EE 657 - Pattern Recognition and Machine Learning

Assignment-1

Problem 1

The data from the file `‘P1 data train.csv’ and ‘P1 labels train.csv’` is being used to find the parameters means $\mu5$, $\mu6$ and covariances $\Sigma5$,$\Sigma6$ and the a priori π respectively for the classes.

Then the data from the files `‘P1 data test.csv’ , ‘P1 labels test.csv’` is being used to get the log (a posteriori probabilities) so as to decide to which class the feature vector belongs to , using the following equation:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \underset{\substack{\uparrow \\ \text{matrix} \\ \text{inverse}}}{\boldsymbol{\Sigma}_i^{-1}} (\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \underset{\substack{\uparrow \\ \text{matrix} \\ \text{determinant}}}{|\boldsymbol{\Sigma}_i|} + \ln \underset{\substack{\uparrow \\ \text{prior} \\ \text{probability}}}{P(\omega_i)}.$$

Then the respective discriminant scores g5(x) and g6(x) are compared. ***Bayesian Decision Theory*** is employed to make the decision. The class which gets the highest value of the log of a posteriori gets the feature vector.

The following cases have been taken and experimented:

1. The Covariance of the both the classes are **different** and the calculated for each class separately. ($\Sigma5 != \Sigma6$)
   In that case the results are as follows:
   misclassification rate for '5' is 31.613%
   misclassification rate for '6' is 15.1685393258 %
   And the Confusion matrix is [[106, 49], [27, 151]]
   The class ‘ 6 ’ is classified better than the class ‘ 5 ’
2. The covariance of the both the classes are **same** and is the weighted mean, weighted by the a priori probabilities:
   $\Sigma5 = \Sigma6 = \Sigma w = \pi_5 \Sigma5 + \pi_6 \Sigma6$
   In such a case , the results are promising and seem to be optimal
   misclassification rate for '5' is 13.5483870968 %

misclassification rate for '6' is 15.1685393258 %
And Confusion matrix is [[134, 21], [27, 151]]
In this case, both the classes are classified correctly to an extent and the error is small .

3. The covariance of the both the classes are equal to the weighted covariance except the fact that all the **non diagonal elements** are equal to **0**
   In such a case, the results are as follows
   misclassification rate for '5' is 14.1935483871 %
   misclassification rate for '6' is 23.0337078652 %
   Respective Confusion matrix is [[133, 22], [41, 137]]
   In this case, the class ‘5’ is classified better than the class ‘6’

Problem 2

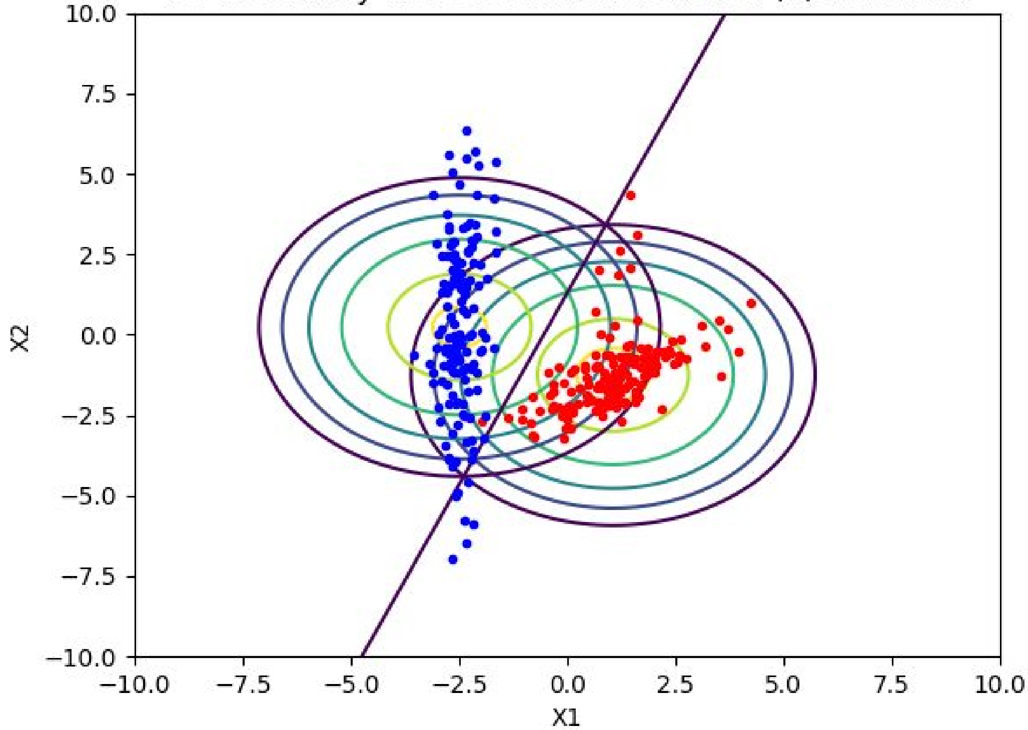
The data from the file ‘`P2 train.csv`’ is extracted and the *mean*s and the *covariance matrice*s of the respective classes are being found.
The same ***Bayesian Decision Theory*** is implemented to classify the data points.
The following cases are being studied :

A. *The covariance matrices of the both the classes are of the form [[a,0] , [0,a]]*
   i.e the diagonal elements are equal and the non diagonal elements are zero
   The log a posteriori values of the respective classes are calculated as follows

