Name: Rohit Kulkarni

USC ID: 5402749044

1.c.i) Time-domain features:

- Minimum
- Maximum
- Mean
- Standard Deviation
- Median
- Variance
- Slope

1.c.ii) Dataset with all the features

	Instance	Min1	Max1	Mean1	Median1	SD1	1st Quartile1	3rd Quartile1	Min2	Max2	...	SD5	1st Quartile5	3rd Quartile5	Min6	Max6	Mean6	Median6	S
0	1	37.25	45.00	40.624792	40.50	1.476967	39.2500	42.00	0.0	1.30	...	2.188449	33.00	36.00	0.0	1.92	0.570583	0.43	0.5829
1	2	38.00	45.67	42.812812	42.50	1.435550	42.0000	43.67	0.0	1.22	...	1.995255	32.00	34.50	0.0	3.11	0.571083	0.43	0.6010
2	3	12.75	51.00	24.562958	24.25	3.737514	23.1875	26.50	0.0	6.87	...	3.693786	20.50	27.00	0.0	4.97	0.700188	0.50	0.6937
3	4	0.00	42.75	27.464604	28.00	3.583582	25.5000	30.00	0.0	7.76	...	5.053642	15.00	20.75	0.0	6.76	1.122125	0.83	1.0123
4	5	24.25	45.00	37.177042	36.25	3.581301	34.5000	40.25	0.0	8.58	...	2.890347	17.95	21.75	0.0	9.34	2.921729	2.50	1.8526

5 rows × 43 columns

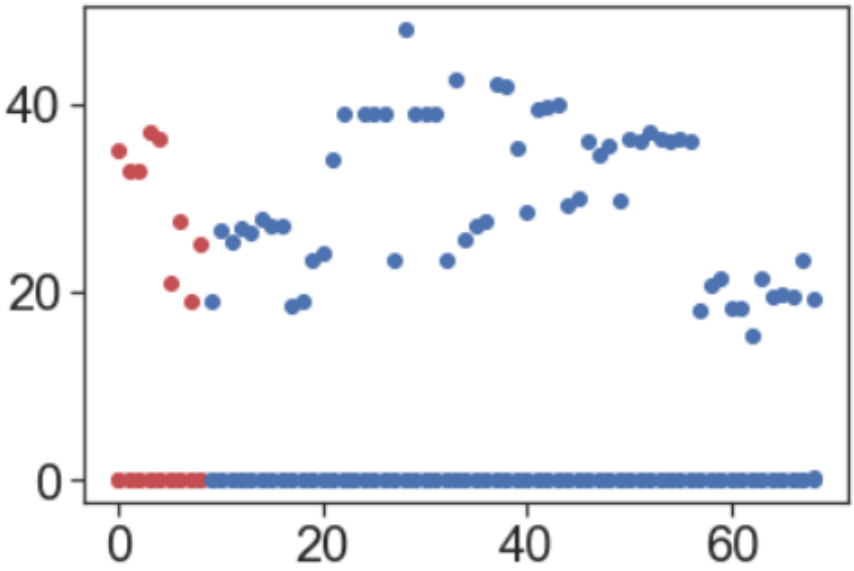
1.c.iii) 90% bootstrap confidence intervals for each feature

- Min: (0.3599570269645862, 6.92345461326823)
- Max: (0.13060205021809712, 2.512014766623053)
- Mean: (0.19051901693737094, 3.6644645552652064)
- Median: (0.1994058194500965, 3.835394330890862)
- Standard Deviation: (0.04257369199504635, 0.8188672596075304)
- 1st Quartile: (0.23785213370672764, 4.57487513516315)
- 3rd Quartile: (0.16490575735632507, 3.171816191924113)

