Programming Assignment 1

Sneha Reddy Aenugu, EE11B059

October 3, 2015

1 Linear Classifier

- The dataset is generated with the required constraints and is attached in the zip file as *DS1*
- Figure 1 gives the estimated coefficients of the regression model.

  Accuracy of the regression classifier = 89.75% Root Mean Squared Error = 0.311
- Table 1 gives accuracy of KNN classifiers for different values of k.

  KNN is not better than the regression.There are few values of k where accuracy goes up slightly as shown in Table 1

2 Linear Regression

- Among all the features there are 22 features which have 1675 features missing out of 1994 features. So these features were discarded because of the uncertainty in the readings.
- The coefficients learned are a vector of 101 features and the model is given in the zip file submitted. The Root Mean Squared Error for the problem is 0.132.

3 Feature Selection

- Figure 2 shows the 3D plot of the dataset. The figure shows the data is seperable in one dimension. PCA is performed on the dataset and a single feature is extracted. The extracted features are then fitted with the indicator variables given by the labels using linear regression. The linear model is then used to predict the classes of the test dataset. Figure 2 plots the data points with their labels. The decision boundary used for classification is also included. Figure 3 plots the predicted value of the datapoints with the threshold line.

Linear regression model:
y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.68623	0.038051	18.035	1.8416e-64
x1	0.15475	0.012195	12.69	1.0739e-34
x2	0.15927	0.011126	14.315	5.2942e-43
x3	0.051078	0.017066	2.993	0.0028199
x4	-0.12018	0.0099568	-12.071	9.9387e-32
x5	-0.15197	0.010451	-14.541	3.2579e-44
x6	0.07827	0.012018	6.5128	1.086e-10
x7	-0.017274	0.010522	-1.6417	0.10091
x8	0.15539	0.015258	10.184	2.0701e-23
x9	-0.23532	0.016329	-14.411	1.6254e-43
x10	-0.045784	0.015263	-2.9996	0.0027598

Figure 1: Estimated coefficients of the regression model

k	Accuracy
1	74
2	74
3	82.25
4	81.25
5	83.75
6	82.75
7	82.5
8	82.25
9	83.75
10	83.5
11	83
12	82.25
13	83

Table 1: Accuracy in knn-classifier

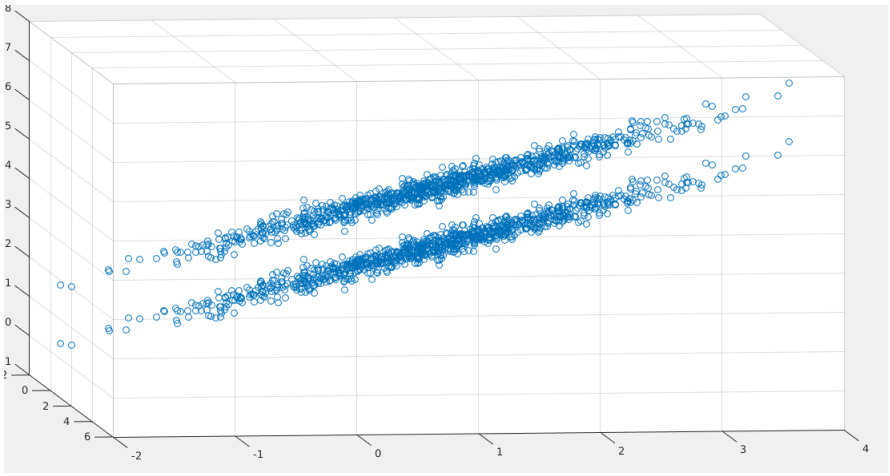

Figure 2: 3D plot of the dataset

$$Confusion\ Matrix = \begin{bmatrix} 150 & 50 \\ 42 & 158 \end{bmatrix} \tag{1}$$

Class 1:

$$Precision = \frac{150}{150 + 42} = 0.78 \tag{2}$$

$$Recall = \frac{150}{150 + 50} = 0.75 \tag{3}$$

$$f - measure = 2 * \frac{Precision * Recall}{Precision + Recall} = 0.76 \tag{4}$$

Class 2:

$$Precision = \frac{158}{158 + 50} = 0.76 \tag{5}$$

$$Recall = \frac{158}{158 + 42} = 0.77 \tag{6}$$

$$f - measure = 2 * \frac{Precision * Recall}{Precision + Recall} = 0.76 \tag{7}$$

