EAS 506 Statistical Data Mining — Jaideep Reddy Kommera

Homework 3 — Class no. 28

Question 1:

<u>Introduction</u>: The Boston dataset has the housing values in the suburbs of Boston from the 1970's. The dataset has 506 rows and 14 columns. To predict whether a given suburb has a crime rate above or below the median.

<u>Pre-processing</u>: The target variable crime rate is initially numerical. The median is found and the values above the median are labelled as 1 and the values below the median are labelled as 0.

```
> str(mydata$crim_med)
 Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2 2 ...
```

<u>Train-Test Split:</u> The dataset has been split into train and test. 75% of the data has been allotted to training and 25% has been allotted to testing. set.seed() has been used before splitting data.

<u>Linear Discriminant Analysis:</u>

Prior Probabilities - Class '0'= 0.2526316

Class '1'= 0.7473684

<u>Coefficients of Linear Discriminants:</u>

If you multiply each coefficient by the corresponding elements of the predictor variables and sum them, we get a score for each respondent. This score along with the prior probabilities are used to compute the posterior probability of class membership. Classification is made based on the posterior probability, with observations predicted to be in the class for which they have the highest probability.

```
Coefficients of linear discriminants:
                  LD1
zn       -0.0369470126
indus     0.0180766574
chas      0.3302545159
nox       0.2490025567
rm       -0.0902813477
age       0.0061455891
dis       0.0500265794
rad      -0.0081234017
tax       0.0028390713
ptratio  -0.0801720799
black    -0.0001783738
lstat     0.0489327489
medv      0.0226883174
```

Train Error:

```
> train_err_lda
[1] 0.1631579
```

Test Error:

```
> test_err_lda
[1] 0.1825397
```

Logistic Regression:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.512084  11.535907   0.911 0.362164
zn          -0.040046   0.012923  -3.099 0.001943 **
indus        0.023953   0.061173   0.392 0.695387
chas         0.780537   1.091324   0.715 0.474473
nox          5.932820   5.169862   1.148 0.251143
rm          -0.220491   0.820032  -0.269 0.788021
age         -0.010723   0.013461  -0.797 0.425680
dis         -0.034214   0.179067  -0.191 0.848472
rad          0.505179   0.130325   3.876 0.000106 ***
tax          0.009607   0.003924   2.448 0.014349 *
ptratio      0.003942   0.125188   0.031 0.974881
black       -0.052305   0.024002  -2.179 0.029318 *
lstat        0.285837   0.095237   3.001 0.002688 **
medv         0.097701   0.069241   1.411 0.158235
```

We can see from the above summary that ‘zn’, ‘rad’, ‘tax’, ‘black’ and ‘lstat’ are the significant variables in this classification problem.

Train Error:

```
> train_err_log
[1] 0.1078947
```

Test Error:

```
> test_err_log
[1] 0.1507937
```

K-Nearest Neighbor:

- K = 3

Test Error:

```
> test_err_knn
[1] 0.1111111
```

Similarly,

Subset 1: Chosen based on EDA done during homework 1 on the Boston Dataset

- Zn- proportion of residential land zoned
- Nox- nitric oxides concentration
- Dis- weighted distances to five Boston employment centers
- Rad- index of accessibility to radial highways
- Ptratio- pupil-teacher ratio by town
- Black- proportion of blacks by town
- Lstat- % lower status of the population
- Medv- Median value of owner-occupied homes in $1000's

Train Error LDA:

```
> train_err_lda_set1
[1] 0.1578947
```

Test Error LDA:

```
> test_err_lda_set1
[1] 0.1746032
```

Train Error Logistic:

```
> train_err_log_set1
[1] 0.1184211
```

Test Error Logistic:

```
> test_err_log_set1
[1] 0.1666667
```

Test Error K-NN:

```
> test_err_knn_set1
[1] 0.1507937
```

Subset 2: Chosen based on significant variables shown by Logistic Regression

- Zn- proportion of residential land zoned
- Rad- index of accessibility to radial highways
- tax- full-value property-tax rate per $10,000
- Black- proportion of blacks by town
- Lstat- % lower status of the population

Train Error LDA:

```
> train_err_lda_set2
[1] 0.1605263
```

Test Error LDA:

```
> test_err_lda_set2
[1] 0.1666667
```

Train Error Logistic:

```
> train_err_log_set2
[1] 0.1131579
```

Test Error Logistic:

```
> test_err_log_set2
[1] 0.1269841
```

Test Error K-NN:

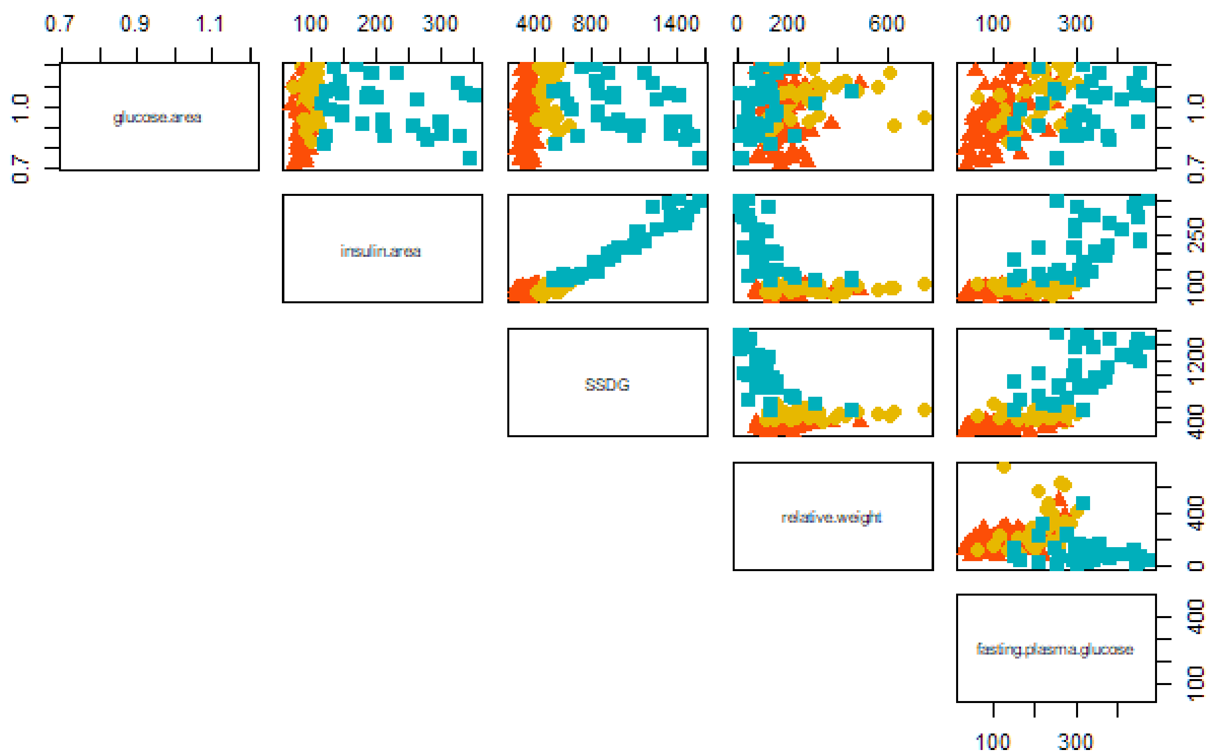