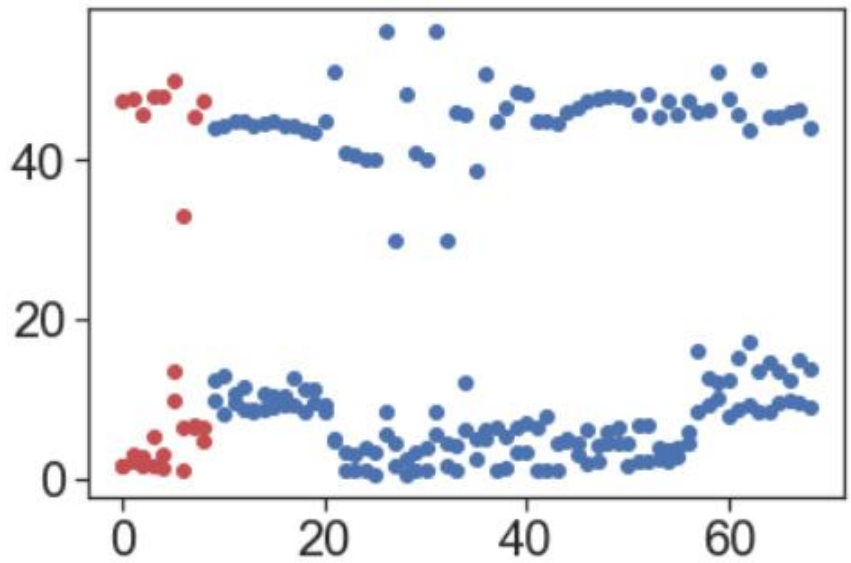
1.c.iv) Three most important time-domain features selected in this assignment are min, max and mean. Because they are most widely used in all the models and it gives better results compared to other feature.

1.d.i)