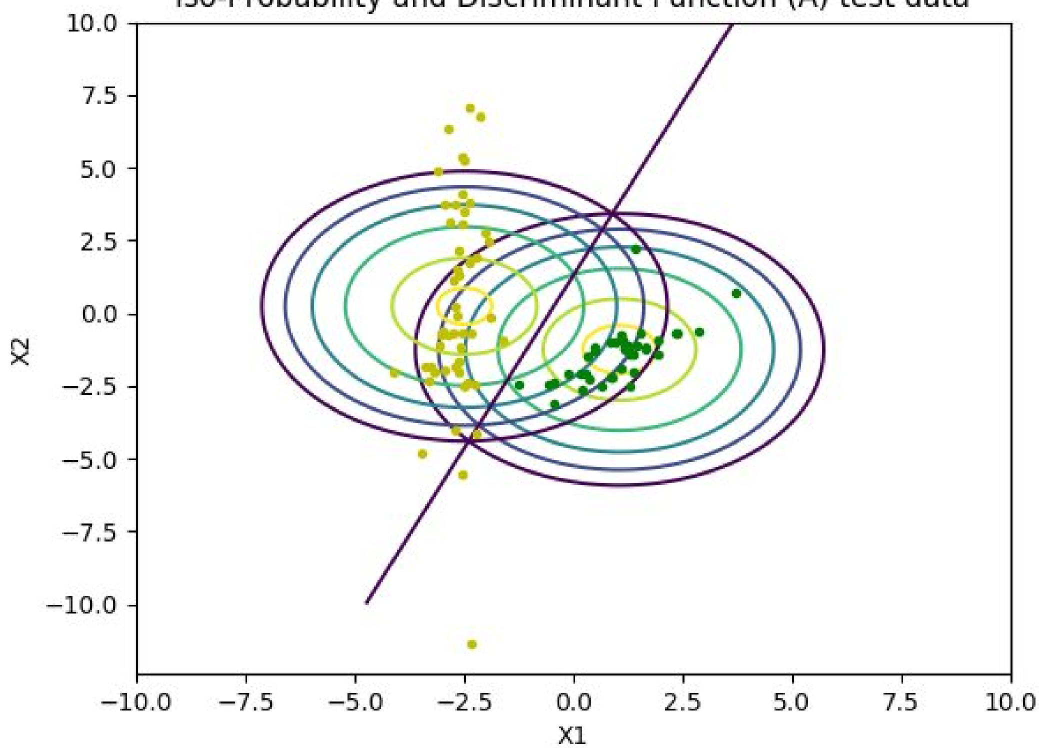
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \underbrace{\boldsymbol{\Sigma}_i^{-1}}_{\text{matrix inverse}}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \underbrace{|\boldsymbol{\Sigma}_i|}_{\text{matrix determinant}} + \ln \underbrace{P(\omega_i)}_{\text{prior probability}}.$$

   We can notice that the determinant doesn't involve in decision making because the determinant is same for both the classes.
   And further we can notice that the first part of the equation (the inverse part) is a function of ‘a’ and is inversely proportional to g(x)
   Hence it can be generalised that the change in value of ‘a’ doesn't make any change in the decision making
   The discriminant function is plotted by taking the difference in g1(x) and g2(x) and plotting the ‘0’ contour of the function
   The iso probability contours are plotted taking the likelihood function
   The discriminant function and the iso-probability contour plots on the train and test data are as follows: (the variance is equal along both the directions)

Iso-Probability and Discriminant Function (A) train data



Iso-Probability and Discriminant Function (A) test data



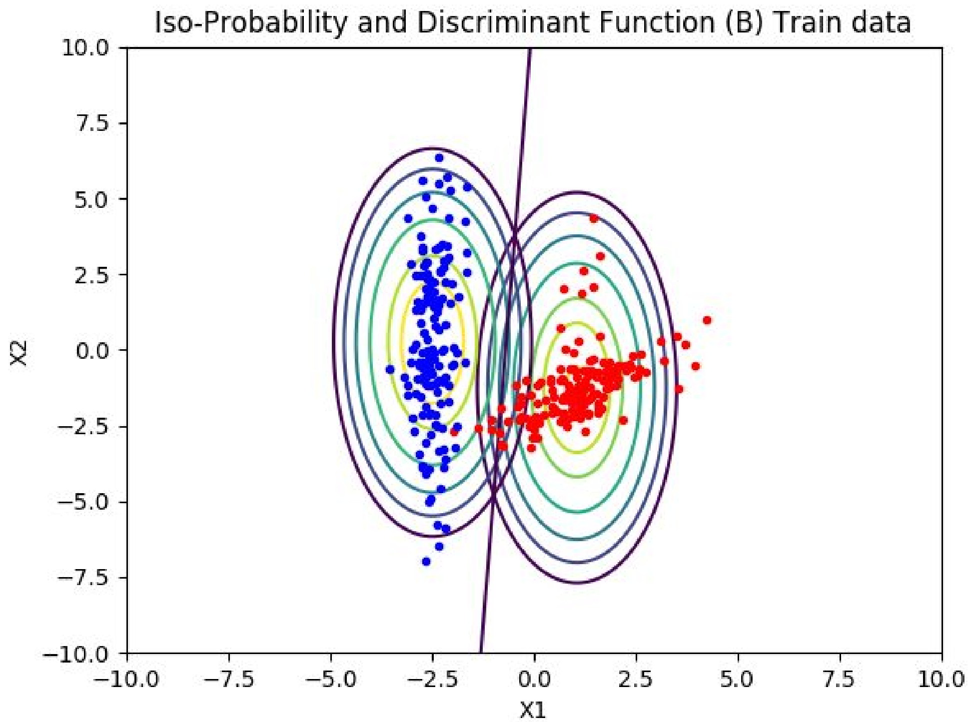
Observation:
From the above graph we can see that the iso probabilistic curves are in circular fashion. It is because the variance of the data is same in both the directions , and id same for the both the the classes. This make the decision difficult, hence the errors would me most in this case
misclassification rate for '1' is 0.0 %
misclassification rate for '0' is 6.0 %
Confusion matrix
[[40, 0], [3, 47]]