```
> test_err_knn_set2
[1] 0.1111111
```

All three models did not show much significant difference in their train or test errors. The subset taken from significant variables of Logistic model performed better than the full dataset on train and test. KNN remained the same because it is non-parametric.

Question 2:

Introduction: The dataset is on Diabetes. The dataset has 145 observations and 6 variables.

Pair-wise Scatterplot:

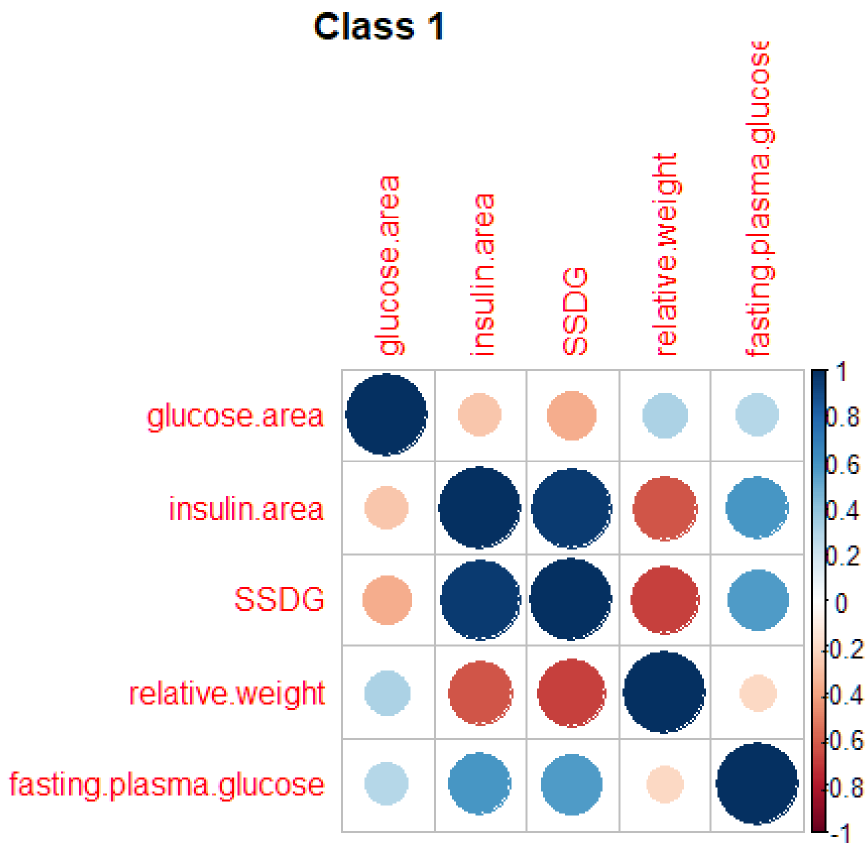


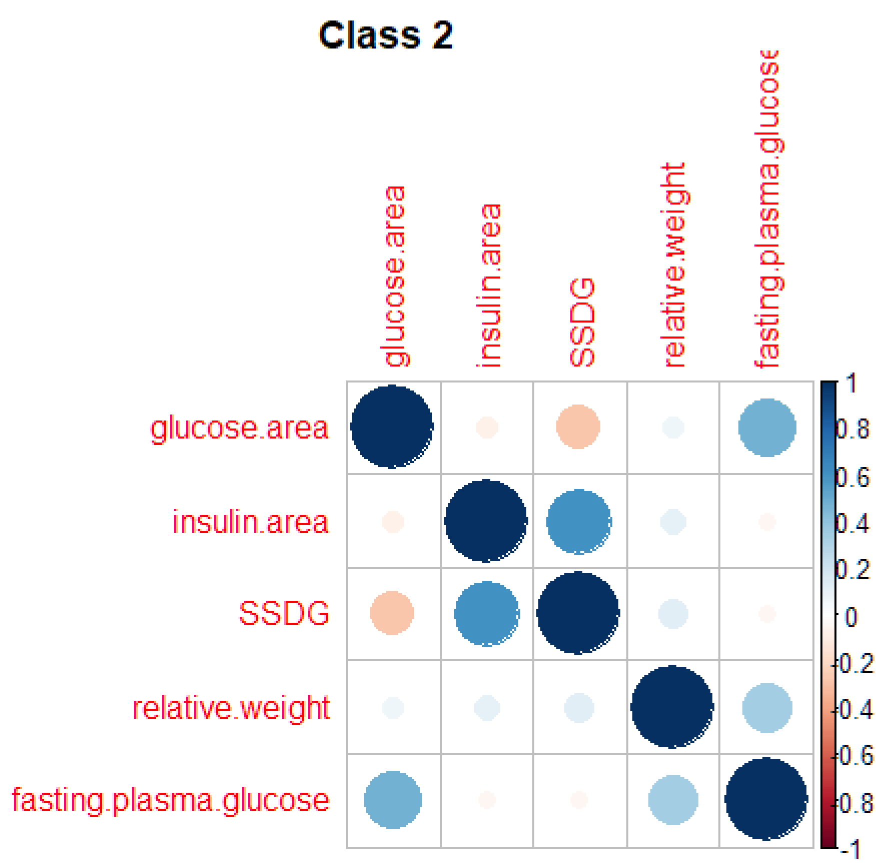
Blue: Class 1

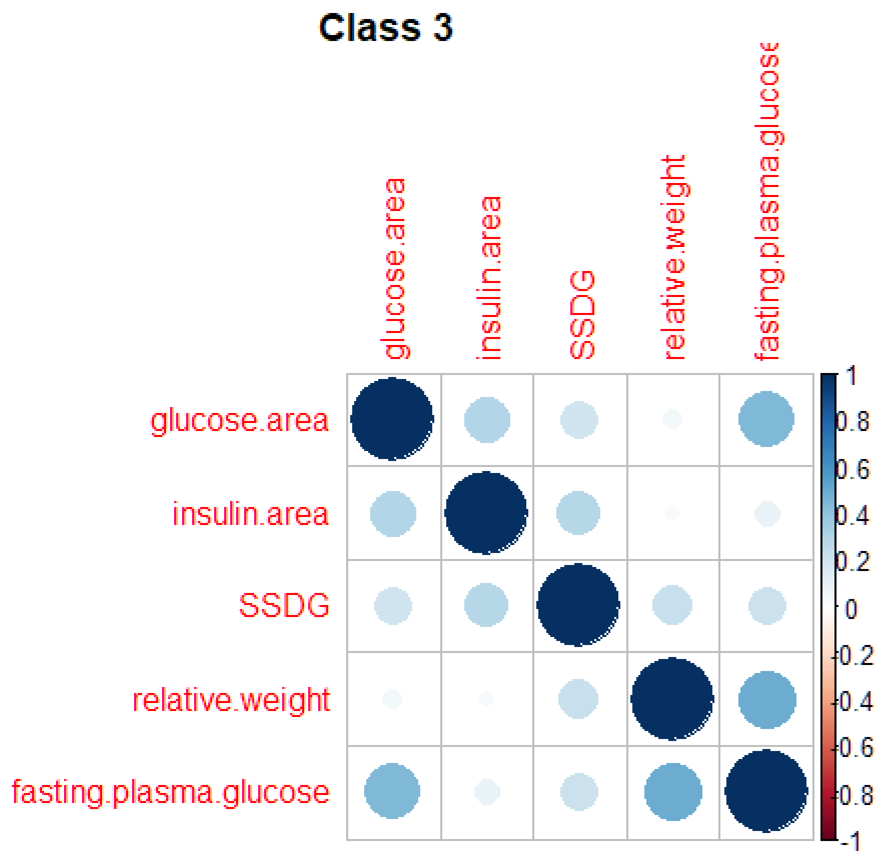