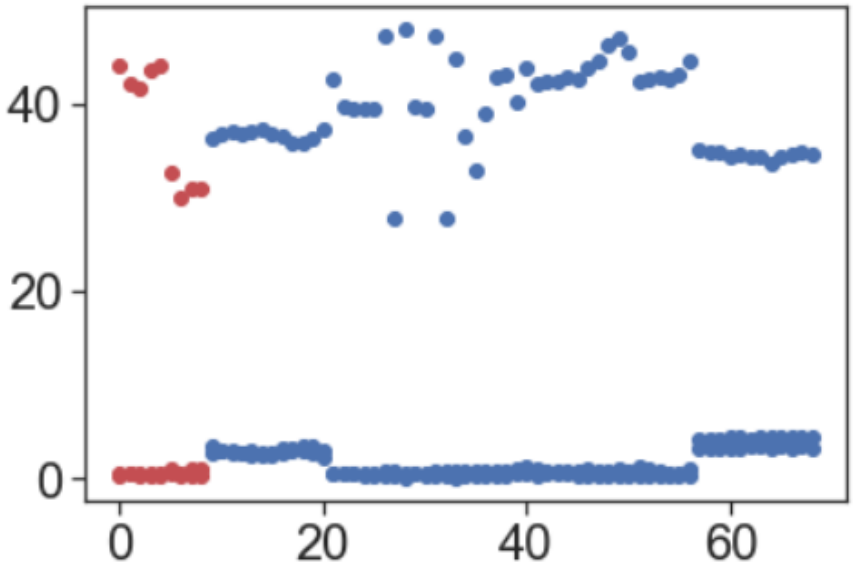
Scatter Plot for min



Scatter Plot for max

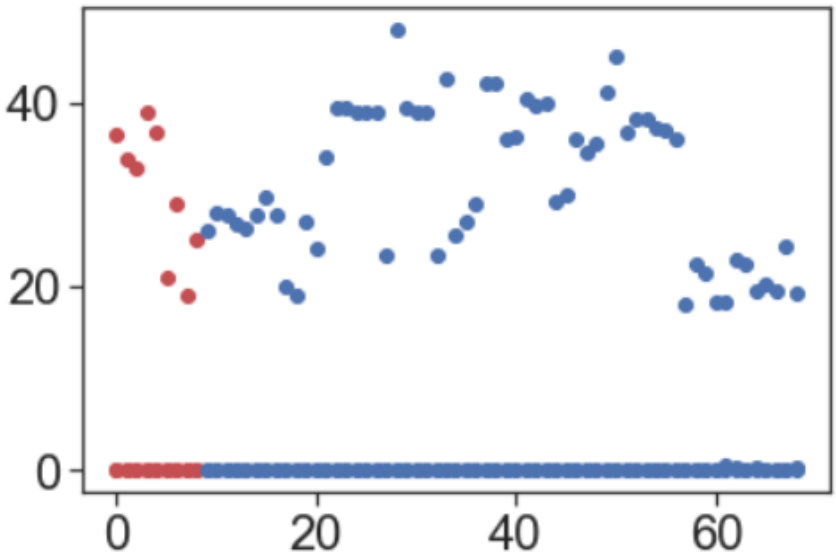


Scatter Plot for mean

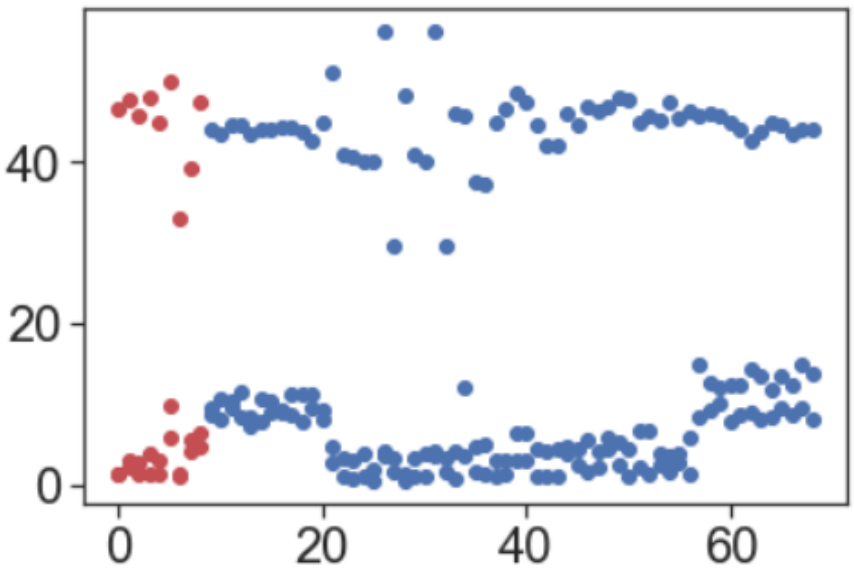


1.d.ii) The scatter plots are same as the previous question, there is no considerable between the scatter plots with the above ones.

