INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY
THIRUVANANTHAPURAM

Assignment #3

Due on 08-10-2014

Akhil P M (SC14M044)

Contents

1.Data Set 1 **3**

1.1 Logistic Regression 3

1.2 Gaussian Discriminant Analysis 4

1.3 Naive Bayes Classifier 4

2.Haberman's Survival Data Set **5**

2.1 Logistic Regression 5

2.2 Gaussian Discriminant Analysis 6

2.3 Naive Bayes Classifier 6

3.Performance Comparison **10**

4.Decision Boundaries **11**

5.Density Function and Contours **13**

6.ANOVA and Students t Distribution **15**

7.Spam Classification Problem **18**

1.Data Set 1

1.1 Logistic Regression

model parameters
Weight values -0.0931 -0.3633 -0.3034
α value :0.01

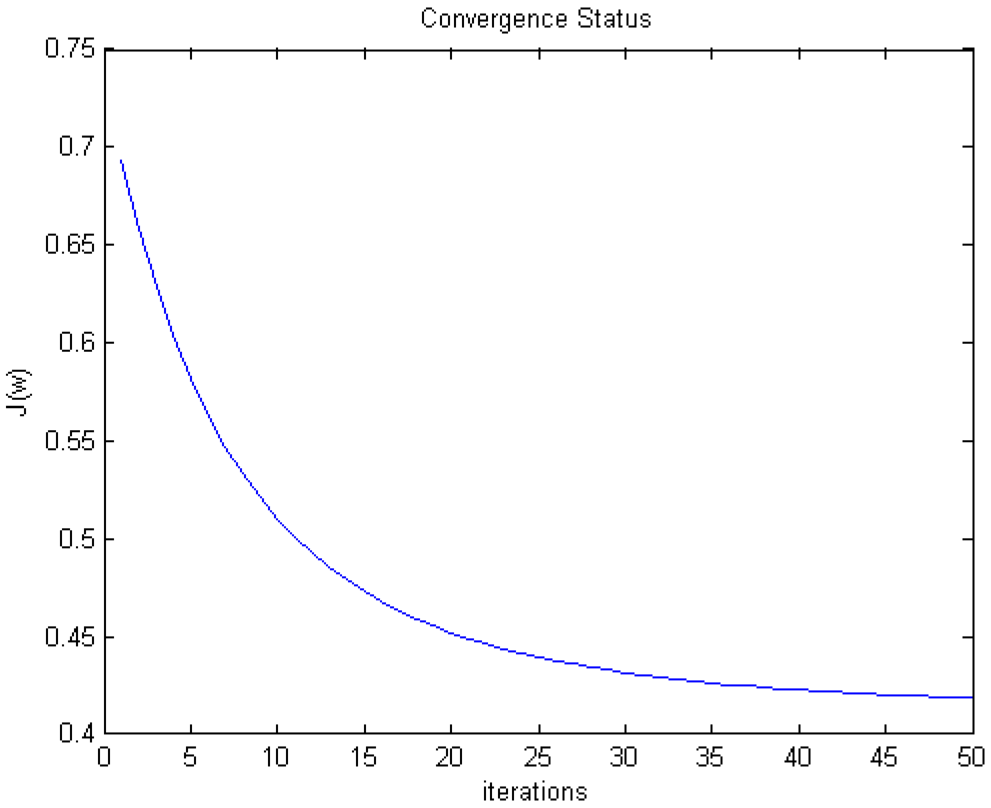

Figure 1: Logistic Regression

The ROC curve of Logistic Regression is shown below

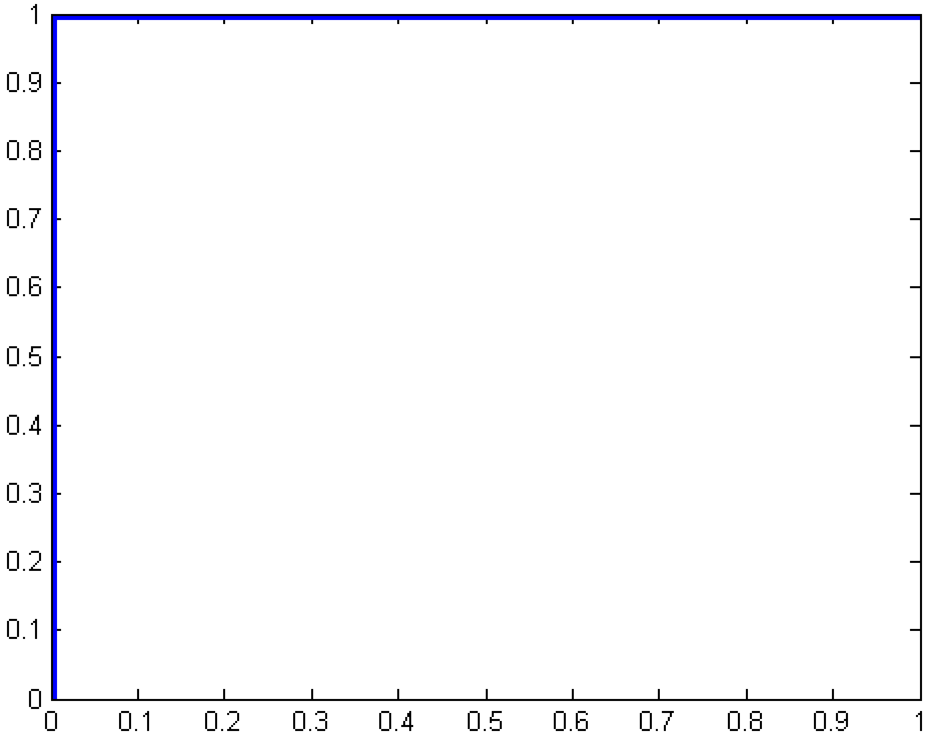

Figure 2: ROC curve for Logistic Regression

1.2 Gaussian Discriminant Analysis

The model parameters are:

```
Mean of Positive Class = -0.1709          -0.0055
Mean of Negative Class = 4.0029           3.9501

Covariance
5.6964          4.1731
4.1731          5.0065
```

1.3 Naive Bayes Classifier

Here the parameters of the model are Φ values

```
Φ(: ,: ,1) =
0.2000          0.1143
