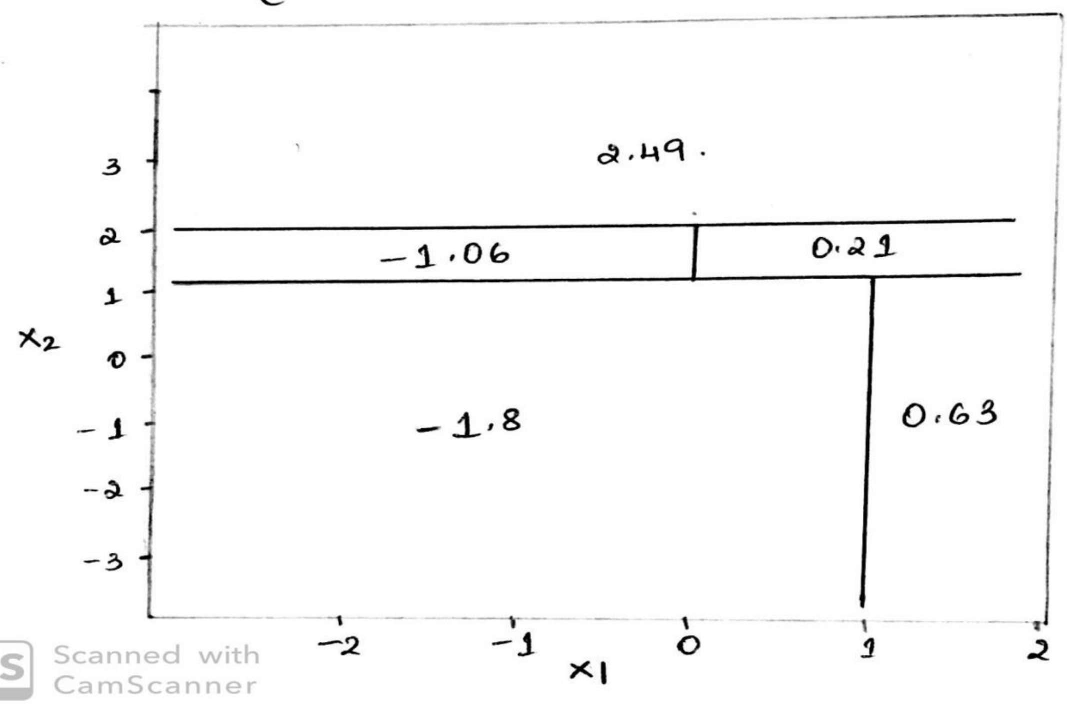
The corresponding decision tree partition for the predictor space is as follows-



- **A diagram corresponding to the tree below -**



3 (c) • A diagram corresponding to the tree -



References-

ISLR book &

https://stats.stackexchange.com/questions/71489/three-versions-of-discriminant-analysis-differences-and-how-to-use-them