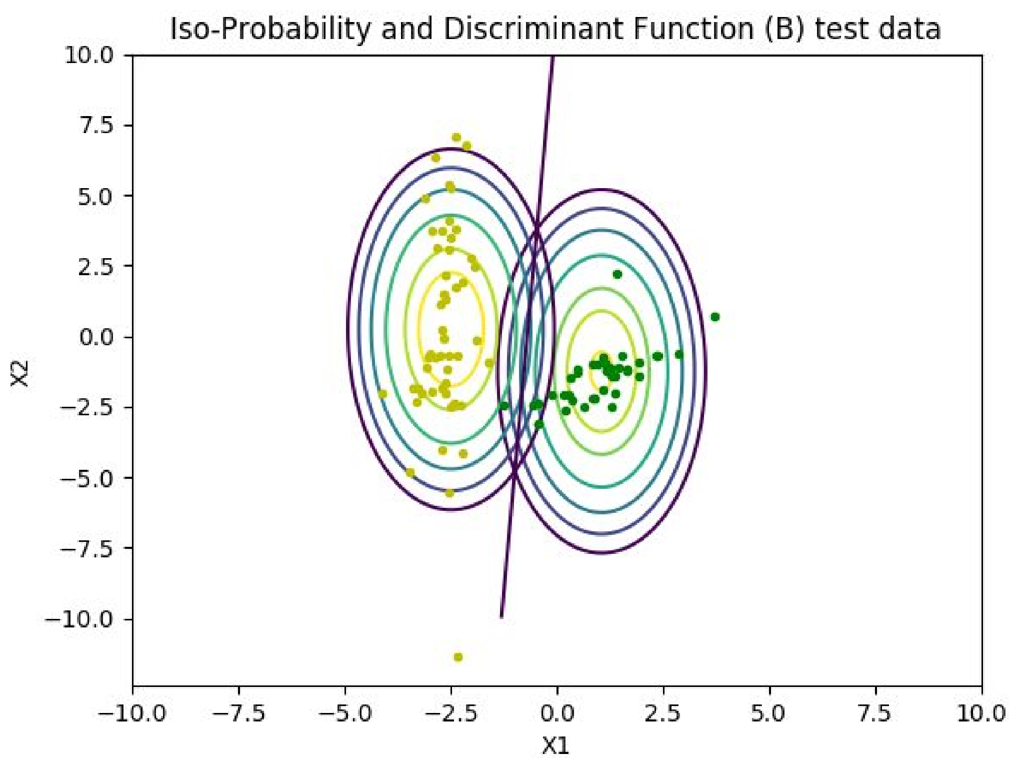
B. *The covariance matrices are equal and are of the form [[a,0],[0,b]]*

the variance is different in different directions

The values of a and b are found by taking log likelihood and maximizing with respect to the parameters.the result comes out to be the same as the general covariance matrix with non diagonal elements set to 0.

This makes the iso probabilistic curves elliptical rather than circles like in previous case
The discriminant function and the iso-probability contour plots on the train and test data are as follows:


Iso-Probability and Discriminant Function (B) Train data
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
X2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
X1


Iso-Probability and Discriminant Function (B) test data
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
X2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
X1

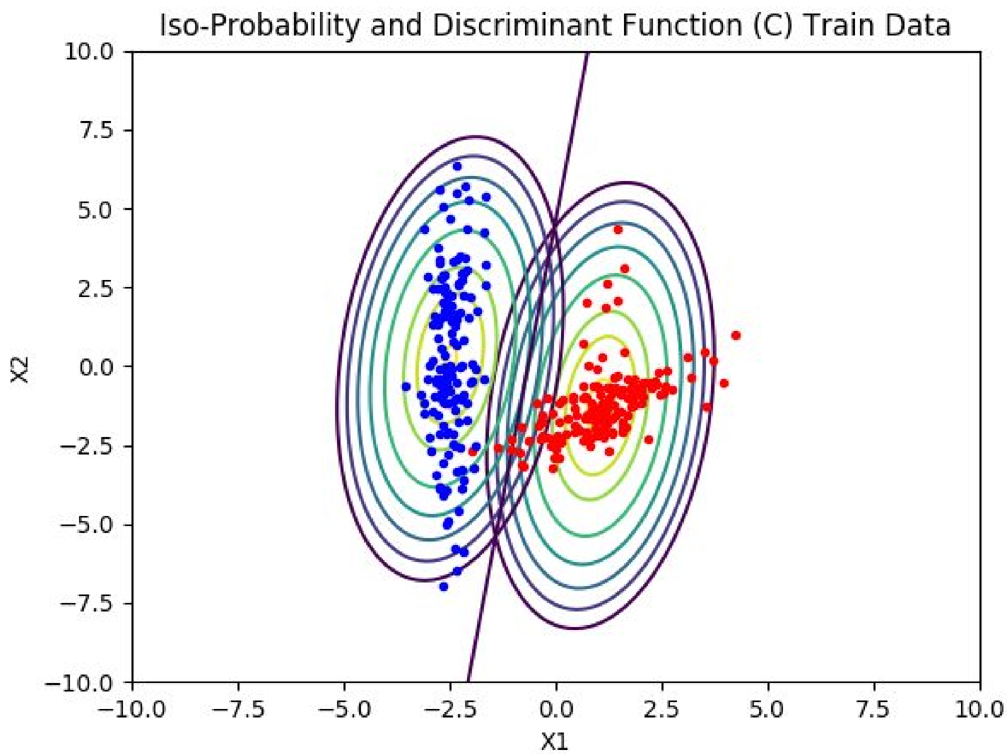
Observation:
The iso probabilistic curves are elliptical because the variances are different in the directions
The iso probabilistic curves of both the classes look alike because the covariance matrices are same for both the classes
When the Covariance is calculated for each class ,and are different
misclassification rate for '1' is 2.5 %
misclassification rate for '0' is 0.0 %
Confusion matrix
[[39, 1], [0, 50]]