0.6286          0.7714
```

```
0.1714          0.1143
0          0
0          0

Φ(: ,: ,2) =
0          0
0          0
0.1714          0.2286
0.6286          0.5429
0.2000          0.2286
```

2.Haberman's Survival Data Set

2.1 Logistic Regression

The parameters of the model are

```
Weight values
0.6705          0.67050     0.0258
α value： 0.01
```

The ROC-curve obtained is shown below

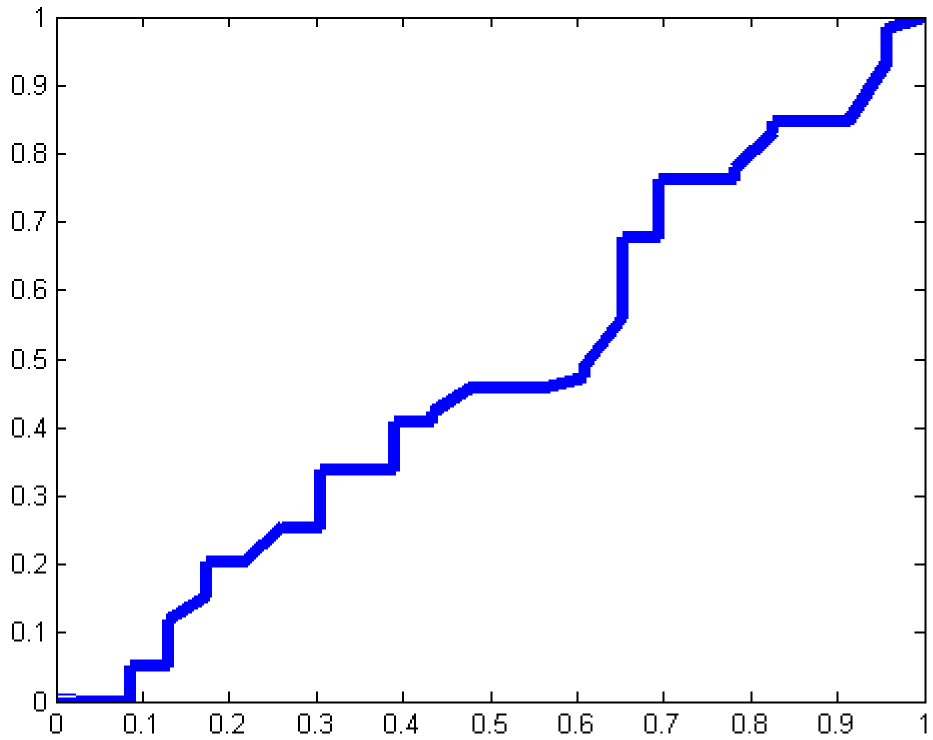

Figure 3: Logistic Regression ROC curve

2.2 Gaussian Discriminant Analysis

The model parameters are:

```
Mean of positive class = 47.0181        62.6265     2.9639
Mean of negative class = 48.5172         63.1034     8.3793
Covariance
55.4441            1.8464      1.4201
1.8464            10.3229     0.3520
1.4201            0.3520      56.0089
```

2.3 Naive Bayes Classifier

Here the parameters of the model are Φ values

```
Φ(: ,: ,1) =
0.2169          0.0058    0.8253
0.3434          0.1867    0.0843
0.4398          0.5241    0.0422
0          0.2892    0.0241
0.0058          0.0058    0.0241

Φ(: ,: ,2) =

0.0690          0.0159    0.4655
0.4655          0.2414    0.2069
0.4483          0.3793    0.1379
0.0172          0.3793    0.0517
0.0159          0.0159    0.1379
```