- Performing LDA to the same data set shows a remarkable increase in the efficiency of classification. The per-class precision and recall of each class equal to 1

$$Confusion\ Matrix = \begin{bmatrix} 200 & 0 \\ 0 & 200 \end{bmatrix} \tag{8}$$

Class 1:

$$Precision = 1 \tag{9}$$

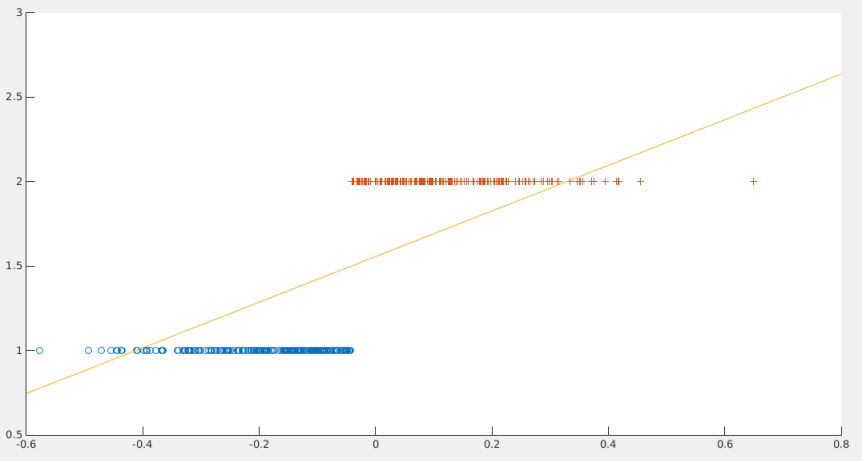

Figure 3: Projected feature space with the decision boundary

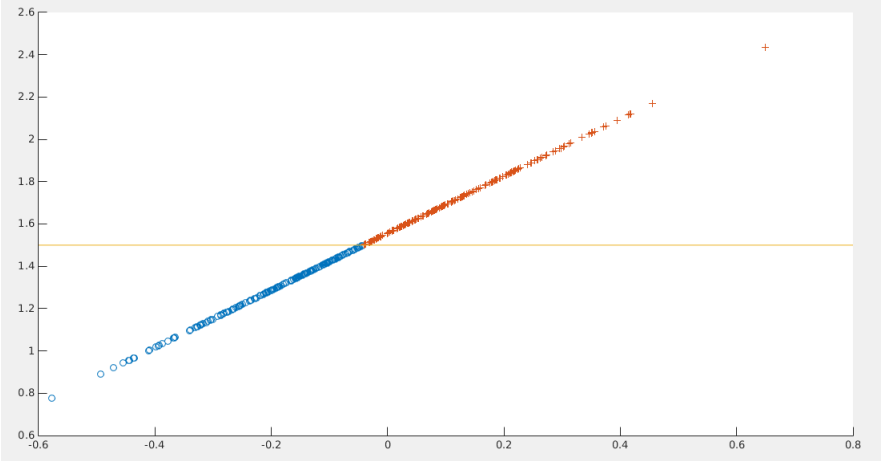

Figure 4: Projected feature space with the threshold

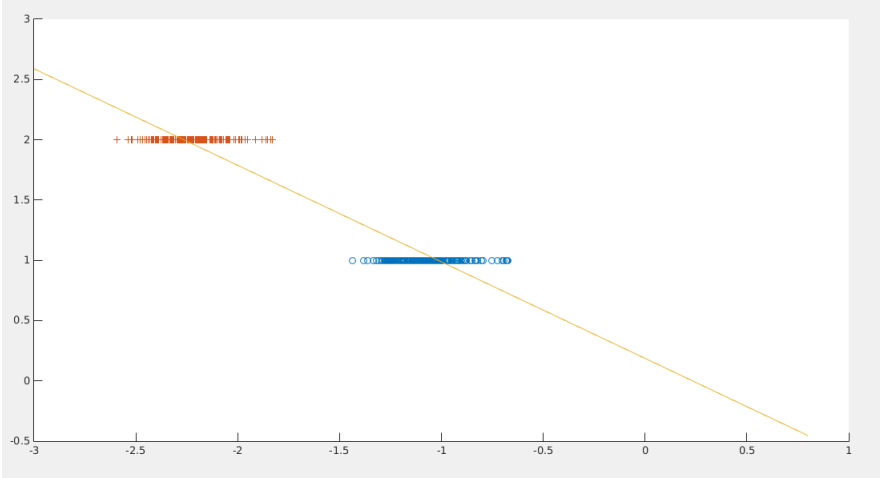

Figure 5: Projected feature space with the decision boundary

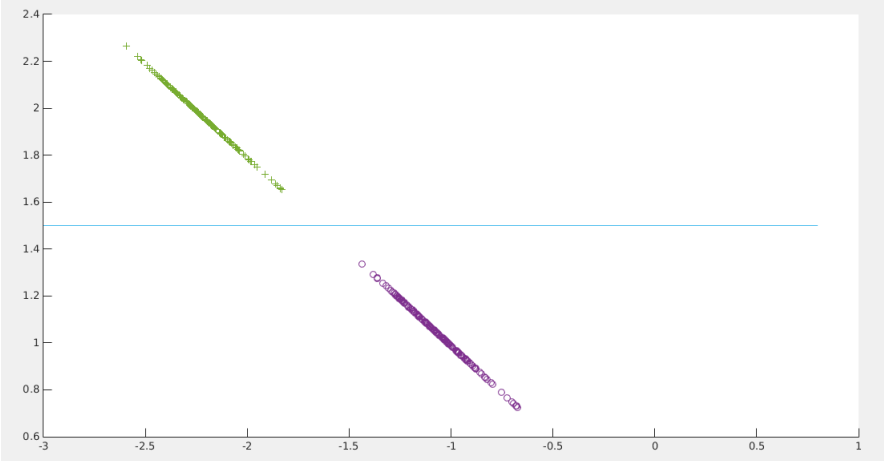

Figure 6: Projected feature space with the threshold

$$Recall = 1 \tag{10}$$

$$f - measure = 1 \tag{11}$$

Class 2:

$$Precision = 1 \tag{12}$$

$$Recall = 1 \tag{13}$$

$$f - measure = 1 \tag{14}$$

Kernelfunction	gamma	coef0	cost	Accuracy
Linear	0.01	0	20	65.06
Polynomial (degree 2)	0.1	10	1	65.06
Radial Basis Function	0.1	0	1	68.67
Sigmoid	0.01	-1	1	66.26

Table 2: Accuracy in knn-classifier

4 Support Vector Machines

A 96 dimensional feature vector is employed for the classification of the images into the 4 categories forest, insidecity, coast, mountain. Four kernel functions are used for the purpose of classification and the kernal parameters and accuracy are tabulated in Table 2.

The training models generated by the each kernel function are provided in the zip file named as model0, model1, model2, model3 respectively for Linear, Polynomial, Radial Basis Function and Sigmoid functions.

5 Bayesian Parameter Estimation

- **Multinomial likelihood:**The 5 confusion matrices obtained through cross validation are given below.

$$Confusion\ Matrix1 = \begin{bmatrix} 119 & 2 \\ 5 & 95 \end{bmatrix} \tag{15}$$

$$Confusion\ Matrix2 = \begin{bmatrix} 184 & 1 \\ 2 & 144 \end{bmatrix} \tag{16}$$

$$Confusion\ Matrix3 = \begin{bmatrix} 243 & 2 \\ 5 & 191 \end{bmatrix} \tag{17}$$

$$Confusion\ Matrix4 = \begin{bmatrix} 303 & 3 \\ 6 & 238 \end{bmatrix} \tag{18}$$

$$Confusion\ Matrix5 = \begin{bmatrix} 367 & 2 \\ 4 & 287 \end{bmatrix} \tag{19}$$

Average per class Precision

Legit mails = 99.09%
Spam mails = 97.44%

Average per class Recall

Legit mails = 97.97%
Spam mails = 98.85%

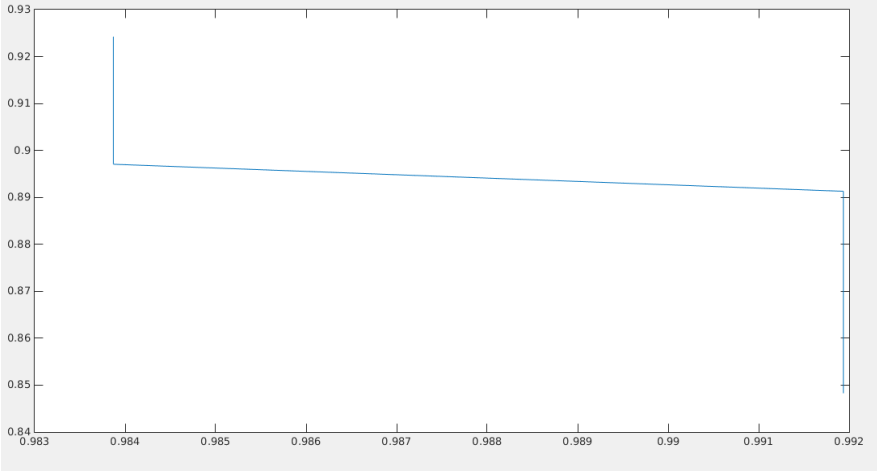

Figure 7: P-R curve for beta prior

- **Bernoulli likelihood:** The 5 confusion matrices obtained through cross validated are given below.

$$Confusion\ Matrix1 = \begin{bmatrix} 123 & 1 \\ 24 & 73 \end{bmatrix} \tag{20}$$

$$Confusion\ Matrix2 = \begin{bmatrix} 186 & 0 \\ 12 & 133 \end{bmatrix} \tag{21}$$

$$Confusion\ Matrix3 = \begin{bmatrix} 248 & 0 \\ 12 & 181 \end{bmatrix} \tag{22}$$

$$Confusion\ Matrix4 = \begin{bmatrix} 309 & 0 \\ 17 & 224 \end{bmatrix} \tag{23}$$

$$Confusion\ Matrix5 = \begin{bmatrix} 369 & 2 \\ 17 & 272 \end{bmatrix} \tag{24}$$

Average per class Precision

Legit mails = 92.68%
Spam mails = 99.58%

Average per class Recall

Legit mails = 99.73%
Spam mails = 89.57%

- **Beta Prior :**The optimum value of α and β which give the best value of $Precision * Recall$ is $\alpha = 9$ and $\beta = 2$.
The P-R curve for the following in given in the figure