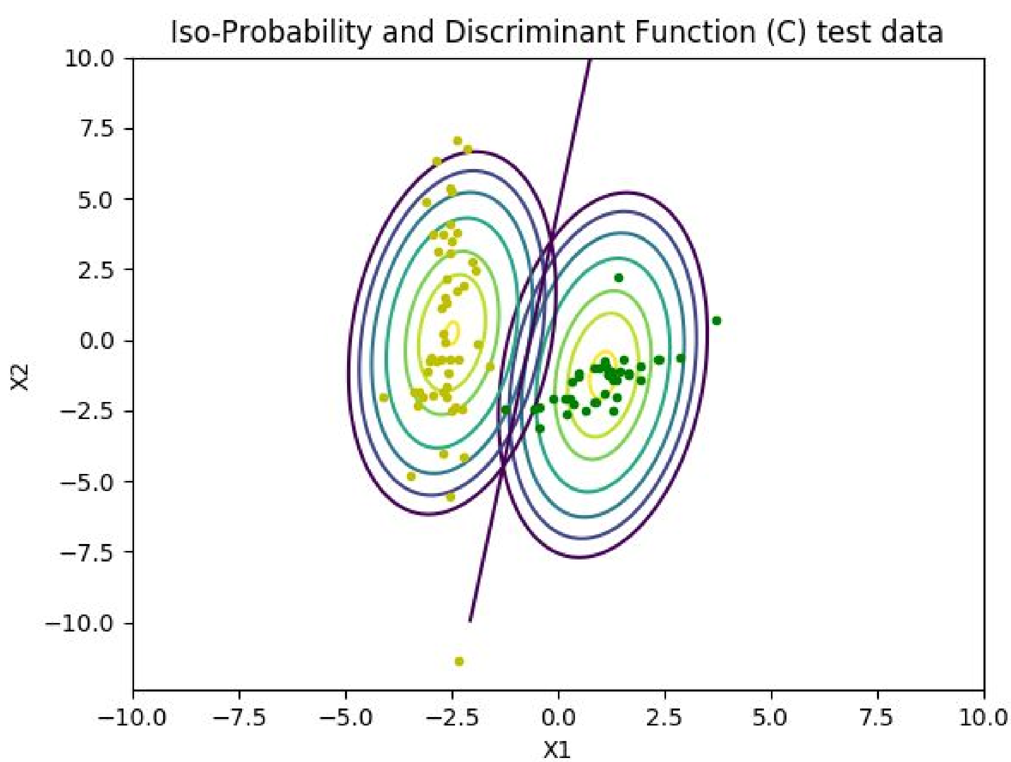
C. *The covariance matrices is same for the both classes* :

The covariances for each class is calculated individually and the weighted mean of the covariance matrices are calculated , which is weighted to the a priori probabilities

$$\Sigma_1=\Sigma_0=\Sigma_w=\pi_1\Sigma_1+\pi_0\Sigma_0$$

The discriminant function and the iso-probability contour plots on the train and test data are as follows:


Iso-Probability and Discriminant Function (C) Train Data
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
X2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
X1


Iso-Probability and Discriminant Function (C) test data
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
X2
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
X1