The ROC-curve obtained for Naive Bayes Classifier is shown below

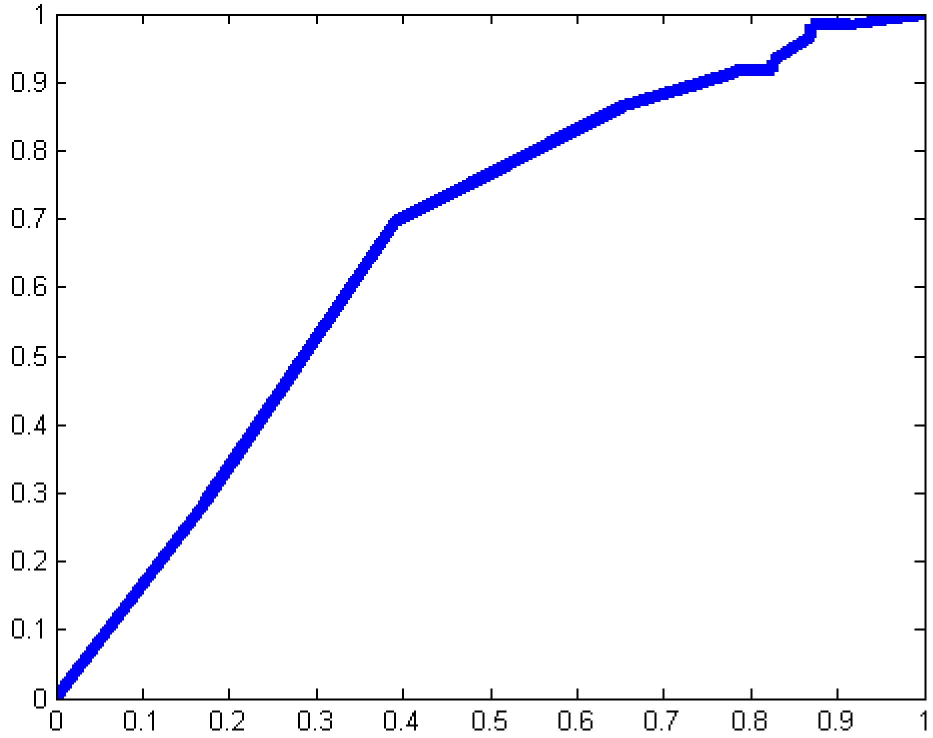

Figure 4: Naive Bayes ROC curve

The preprocessing that i did on the data are
1.**adding bias to the feature matrix X(for logistic regression)**.
it is done as follows

```
    [m n] = size(X);
    X = [ones(m,1) X];
```

2.**Attribute Discretizing(for Naive Bayes Classifier)**
Naive Bayes Classifier cannot work with continuous valued attributes,so attributes are to be discretized before the classification starts.The code that performs attribue discretization is shown below.

```
function [disc,classes] = AttributeDiscretizer(X,prob)

    % this function is hardcoded since the range of values that an
    % attribute takes is varying a lot for different attributes.
    [m n]=size(X);
    disc=zeros(m,n);
    classes=5;
    if(prob==1),

        for ii=1:m,
             for jj=1:2,
              val=X(ii,jj);
                if (val>=-3 && val< -1),
                     disc(ii,jj)=1;
                elseif (val>=-1 && val< 1),
                     disc(ii,jj)=2;
               elseif (val>=1 && val< 3),
                     disc(ii,jj)=3;
               elseif (val>=3 && val< 5),
                     disc(ii,jj)=4;
               elseif (val>=5),
                     disc(ii,jj)=5;
               end;
           end;
        end;


    else,
     for ii=1:m,
        val1=X(ii,1);
        val2=X(ii,2);
        val3=X(ii,3);
        if(val1>=30 && val1<40),
            disc(ii,1)=1;
         elseif(val1>=40 && val1<50),
            disc(ii,1)=2;
         elseif(val1>=50 && val1<60),
            disc(ii,1)=3;
         elseif(val1>=60 && val1<70),
            disc(ii,1)=4;
         elseif(val1>=70),
            disc(ii,1)=5;
        end;

        if(val2>=50 && val2<55),
            disc(ii,2)=1;
         elseif(val2>=55 && val2<60),
            disc(ii,2)=2;
         elseif(val2>=60 && val2<65),
            disc(ii,2)=3;
         elseif(val2>=65 && val2<70),
            disc(ii,2)=4;
         elseif(val2>=70),
            disc(ii,2)=5;
        end;

        if(val3>=0 && val3<5),
            disc(ii,3)=1;
         elseif(val3>=5 && val3<10),
            disc(ii,3)=2;
         elseif(val3>=10 && val3<15),
            disc(ii,3)=3;
         elseif(val3>=15 && val3<20),
            disc(ii,3)=4;
         elseif(val3>=20),
            disc(ii,3)=5;
        end;
      end;
    end;
end
```