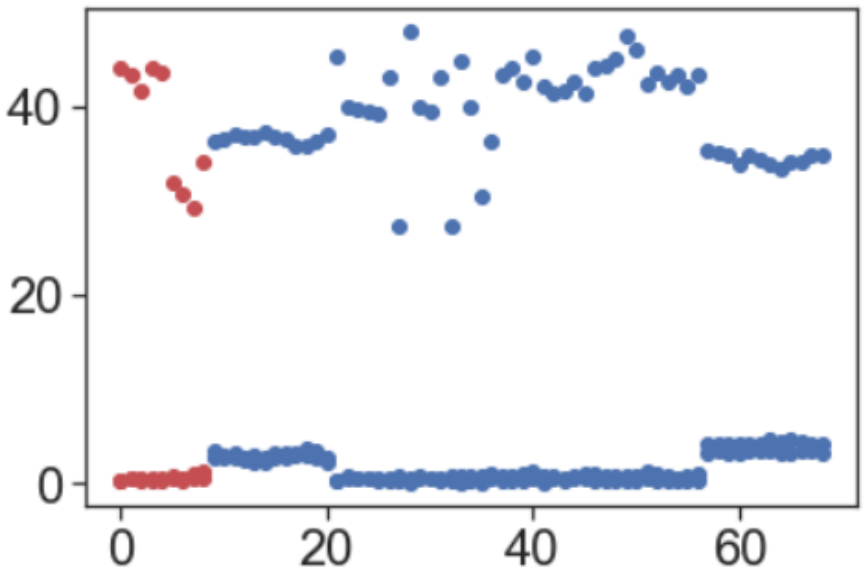
Scatter Plot for min



Scatter Plot for max



Scatter Plot for mean



1.d.iii) After splitting the data and running the RFE model we got the below scores for each value of L, Then we chose L=1 based on the best score and fit the logistic regression model.

```
[False False False False  True False False  True False False  True False
  True False False False  True False False False False False False False
 False False False False False False False False False False  True False
  True False False False False False]
[ 9 36 22 35  1 34 33  1  6 25  1  3  1 15 18 16  1 11 30 27  2 13 12 14
 21 20 24 32 31 26 19 28 29 23  1 17  1  4  7  8 10  5]
P value scores
0.9857142857142858
[False False False False False False False False  True False  True False
 False False False False False False False False False  True False False
 False False False False  True False False False False False False False
  True False  True  True False False]
[33 35 17 36  4 34  8  7  1 18  1 14  6  2 10 11 12 28 22 30 24  1 21 15
  9  5 29 20  1 26 27 32 23 16 13 19  1  3  1  1 31 25]
P value scores
0.8997354497354497
[False False False False False False False False False False  True False
 False False False False False False False False False  True False False
  True  True False False  True False False False False  True False False
 False  True False False False False]
[25 34 33 36 13 32 12 35  3 23  1 11  4  7  6 19 16 29 14 24  5  1 26 31
  1  1 17  8  1 20 18 22 30  1 10 28 27  1  2 15  9 21]
P value scores
0.8687572590011614
[False False False False  True False False False False  True False False
 False  True False False False False False False False False False False
 False False False False False False False  True False False False  True
 False  True False  True False False]
[25 32  2 33  1 36 22  7  3  1  8 31 17  1 13 15 29 34 26 12  6 23 11 20
 21 16  4 28  9 24  5  1 27 19 30  1 18  1 14  1 10 35]
P value scores
0.890909090909091
[False False False False False False False False False  True False False
 False  True False False False False False False False  True False False
 False False False False  True False False  True False  True False False
 False False False False  True False]
[29 11 10 35 22 28 32 13  3  1  9 18 31  1  2  8 30 17 27 36  5  1 21 23
 14 19 34 12  1 25 16  1 33  1 20 26  6 24  4  7  1 15]
P value scores
```

Cross validation is one of the most popular methods for estimating test error and selecting tuning parameters, however, one can easily be mislead/confused by it without realizing it.

The Wrong Way

A model takes the dataset and finds that there are moderate correlations between some xs and y, and that the top 5 variables are not correlated. In this model noise can be fitted as well resulting in higher value in error e.

The Right Way

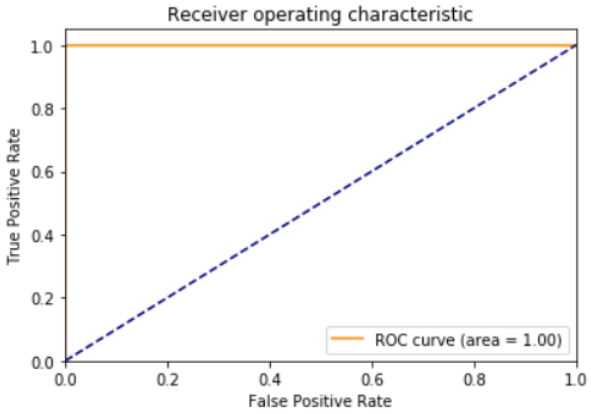
The right way to do cross validation is to include the feature screening process in the cross validation as well.

Hence we have used feature to do the cross validation in this assignment.

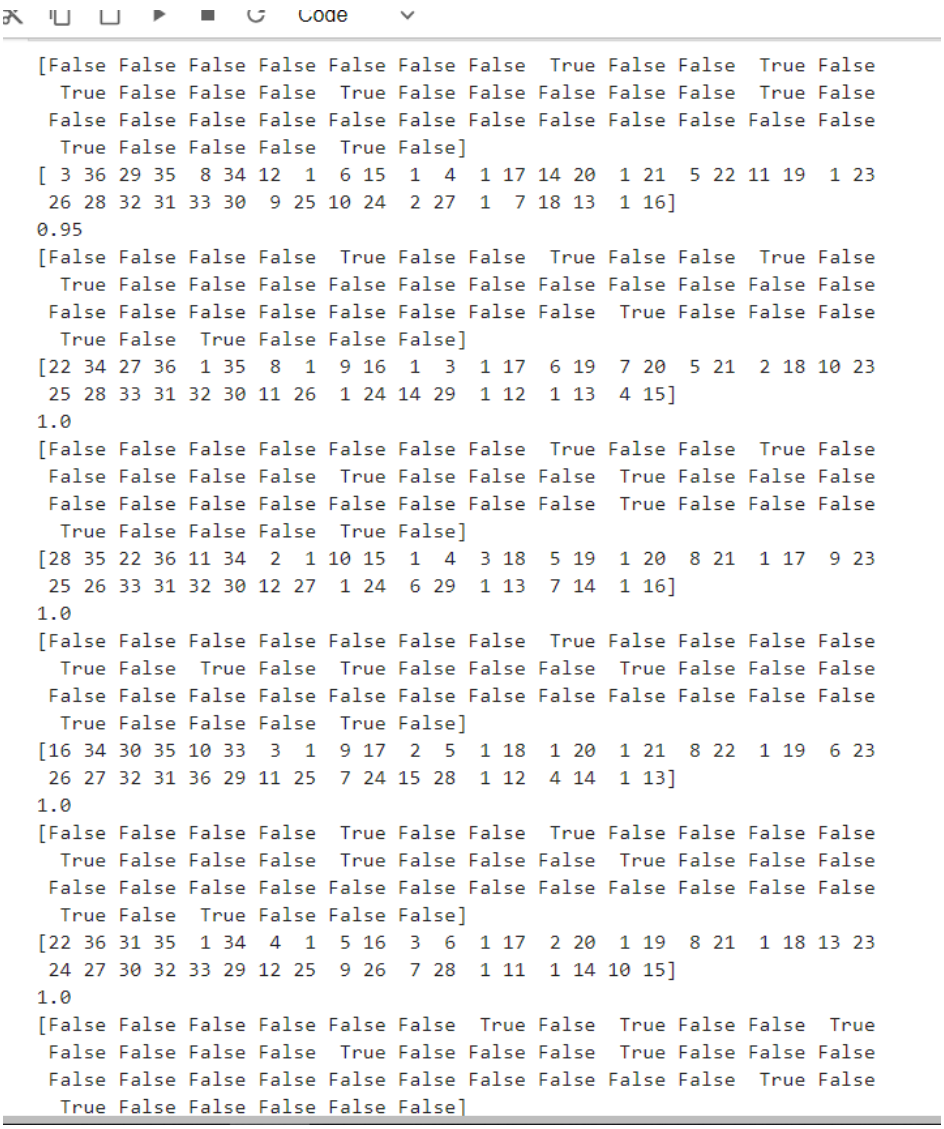
1.d.iv)