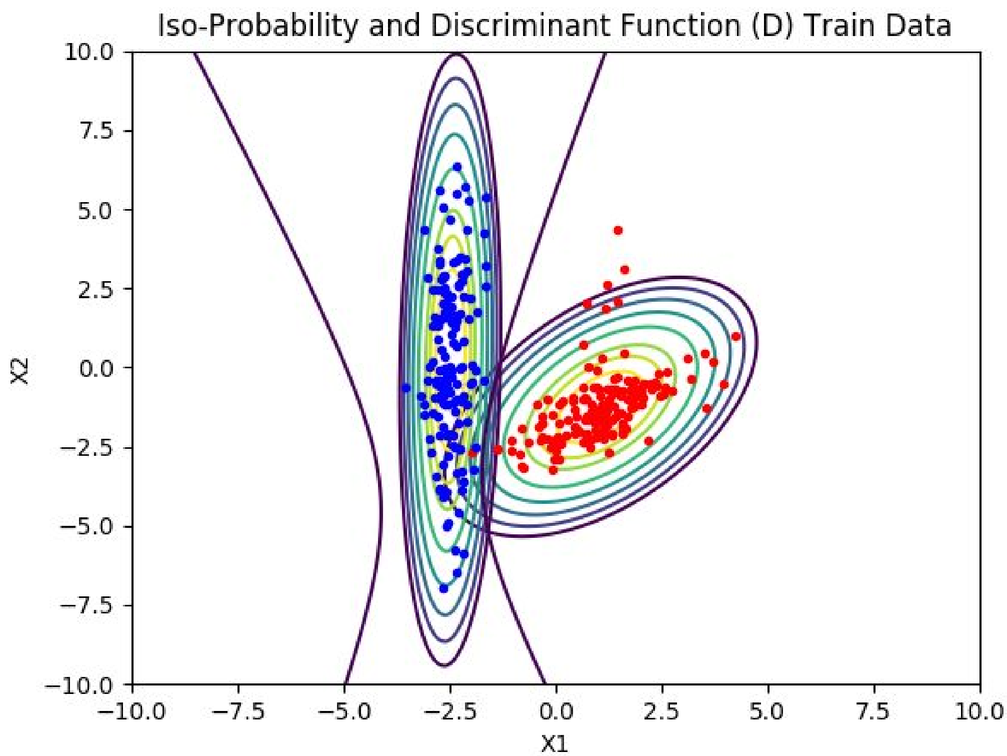
Observations:
Even now the iso probabilistic contours look the same because the covariance matrix is same for both the classes

When the Covariance is calculated for each class and are weighted mean is calculated
misclassification rate for '1' is 2.5 %
misclassification rate for '0' is 0.0 %
Confusion matrix
[[39, 1], [0, 50]]

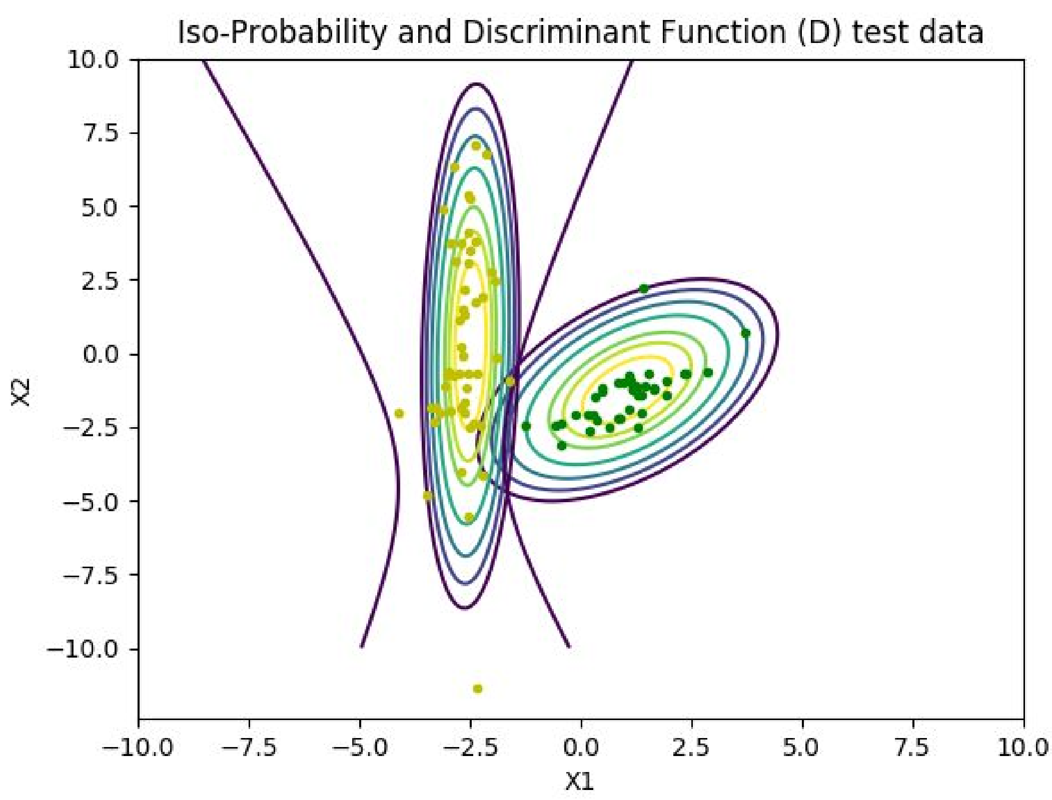
D.*The weighted mean of the covariance matrices are different for different classes*

The covariance matrices are calculated individually for both the classes, this gives the best results because the variances are different for the different classes and even in different directions. Hence the date is being represented by the iso probabilistic contours

The discriminant function and the iso probabilistic contours are as follows



Iso-Probability and Discriminant Function (D) Train Data
X2
X1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0