1.Attribute Discretizing

3.**Feature scaling**

The basic idea of feature scaling is to make sure that features are on a similiar scale. This will in turn results in faster convergence of gradient descent[1]. The common technique is to make every feature in the range 0≤ X_i ≤1. Feature scaling for a vector x is done as follows

$$x_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Feature scaling is only applied for logistic regression with the second data set.The code that performs the feature scaling is shown below.

```
% This fuction performs scaling of the features to the range 0<=i<=1.
% feature scaling facilitaes easy convergence


function X = featureScale(x)

   % debug_on_warning(1);
    %debug_on_error(1);

    [m,n]=size(x);
    for j=2:n,
        maxval(j-1)=max(x(:,j));
        minval(j-1)=min(x(:,j));

    end;

    for j=2:n,
        difference=maxval(j-1)-minval(j-1);
        for i=1:m,
            x(i,j)=(x(i,j)-minval(j-1))/difference;
        end;
    end;
    X=x;
    %disp(x);

end
```

2.Feature Scaling

3.Performance Comparison

Logistic regression gives poor results on data set 1.

```
no of test datasets :30
no of misclassifications :10
accuracy :0.667
precision :1.000
recall/sensitivity :0.333
F-Measure :0.500
```

GDA gives 100% classification accuracy on data set 1.

```
no of test datasets :30
no of misclassifications :0
accuracy :1.000
precision :1.000
recall/sensitivity :1.000
F-Measure :1.000
```

Naive Bayes Classifier also give a good performance on data set 1.

```
no of test datasets :30
no of misclassifications :1
accuracy :0.967
precision :1.000
recall/sensitivity :0.933
F-Measure :0.966
```

The results for Haberman‘s survival data set is as shown below.

Logistic Regression gives best result for classification of these data set

```
no of test datasets :82
no of misclassifications :0
accuracy :1.000
precision :1.000
recall/sensitivity :1.000
F-Measure :1.000
```

GDA also performs well in case of second data set.

```
no of test datasets :82
no of misclassifications :4
accuracy :0.932
precision :1.000
recall/sensitivity :0.932
F-Measure :0.965
```

But Naive Bayes classifier is giving poor performance on this data set

```
no of test datasets :82
no of misclassifications :46
accuracy :0.220
precision :1.000
recall/sensitivity :0.220
F-Measure :0.361
```