Yellow: Class 2

Red: Class 3

Covariance/ Correlation Matrices:





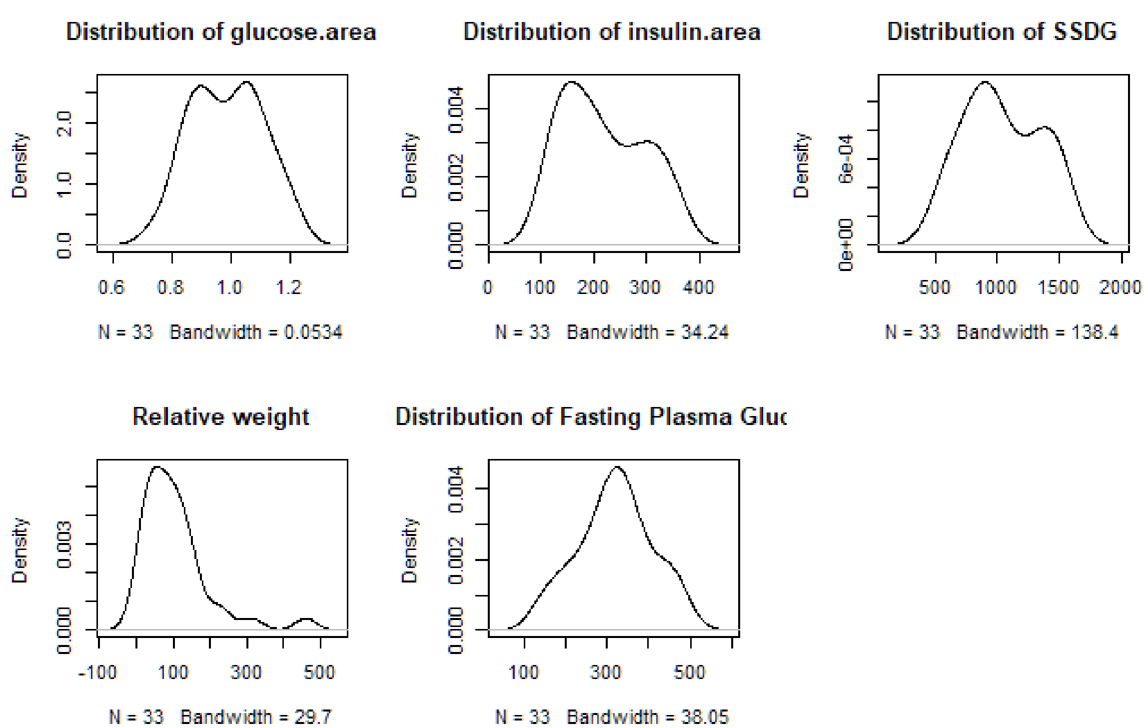


From the above correlation plots we can infer that the classes have **different** covariance matrices.

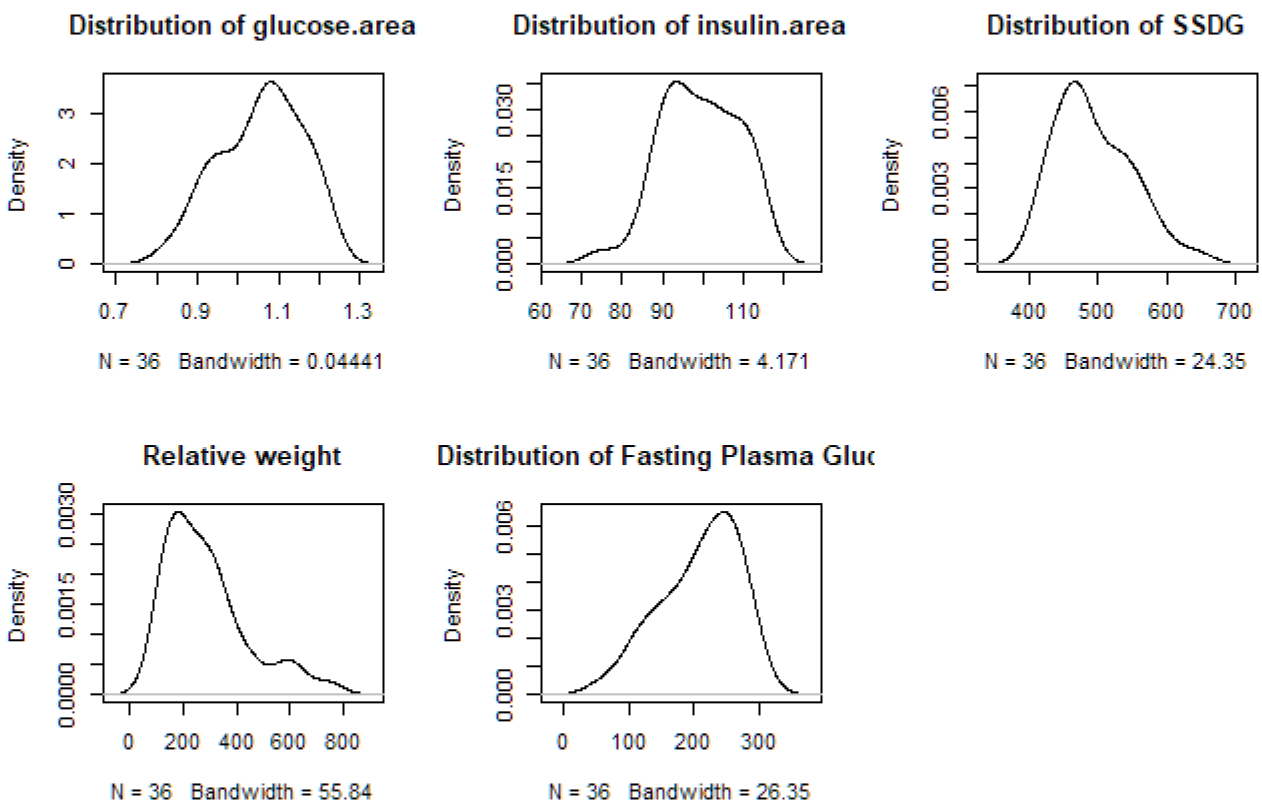
Multivariate normal:

Distribution of variables by class

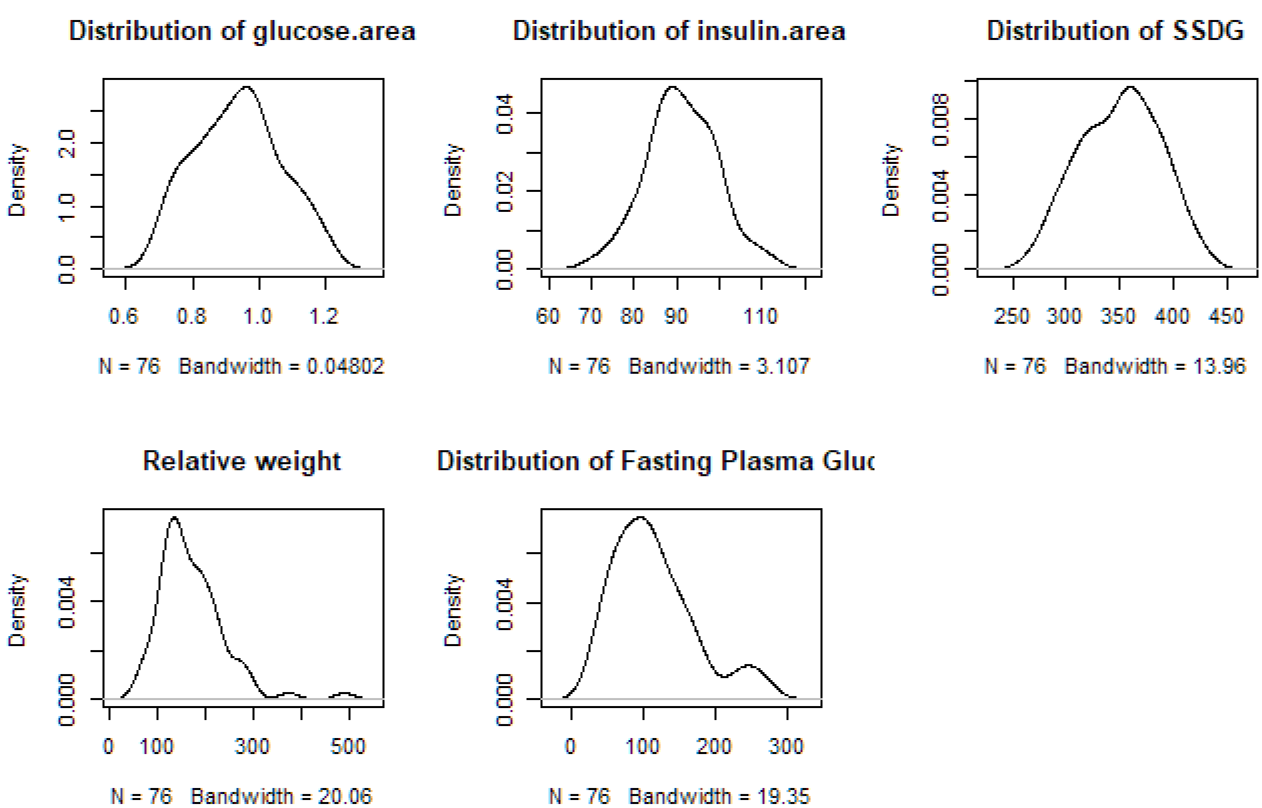
Class 1:



Class 2:



Class 3:



ANALYSIS:

From the above plots we can see that for each class the predictors in that class follow multivariate normal distribution.

Part b)

Train and Test split : 75% and 25%

LDA:

Train error:

```
> train_err_lda
[1] 0.1192661
```

Test error:

```
> test_err_lda
[1] 0.08333333
```

QDA:

Train error:

```
> train_err_qda
[1] 0.06422018
```

Test error:

```
> test_err_qda
[1] 0.05555556
```

ANALYSIS: QDA performs better than LDA. This due to the covariance matrices **not** being same for all the classes.

Part c)

Classification of the new sample.

LDA:

Class no: **3**

QDA:

Class no: **2**

Question 3:

Part a)

Assuming:

$$\sum_{k=1}^{k} P_r\,(G = k \mid X = x) = 1$$

$$P_r(G = k \mid X = x) = \frac{e^{(\beta_{k0} + \beta_k^T x)}}{1 + \sum_{l=1}^{k-1} e^{(\beta_{l0} + \beta_l^T x)}} \quad for\ k = 1,2 \ldots\ldots k-1$$

$$P_r(G = K \mid X = x) = \frac{1}{1 + \sum_{K=1}^{K-1} e^{(\beta_{l0} + \beta_l^T x)}}$$

$$\log \frac{P_r(G = k \mid X = x)}{P_r(G = K \mid X = x)} = \beta_{k0} + \beta_k^T x$$

$$e^{(\beta_{k0} + \beta_k^T x)} = \frac{P_r(G = k \mid X = x)}{P_r(G = K \mid X = x)}$$

$$P_r(G = k \mid X = x) = \frac{\frac{P_r(G = k \mid X = x)}{P_r(G = K \mid X = x)}}{\frac{1 + \sum_{l=1}^{k-1} P_r(G = l \mid X = x)}{P_r(G = K \mid X = x)}}$$

$$P_r(G = k \mid X = x) = \frac{P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

$$P_r(G = k \mid X = x) = \frac{P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

$$\sum_{k=1}^{k} P_r(G = k \mid X = x) = \sum_{k=1}^{k-1} P_r(G = k \mid X = x) + P_r(G = K \mid X = x)$$

$$\sum_{k=1}^{k} P_r(G = k \mid X = x) = \frac{\sum_{k=1}^{k-1} P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)} + \frac{P_r(G = K \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)}$$

$$\frac{\sum_{k=1}^{k} P_r(G = k \mid X = x)}{\sum_{k=1}^{k} P_r(G = k \mid X = x)} = 1$$

Part b)

Given logistic function $p(X)$:

$$1 - p(X) = 1 - \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} = \frac{1}{1 + e^{(\beta_0 + \beta_1 X)}}$$

$$\frac{1}{1-p(X)} = 1 + e^{(\beta_0 + \beta_1 X)}$$

$$p(X) * \frac{1}{1-p(X)} = \frac{e^{(\beta_0 + \beta_1 X)}}{1+ e^{(\beta_0 + \beta_1 X)}}\left(1 + e^{(\beta_0 + \beta_1 X)}\right),$$

$$\frac{p(X)}{1-P(X)} = e^{(\beta_0 + \beta_1 X)}$$

Therefore, the Logistic Representation and Logit Representation of Logistic Regression model are equivalent