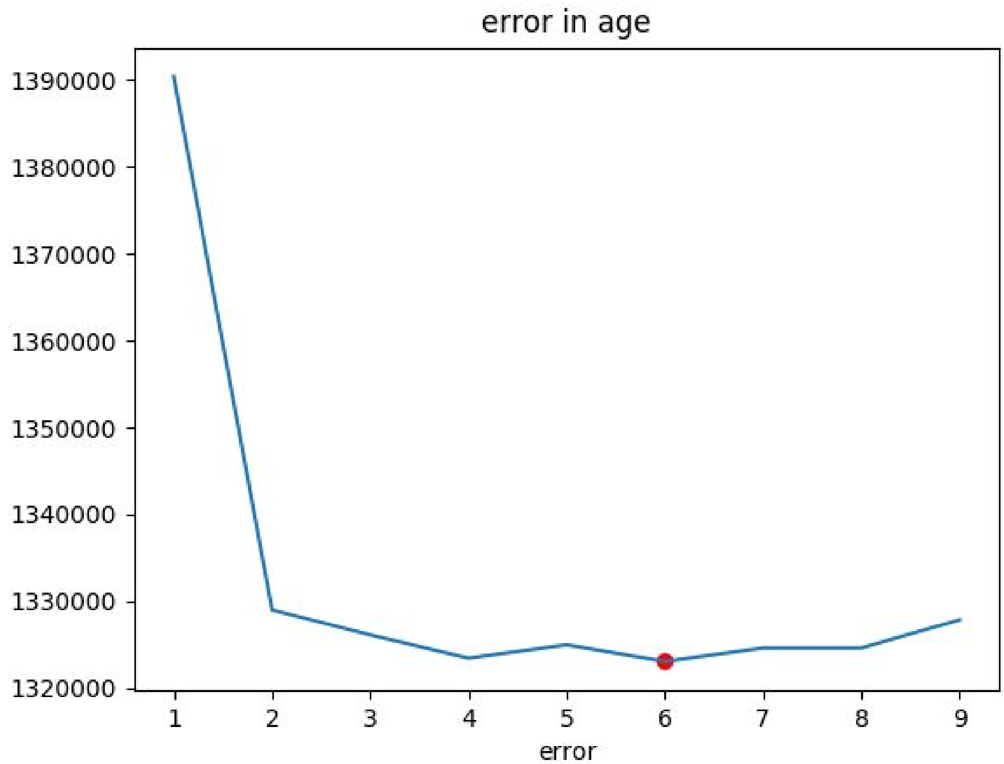
Iso-Probability and Discriminant Function (D) test data
X2
X1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