4.Decision Boundaries

1.Logistic Regression Decision Boundary

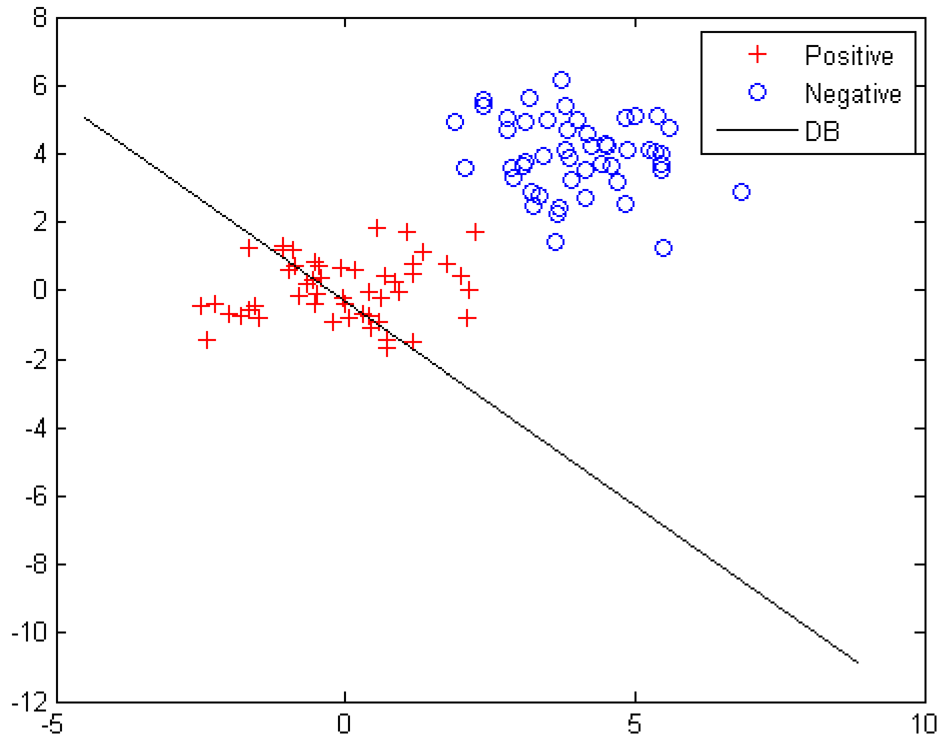

Figure 5: Logistic Regression Data 1

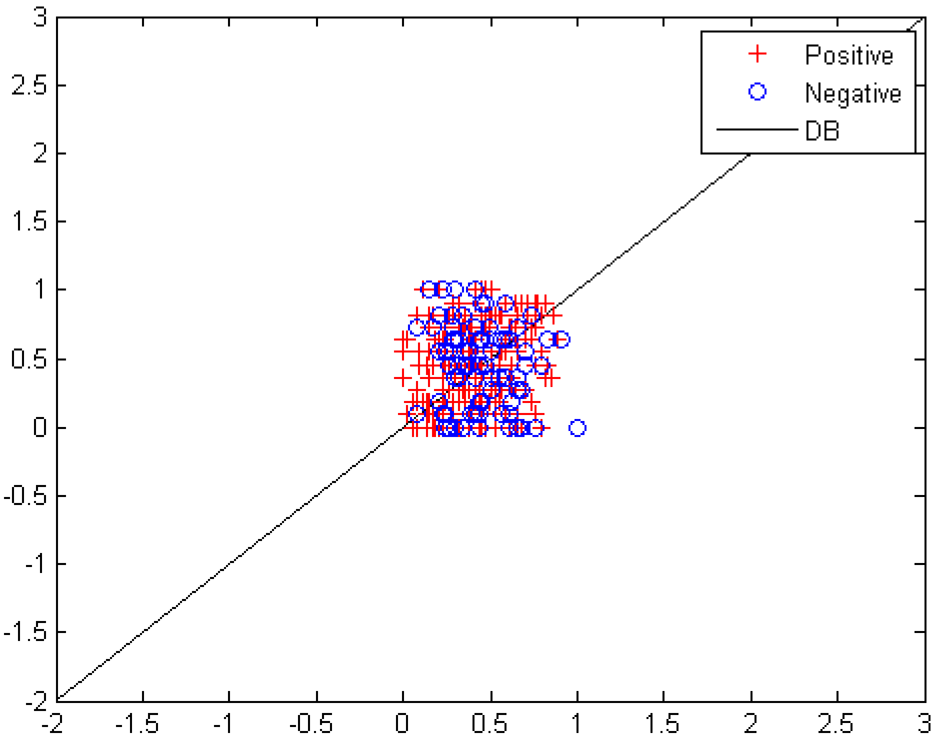

Figure 6: Logistic Regression Data 2

2.GDA Decision Boundary

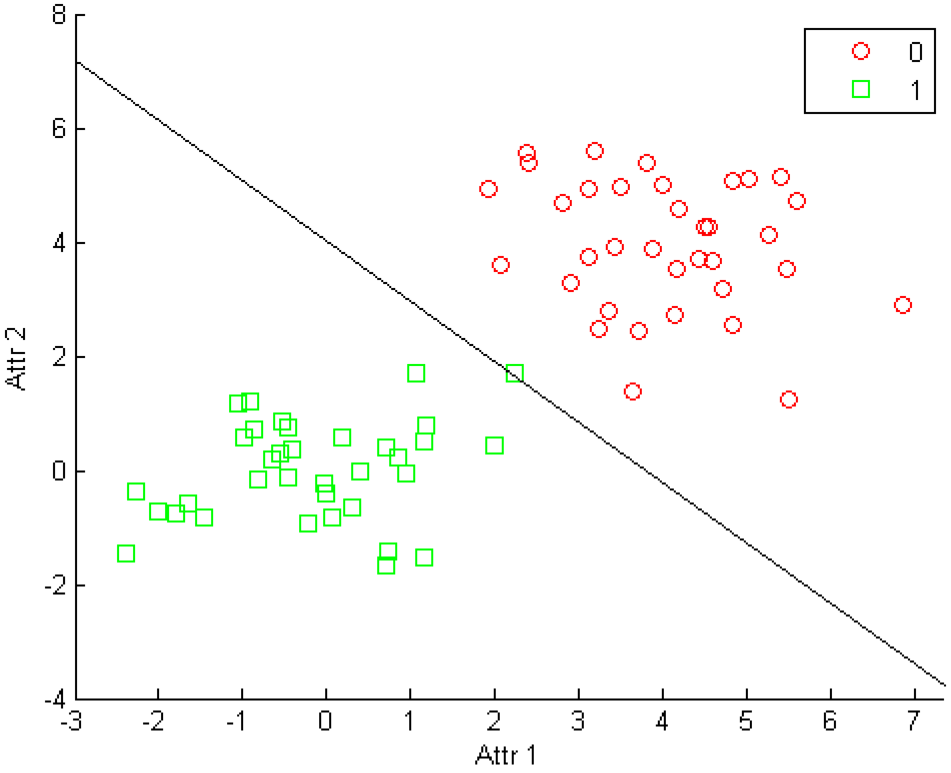

Figure 7: GDA

5.Density Function and Contours

1.Multi Variate Gaussian Density Function for Data 1

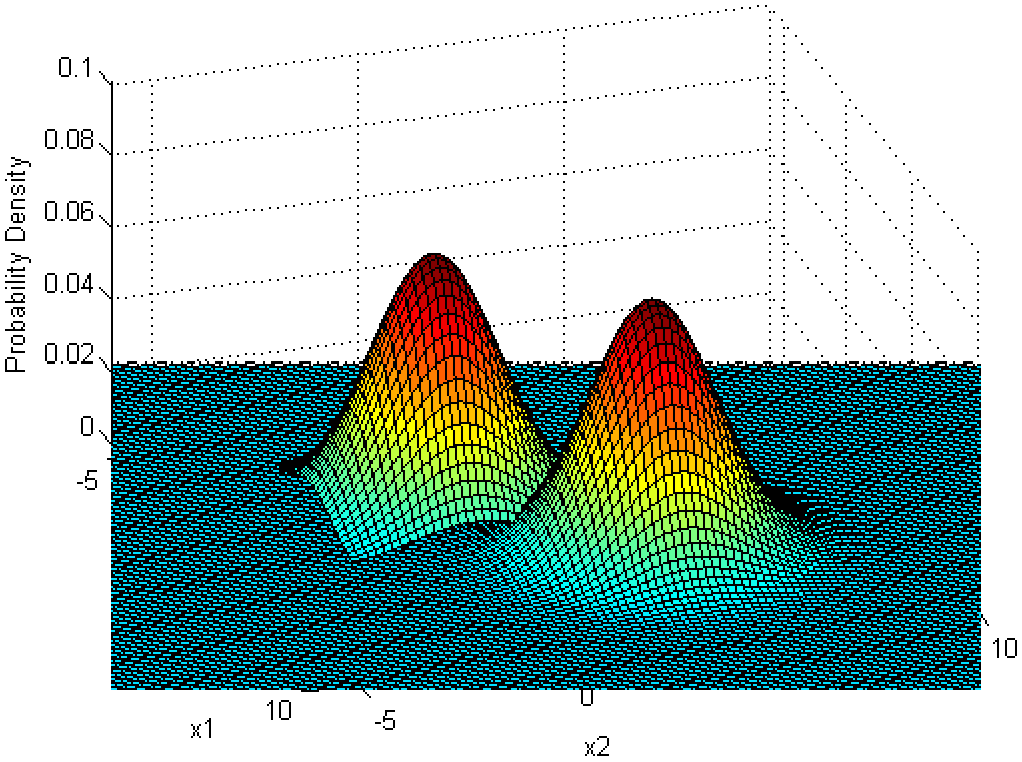

Figure 8: Multi Variate Gaussian Density Function

2.Contours for Data 1

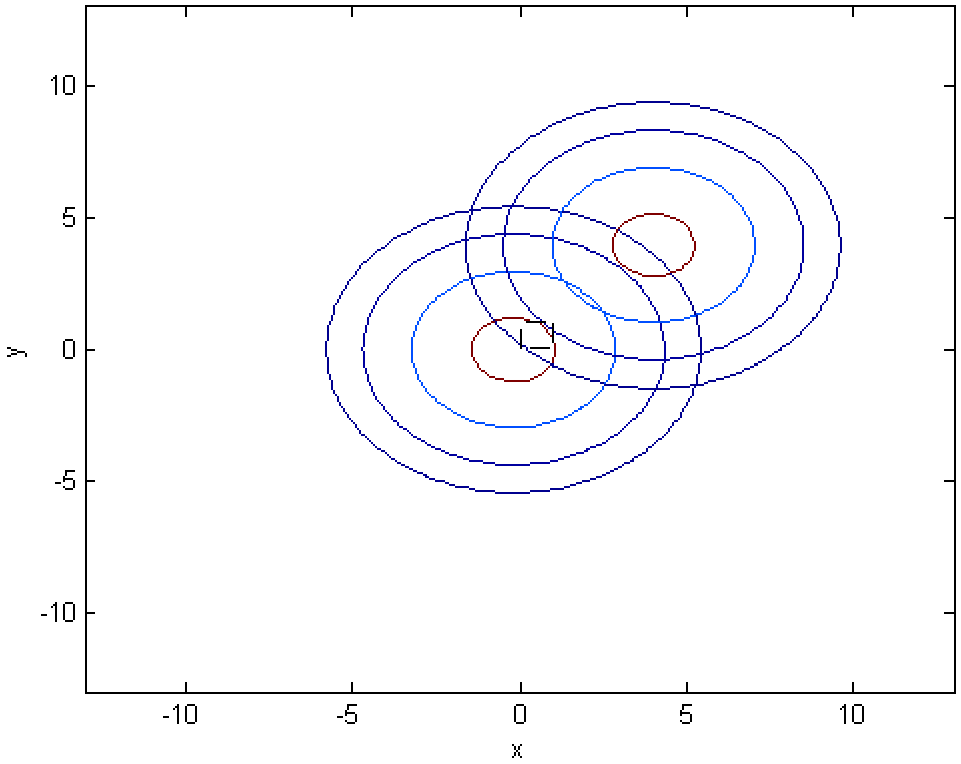

Figure 9: Contours of Data 1

6.ANOVA and Students t Distribution

ANOVA Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called ”Analysis of Variance” rather than ”Analysis of Means.” The name is appropriate because inferences about means are made by analyzing variance.