the confusion matrix with AUC and ROC curve

col_0	0	1
row_0		
0	1	0
1	0	11



1.d.v) Same as above is done for the test set, splitting the data and finding out the best L value which is 2



```
[False False False False False False False  True False False  True False
  True False False False  True False False False False False  True False
 False False False False False False False False False False False False
  True False False False  True False]
[ 3 36 29 35  8 34 12  1  6 15  1  4  1 17 14 20  1 21  5 22 11 19  1 23
 26 28 32 31 33 30  9 25 10 24  2 27  1  7 18 13  1 16]
0.95
[False False False False  True False False  True False False  True False
  True False False False False False False False False False False False
 False False False False False False False False  True False False False
  True False  True False False False]
[22 34 27 36  1 35  8  1  9 16  1  3  1 17  6 19  7 20  5 21  2 18 10 23
 25 28 33 31 32 30 11 26  1 24 14 29  1 12  1 13  4 15]
1.0
[False False False False False False False  True False False  True False
 False False False False  True False False False  True False False False
 False False False False False False False False  True False False False
  True False False False  True False]
[28 35 22 36 11 34  2  1 10 15  1  4  3 18  5 19  1 20  8 21  1 17  9 23
 25 26 33 31 32 30 12 27  1 24  6 29  1 13  7 14  1 16]
1.0
[False False False False False False False  True False False False False
  True False  True False  True False False False  True False False False
 False False False False False False False False False False False False
  True False False False  True False]
[16 34 30 35 10 33  3  1  9 17  2  5  1 18  1 20  1 21  8 22  1 19  6 23
 26 27 32 31 36 29 11 25  7 24 15 28  1 12  4 14  1 13]
1.0
[False False False False  True False False  True False False False False
  True False False False  True False False False  True False False False
 False False False False False False False False False False False False
  True False  True False False False]
[22 36 31 35  1 34  4  1  5 16  3  6  1 17  2 20  1 19  8 21  1 18 13 23
 24 27 30 32 33 29 12 25  9 26  7 28  1 11  1 14 10 15]
1.0
[False False False False False False  True False  True False False  True
 False False False False  True False False False  True False False False
 False False False False False False False False False False  True False
  True False False False False False]
```

Splitting the test data into 2 and fitting the logistic regression model and displaying the confusion matrix.

col_0	0	1
row_0		
0	1	0
1	0	11

1.d.vi) No the classes did not cause any instability in calculating logistic regression as the classes were able to fit as expected.

1.d.vii) As you can see from the confusion matrix the values are the same hence there are no imbalanced classes.