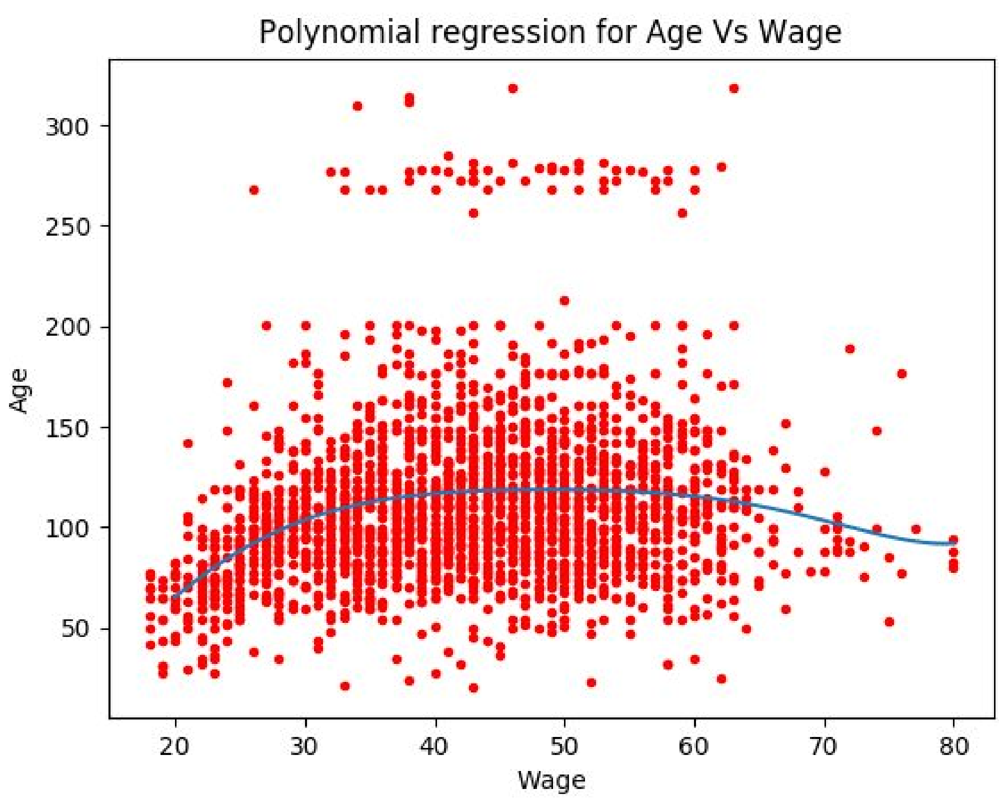
Problem 3

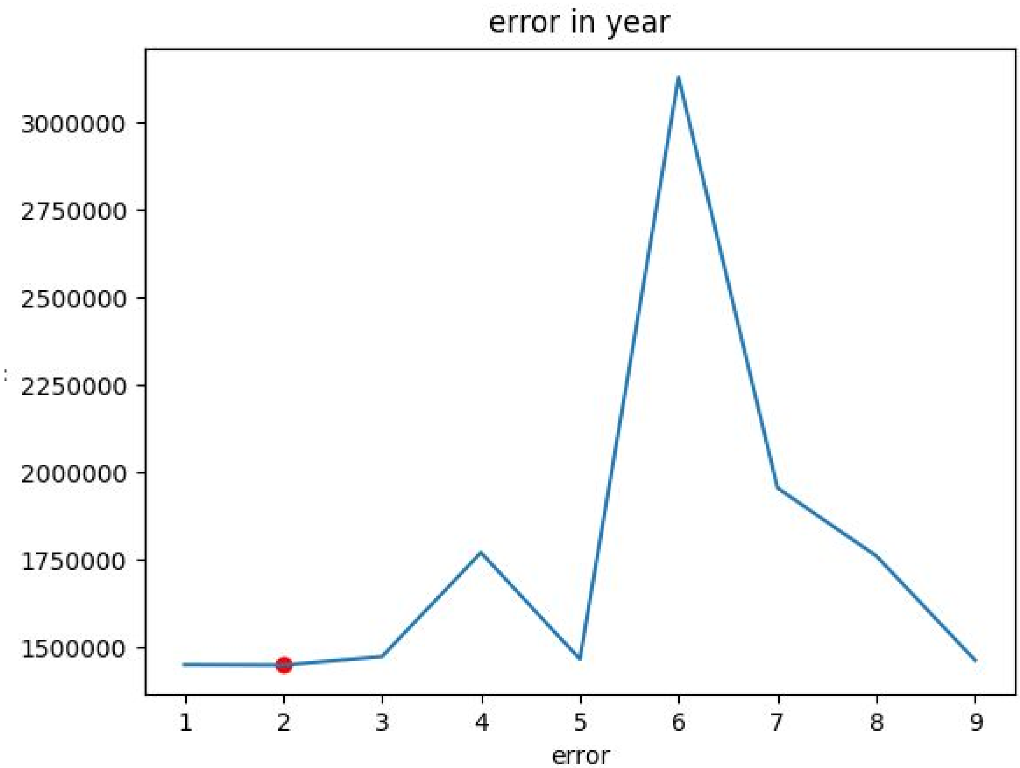
Initially the data set is divided into training data set and test data set {2250,750}
The polynomial regression of Age Vs Data is found by the following equation inverse(A’A).(A'r)
The optimal degree is found out by taking the error on the test data set

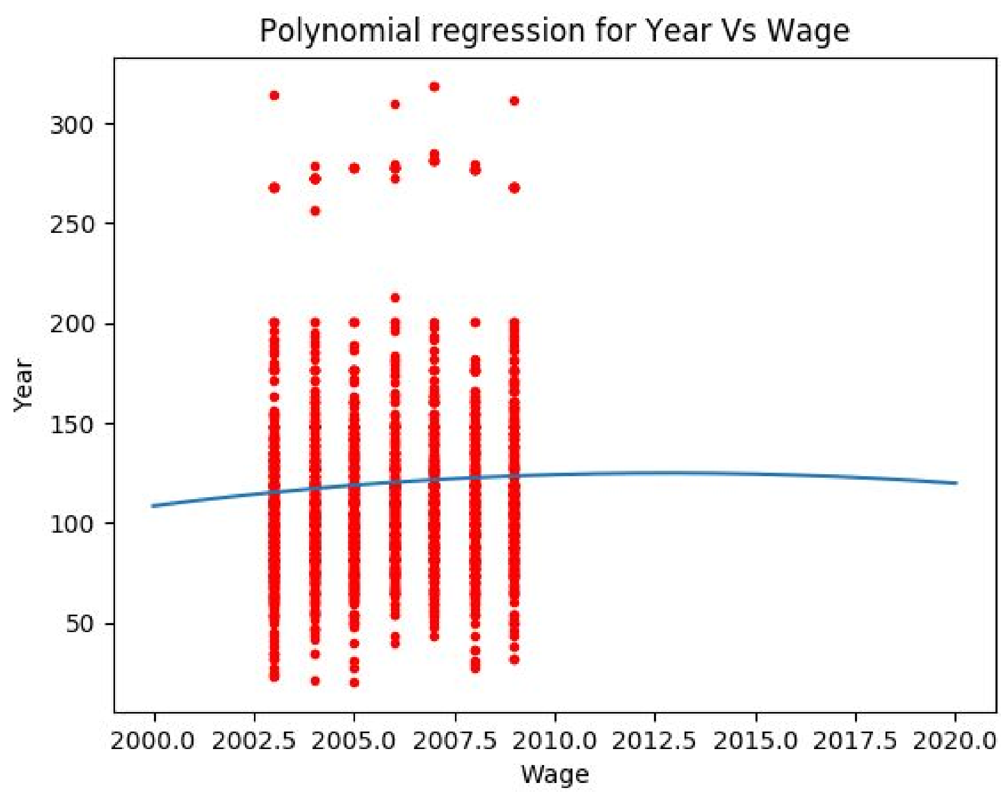
For the age vs wage , the optimal value of n is 6
For the year vs wage , the optimal value of n is 2
For the Education vs wage , the optimal value of n is 3
Given only one of the three attributes, it is not possible to estimate the wage of the person because the variance is very large and the error is very large

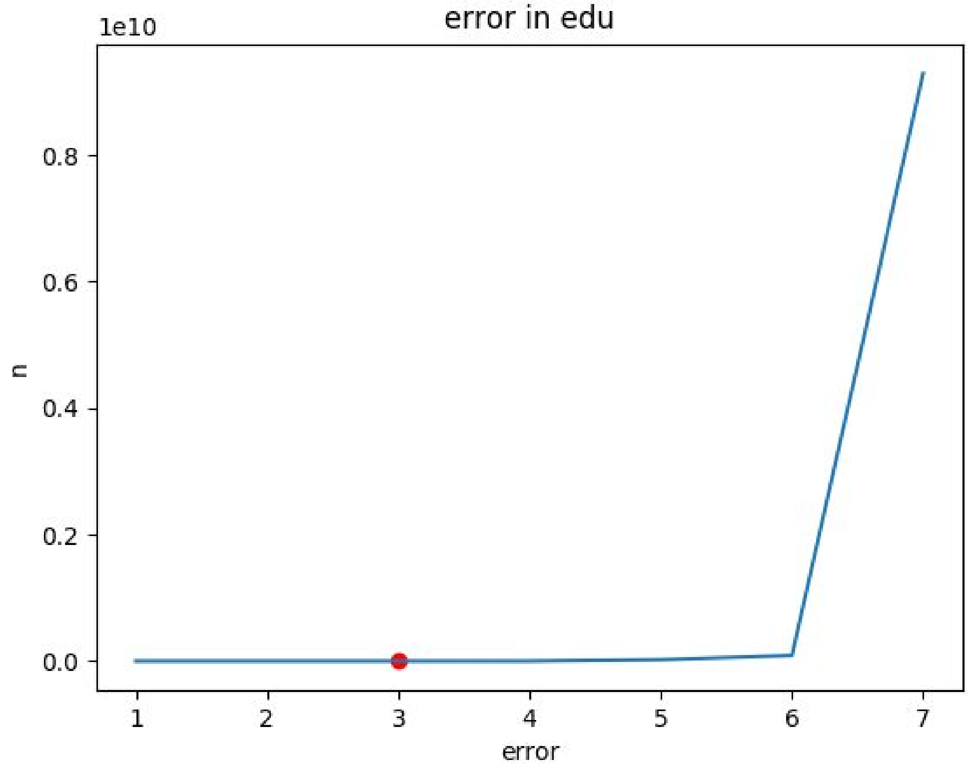
The plots of error vs value of n
The plots of polynomial regression are in the next pages

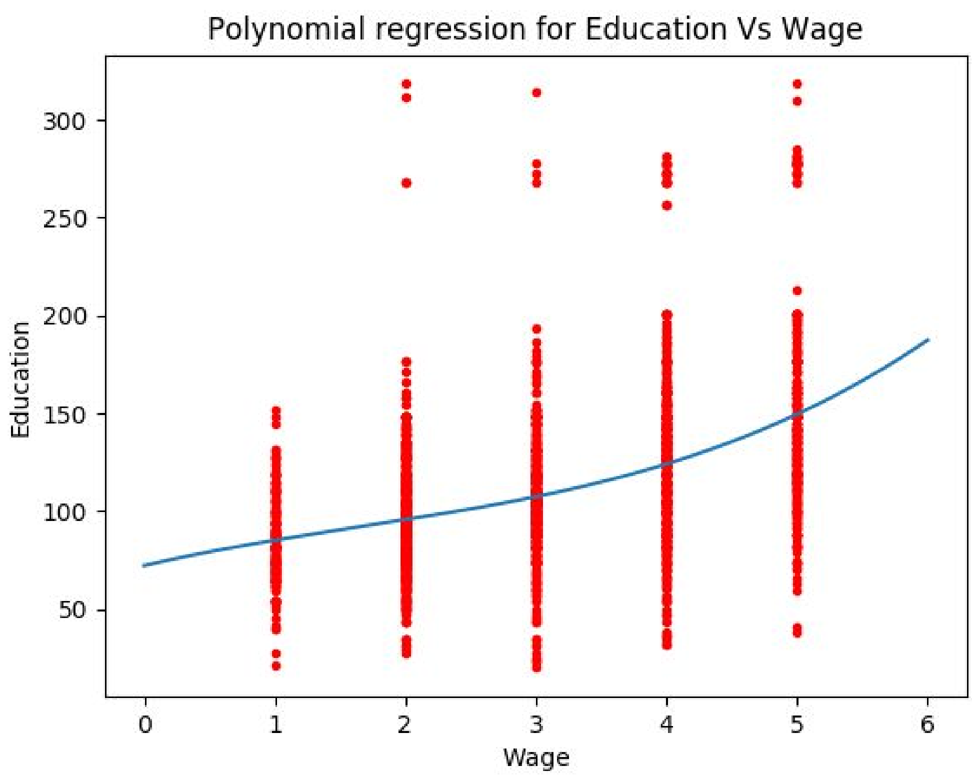

error in age
1390000
1380000
1370000
1360000
1350000
1340000
1330000
1320000
1
2
3
4
5
6
7
8
9
error


Polynomial regression for Age Vs Wage
Age
300
250
200
150
100
50
20
30
40
50
60
70
80
Wage


error in year
3000000
2750000
2500000
2250000
2000000
1750000
1500000
1
2
3
4
5
6
7
8
9
error


Polynomial regression for Year Vs Wage
300
250
200
150
100
50
Year
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
Wage


error in edu
1e10
0.8
0.6
0.4
0.2
0.0
n
1
2
3
4
5
6
7
error


Polynomial regression for Education Vs Wage
300
250
200
150
100
50
Education
0
1
2
3
4
5
6
Wage

Observations:
From the graphs, we can observe that the plot age vs graph is more likely to get an estimate of wage from age to some extent, but the rest of the two graphs do not give much information about the wage .