When we have only two samples we can use the t-test to compare the means of the samples but it might become unreliable in case of more than two samples. If we only compare two means, then the t-test (independent samples) will give the same results as the ANOVA[2].
The ANOVA test makes the following assumptions about the data in X:

- All sample populations are normally distributed.
- All sample populations have equal variance.
- All observations are mutually independent.

The ANOVA test is known to be robust with respect to modest violations of the first two assumptions.
The ANOVA test returns a measure called p-value.If p is near zero, it casts doubt on the null hypothesis and suggests that at least one sample mean is significantly different than the other sample means.ANOVA‘s use an F-ratio as its significance statistic which is variance because it is impossible to calculate the sample means difference with more than two samples[3].
T-tests are easier to conduct, so why not conduct a t-test for the possible interactions in the experiment? A Type I error is the answer because the more hypothesis tests you use the more you risk making a type I error and the less

power a test has. There is no disputing the t-test changed statistics with its ability to find significance with a small sample, but as previously mentioned the ANOVA allowed for testing more than 2 means. ANOVA‘s are used a lot professionally when testing pharmaceuticals and therapies.

ANOVA result for Data1 : p=0.9601

P is greater than significance level, hence the means are almost same.It can be analyzed from the boxplot also.Hence null hypothesis is accepted

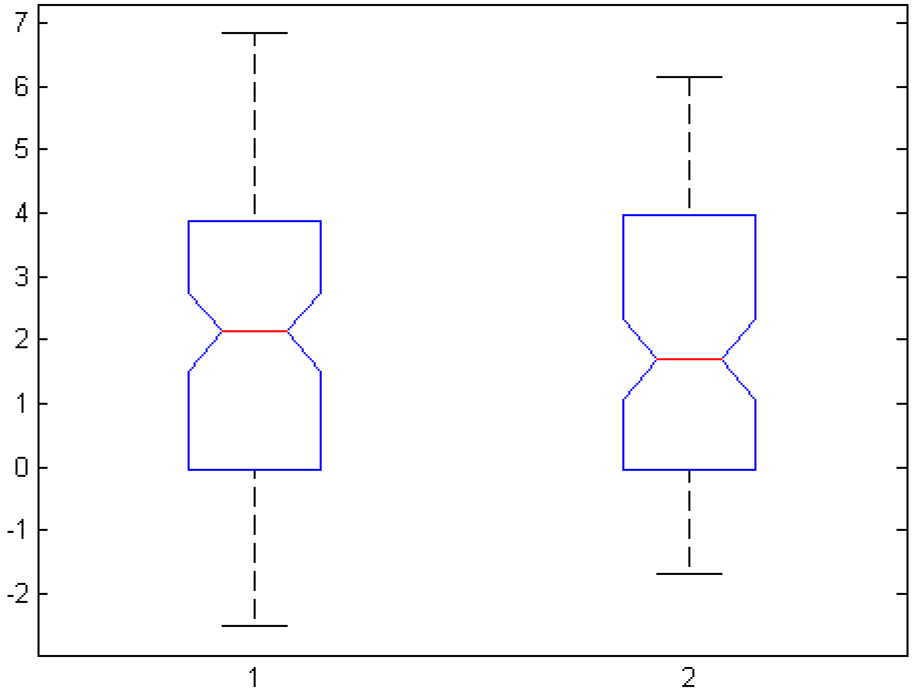

Figure 10: Box plot for data1

ANOVA result for Haberman‘s survival data set: p =0

since P is a small value we reject the null hypothesis.ie,atleast one of the mean is significantly different from all other means.The boxplot also gives the same result.

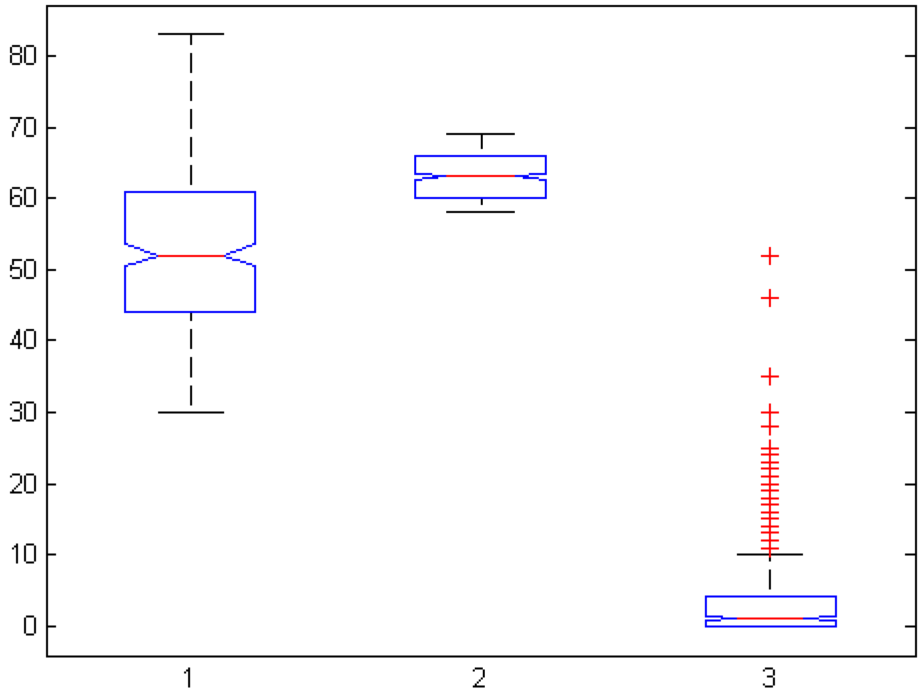

Figure 11: Box plot for Data2

T Distribution

The t distribution (or, Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.It is a type of probability distribution that is theoretical and resembles a normal distribution. A T distribution differs from the normal distribution by its degrees of freedom. The higher the degrees of freedom, the closer that distribution will resemble a standard normal distribution with a mean of 0, and a standard deviation of 1.

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean[4].

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$$\frac{\overline{x}-\mu}{s/\sqrt{n}}$$

where $\overline{x}$ is the sample mean,μ is the population mean,s is the standard deviation of the sample, and n is the sample size.

There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom. The degrees of freedom refers to the number of independent observations in a set of data.The t distribution has the following properties:

- The mean of the distribution is equal to 0.
- The variance is equal to $\frac{v}{(v-2)}$, where v is the degrees of freedom and v≥2.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The t distribution should not be used with small samples from populations that are not approximately normal[4].

The result of the t-test for DATA 1 is [1 1],hence it rejects the null hypothesis at the 5% significance level.
T-test for Haberman‘s survival data set gives the result as [1 1 1], and hence it also rejects the null hypothesis.
Thus T-test rejects null hypothesis for both the data sets,but ANOVA rejects null hypothesis for the second data set only.

7.Spam Classification Problem

The dictionary for this problem can be constructed as

word	**total**	**+ve**	**-ve**
send	4	3	1
us	4	3	1
your	5	3	2
internet	3	2	1
banking	3	2	1
password	3	2	1
review	3	1	2
account	1	1	0
details	1	1	0

The prior probabilities are
P(spam)=$\frac{2}{3}$ and
P(ham)=$\frac{1}{3}$

Now consider the first mail - ”Review your account” denote is as T_1
Applying Naive bayes classification techniques,

P(spam/T_1) = P(spam)*P(T_1/spam)
=$\frac{2}{3}*\frac{1}{3}*\frac{3}{5}*1=\frac{2}{15}$
P(ham/T_1) = P(ham)*P(T_1/ham)
=$\frac{1}{3}*\frac{2}{3}*\frac{2}{5}*0=0$

Hence P(spam/T_1) > P(ham/T_1).
So the given mail might be a spam.

consider the second mail - ”Review us now” denote is as T_2
Here the word ”now” is not present either in the positive or negative class.So we have to perform Laplace smoothing first.

P(now in +ve class) =$\frac{1}{4+9}$ = $\frac{1}{13}$
P(now in -ve class) =$\frac{1}{2+9}$ = $\frac{1}{11}$

P(spam/T_2) = P(spam)*P(T_2/spam)
=$\frac{2}{3}*\frac{1}{3}*\frac{3}{4}*\frac{1}{13}=\frac{1}{78}$
P(ham/T_2) = P(ham)*P(T_2/ham)
=$\frac{1}{3}*\frac{2}{3}*\frac{1}{4}*\frac{1}{11}=\frac{1}{198}$

Hence P(spam/T_2) > P(ham/T_2).
So the given mail might be a spam.

References

[1] Machine Learning - Coursera, *https://class.coursera.org/ml-005/lecture?lecture_player=html5*

[2] *http://www.mathworks.in/help/stats/anova1.html*

[3] *http://www.nku.edu/ statistics/Analysis*$_{o}f_{V}ariance_{E}xample.htm$

[4] *http://pages.uoregon.edu/aarong/teaching/G4074*$_{O}utline/node23.html$