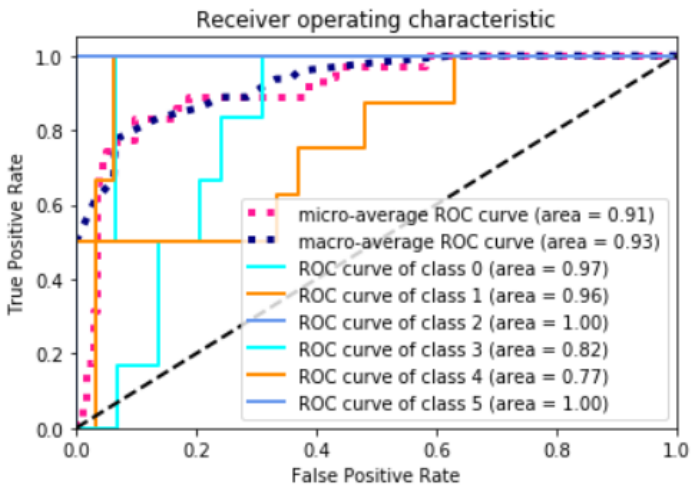
1.e.i) Now we create a model with L1-penalized logistic regression which gives out the below results

```
L1-penalized scores
0.9714285714285715
L1-penalized scores
0.9137566137566138
L1-penalized scores
0.8929152148664343
L1-penalized scores
0.9164935064935065
L1-penalized scores
0.8840579710144928
L1-penalized scores
0.9010578900969733
L1-penalized scores
0.9028994845360824
L1-penalized scores
0.9092055692055692
L1-penalized scores
0.9080645161290324
L1-penalized scores
0.9028985507246376
L1-penalized scores
0.9038602300453119
L1-penalized scores
0.9083242059145673
L1-penalized scores
0.8961576660459342
L1-penalized scores
0.8984455958549222
L1-penalized scores
0.8946859903381641
L1-penalized scores
0.8858988070752776
L1-penalized scores
0.8977777777777778
L1-penalized scores
0.8960195621194454
L1-penalized scores
0.8923722172234638
L1-penalized scores
```

1.e.ii) From the scores of both models as shown below, we can see that the score is better for variable selection using p-values model.

P value scores	L1-penalized scores
0.9857142857142858	0.9714285714285715

1.f.i) Best L value from the L1-penalized multinomial regression model is 1. ROC curves and confusion matrices for multiclass classification




```
col_0   0  1
row_0
0      31  0
1       2  2
col_0   0  1
row_0
0      30  2
1       0  3
col_0   0  1
row_0
0      27  1
1       0  7
col_0   0   1
row_0
0      19  10
1       0   6
col_0   0  1
row_0
0      27  0
1       6  2
col_0   0  1
row_0
0      28  0
1       0  7
```

1.f.ii) Naive Bayes' classifier using Gaussian

```
0.9855072463768116
0.9492753623188406
0.8840579710144928
0.8695652173913043
0.855072463768116
0.8357487922705314
0.8612836438923396
0.822463768115942
0.8405797101449275
0.8014492753623188
0.8063241106719368
0.8043478260869565
0.7959866220735786
0.7888198757763976
0.7942028985507247
0.7880434782608695
0.783461210571185
0.7954911433172303
0.7971014492753623
0.7934782608695652
```

Naive Bayes' classifier using Multinomial priors

```
0.927536231884058
0.8115942028985508
0.8115942028985508
0.7862318840579711
0.8028985507246377
0.7801932367149759
0.7701863354037267
0.7844202898550725
0.7665056360708534
0.7724637681159421
0.761528326745718
0.7705314009661836
0.7658862876254181
0.7577639751552795
0.7603864734299517
0.7608695652173914
0.7621483375959079
0.7616747181964574
0.7589626239511823
0.7681159420289855
```

1.f.iii) Naive Bayes' classifier using Gaussian is better for multi-class classification in this problem

3.7.4) I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta 0 + \beta 1X + \beta 2X2 + \beta 3X3 + e$.

a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta 0 + \beta 1X + e$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Ans: It is expected the polynomial regression to have a lower training RSS than the linear regression because it could make a tighter fit against data that matched with a wider irreducible error e.

b) Answer (a) using test rather than training RSS.

Ans: It is the opposite of the answer for a), It is expected the polynomial regression to have a higher test RSS than the linear regression as the training would be overfitted and will increase the error e.

c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Ans: Because of higher flexibility, Polynomial regression has lower train RSS than the linear fit :no matter what the underlying true relationship is, the more flexible model will closer follow points and reduce train RSS.

