DPA ASSIGNMENT-2

Chapter 3

Part 1

1) If the p value is less than α a then we reject the null hypothesis otherwise we can hold them true. α is a constant value where the value of it is decided before. Accordingly, we reject or accept the null hypothesis.
   Based on the p-values in the table, we can conclude whether we reject or accept null hypothesis.
   - The p-value for TV is <0.0001, so we can reject null hypothesis and therefore we can say that TV will be impacting the sales when others (radio and newspaper) are constant.
   - The p-value for radio is <0.0001, so we can reject null hypothesis and therefore we can say that radio will be impacting the sales when others (TV and newspaper) are constant.
   - The p-value for newspaper is 0.8599, so we can accept null hypothesis and therefore we can say that newspaper will not be impacting the sales when others (radio and TV) are constant.

3)a)Least square equation is given by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$

$\hat{y}$=50 + 20*GPA + 0.07*IQ + 35*GENDER + 0.01*GPA*IQ – 10 *GPA * GENDER

For FEMALE,

$\hat{y}$=50 + 20*GPA + 0.07*IQ + 35*1 + 0.01*GPA*IQ – 10 *GPA*1

$\hat{y}$ = 85 + 10*GPA + 0.07*IQ + 0.01*GPA*IQ

For MALE,

$\hat{y}$=50 + 20*GPA + 0.07*IQ + 35*0 + 0.01*GPA*IQ – 10 *GPA*0

$\hat{y}$ = 50 + 20*GPA + 0.07*IQ +0.01*GPA*IQ

Solving both the above equations, we get

85+10*GPA = 50 + 20*GPA

When we solve this equation, we get the GPA value as 3.5, then males earning is equal to females earning.

If we have higher value of GPA, then male earn more on average as compared to females. So, Option 3 is the correct answer.

b) Least square model for females is given by

$\hat{y}$ = 85 + 10*GPA + 0.07*IQ + 0.01*GPA*IQ

Here, IQ=110, GPA=4.0

So, substituting the above values in the equation we get
85+40+(0.07*110) +(0.01*4*110) =137.1=$137100

So, the salary of female is $137100.

c)False, we need p value to decide whether interaction is significant or not. We cannot predict based on the coefficient of GPA/IQ. We need hypothesis testing to get the p-value.

4)a)The training RSS for cubic regression will be lower than that of linear regression because when the polynomial degree increases then we have more number of data points for plotting which appear to be very closely fitted and accurately. This results in less training errors.

b) In the testing data, the RSS for cubic regression is larger as compared to linear regression as we have a higher degree polynomial which will overfit the training data. This result is more testing errors.

c)According to the question we don't know the relationship between X and Y, whether it is linear or not. Generally, the training RSS is less for higher polynomial degree and it helps is plotting more accurately and closely. This will result in less training error for cubic regression. RSS for cubic regression has less training error.

d)We don't know the true relationship of X and Y for testing data. We don't know exactly how far it is from linear. If the true relationship between X and Y is linear then we will have lower Residual Sum of Squares(RSS) for Linear regression for testing data whereas if the true relationship between X and Y is non linear then Cubic regression will have lower RSS.

Chapter 4

Part 2

4)a)Given that X is in the range of [0,1].We are dividing into 3 different ranges where X ϵ [0.05,0.95] and the fraction of the observations is 10%.If X ϵ [0,0.05] then the faction of observations is (100x+5)%.If X ϵ [0.95,1] then the fraction of the observation is (105-100x)%. To calculate the fraction, we have to make this prediction and calculate with the following expression.

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

=10(0.95-0.05) +50(0.05*0.05) +5(0.05) +105(0.05) +50[(1*1)-(0.95*0.95)]
=9+0.375+0.375=9.75%

To make the prediction, the fraction used for the observation is 9.75%.

b) In the above question, we have calculated the value for X which is 9.75% for single variable. Now we need for 2 features $X_1$ and $X_2$ where they both are independent variables. So, the fraction of observations is 9.75 * 9.75=0.95%.

c)Given that p=100 observations, we assume the same as the above case where all the 100 predictors are independent of each other. So $(9.75)^{100}$ is approximately equal to 0. So, the fraction of observations for 100 features is equal to 0.

d)From the above cases, we can observe that as the number of features increases the fraction of observations is decreasing which is nearly equal to 0. The fraction of observations for n features is $(9.75\%)^{n.}$ Therefore, if there are a greater number of features then there will be a smaller number of neighbors for KNN.

e) As there are 10 observations, volume of hypercube is 10%.
When p=1, length is 0.1=0.1
When p=2, length is $(0.1)^{1/2}$ = 0.316
When p=100, length is $(0.1)^{1/100}$ = 0.977
As the number of features increases, we use all the data to make predictions with the available data.

6)Logit function is given by the following formula:

$$p(x) = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2}}{1 + e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2}}$$

Here $\hat{\beta}_0$= -6, $\hat{\beta}_1$ = 0.05, $\hat{\beta}_2$ = 1, X1= Hours studied, X2 = undergrad GPA.

a)Given, X1=40 hrs, X2=3.5

$$p(x) = \frac{e^{-6+0.05*40+1*3.5}}{1 + e^{-6+0.05*40+1*3.5}}$$

$$= \frac{e^{-0.5}}{1+e^{-0.5}} = 0.3775 = 37.75\%$$

The probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class is 37.75%.

b) Given P(X) = 0.5 , X2= 3.5

$$0.5 = \frac{e^{-6+0.05*X1+1*3.5}}{1+e^{-6+0.05*X1+1*3.5}}$$

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

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

$0.05 * X1 = 2.5$

$X1 = 50$ hrs.

7)According to Bayes Theorem,

$$P_k(X) = \frac{\Pi_K \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}$$

Given, $\Pi_{Yes}$=0.8 , $\Pi_{No} = 0.2$, $\mu_{Yes}$=10, $\mu_{No}$=0,x=4, $\sigma^2$= 36

Substituting all the above values in the equation, we get

$$P_{Yes}(X) = \frac{0.8*\frac{1}{\sqrt{2\pi(36)}} e^{-(\frac{1}{2(36)})(4-10)^2}}{0.8*\frac{1}{\sqrt{2\pi(36)}} e^{-(\frac{1}{2(36)})(4-10)^2} + 0.2*\frac{1}{\sqrt{2\pi(36)}} e^{-(\frac{1}{2(36)})(4-0)^2}}$$

$$P_{Yes}(X) = \frac{0.8*e^{-(\frac{1}{2})^1}}{0.8*e^{-(\frac{1}{2})^1} + 0.2*e^{-(\frac{2}{9})^1}}$$

$P_{Yes}(X)=0.752$

The probability that a company will issue a dividend this year given that its percentage return was X=4 last year is 0.752

9) $a)$ $Odds = \frac{p(x)}{1-p(x)}$

Given odds=0.37

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

0.37*(1-p(x))=p(x)

0.37 - 0.37*p(x)=p(x)

P(x)=0.27=27%

Therefore, an average of 27% people defaulting on their credit card payment.

b) Given p(x) = 0.16

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

$Odds=\frac{0.16}{0.84}$

Odds=0.19

The odds that she will default is 19%.