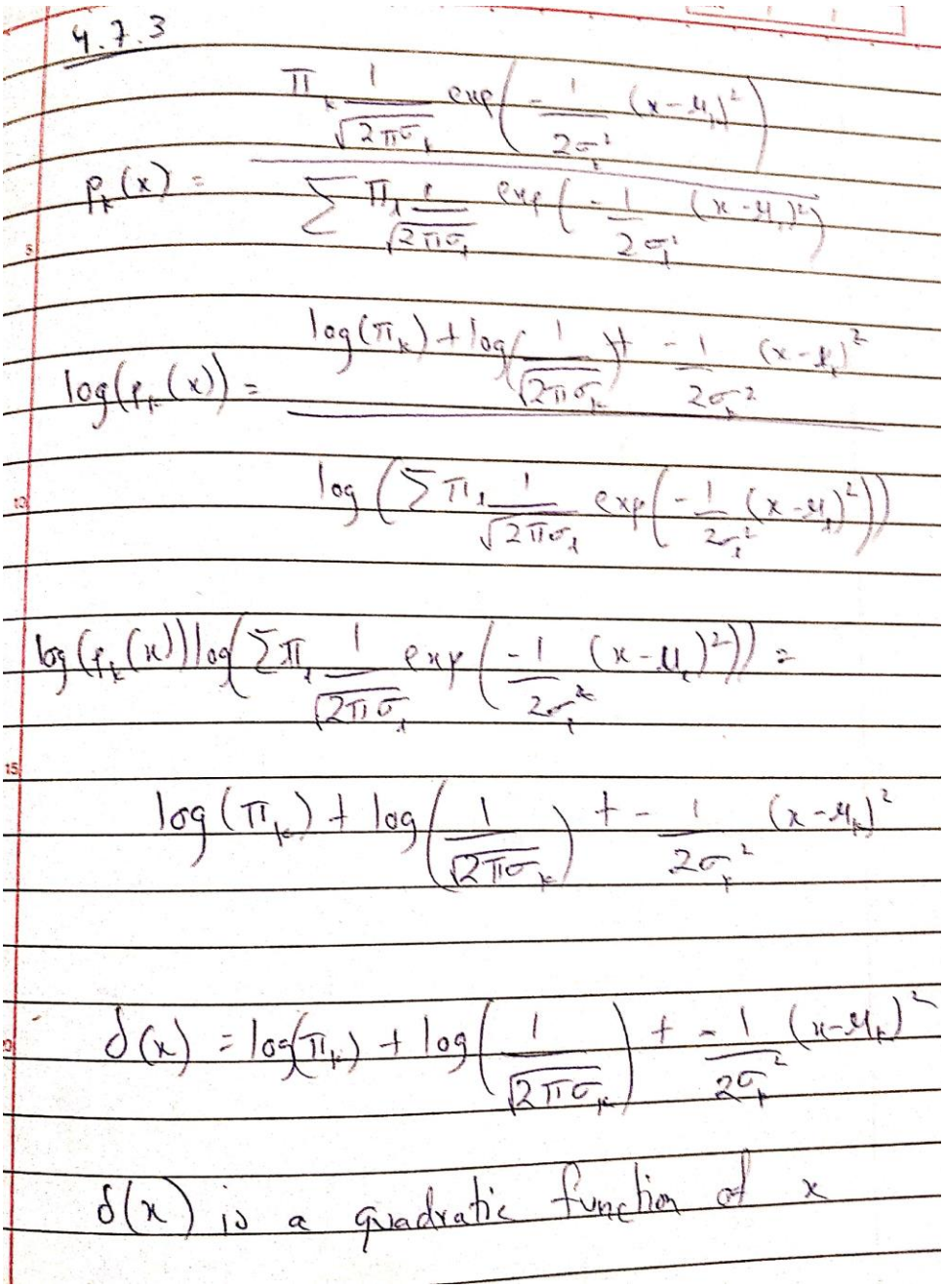
d) Answer (c) using test rather than training RSS.

Ans: There is not enough evidence from the information provided to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing how far it really is from linear. If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. If it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is due to bias-variance tradeoff: it is not clear what level of flexibility will fit data better.

4.7.3) This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class specific mean vector and a class specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature.

Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, X ~ N(μk, σ2 k). Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Ans:



4.7.7) 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.

Ans: