CS584 Assignment 2: Report

Gady Agam
Department of Computer Science
Illinois Institute of Technology
March 10, 2016

Abstract

This report is for assignment in CS584. The problems that we confront Classification problem by different algorithms viz. Gaussian Discriminant Analysis - 1- Feature 2 Class, Gaussian Discriminant Analysis - N- Features 2 Class, Gaussian Discriminant Analysis- N-Feature k Class, Naïve Bayes – Bernoulli distribution, Naïve Bayes –Binomial distribution. Their performance in terms of mean squared error, confusion matrix is evaluated.

Problems Solved:

1. 1D 2-Class Gaussian Discriminant Analysis
    a. To do this analysis, I have chosen the credit_Score data set from UCI Repository. The data has been cleaned to omit the samples with missing values and equal amount of samples for both the classes (i.e. Class 0 and Class 1) have been used as Training Dataset.
    b.

The Model Parameters are μ and σ

Compute the mean and variance of samples with y=0 and y= 1 separately on X_Train

$$\mu_j = 1/m_j \sum_{i=1}^{m_j} X^{(i)}$$

$$\sigma_j^2 = 1/m_j \sum_{i=1}^{m_j} (X^{(i)} - \mu_j)^2$$

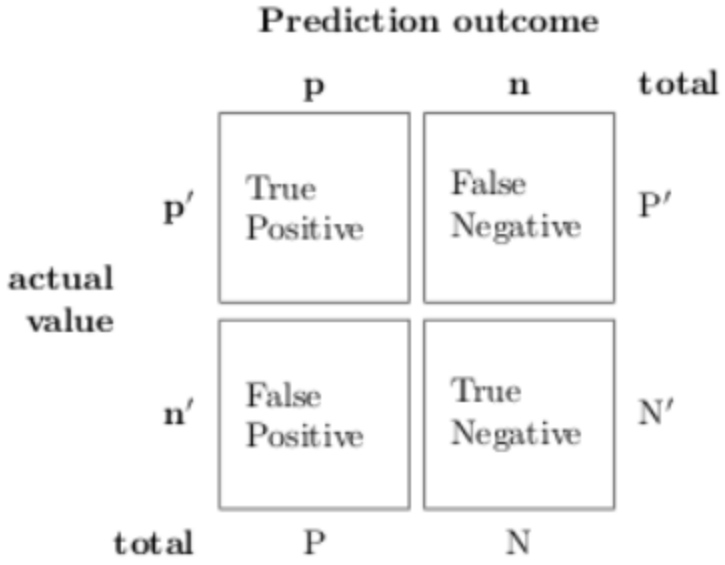
    c. From the Model Parameters we can predict the labels for the new Test Data

Univariate Gaussian Discriminant

$$g_j(X) = -log(\sigma_j) - (X - \mu_j)^2/\sigma_j^2 + log(\alpha_j)$$

using,

    d. We compute the confusion matrix using the predicted and actual classes.
    Confusion Matrix:



Mean Squared Error 0.438735177866

Evaluation Measures from Confusion Matrix for label 0.0 :
Accuracy : 0.561264822134
Recall : 0.302083333333
False Negative : 0.697916666667
Precision : 0.397260273973
False Positive : 0.28025477707
True Negative : 0.71974522293
F Square : 0.343195266272

Evaluation Measures from Confusion Matrix for label 1.0 :
Accuracy : 0.561264822134
Recall : 0.71974522293
False Negative : 0.28025477707
Precision : 0.627777777778
False Positive : 0.697916666667
True Negative : 0.302083333333
F Square : 0.670623145401

2. nD 2-Class Gaussian Discriminant Analysis

a. To do this analysis, I have chosen the credit_Score data set from UCI Repository. The data has been cleaned to omit the samples with missing values and equal

amount of samples for both the classes (i.e. Class 0 and Class 1) have been used as Training Dataset. The Training Set has 6 continuous features.

b. Model Parameter for multi feature Gaussian distribution.

$$\mu_j = 1/m_j \sum_{i=1}^{m_j} I(Y^j = i) X^{(i)}$$

$$\sum_j = 1/m_j \sum_{i=1}^{m_j} (X^{(i)} - \mu_j)(X^{(i)} - \mu_j)^T$$

c. From the computed model parameter mu and sigma, we can predict the labels for any new X. d.

$$g_j(X) = -log(|\sum_j|) - 1/2(X - \mu_j)^T \sum^{-1} (X - \mu_j) + log(\alpha_j)$$

d. Performance Evaluation:

Mean Squared Error 0.237154150198

Evaluation Measures from Confusion Matrix for label 0.0 :

Accuracy : 0.762845849802
Recall : 0.552083333333
False Negative : 0.447916666667
Precision : 0.757142857143
False Positive : 0.108280254777
True Negative : 0.891719745223
F Square : 0.638554216867

Evaluation Measures from Confusion Matrix for label 1.0 :

Accuracy : 0.762845849802
Recall : 0.891719745223
False Negative : 0.108280254777
Precision : 0.765027322404
False Positive : 0.447916666667
True Negative : 0.552083333333
F Square : 0.823529411765

Area Under the PR Curve using Trapezium rule 7.34445326274

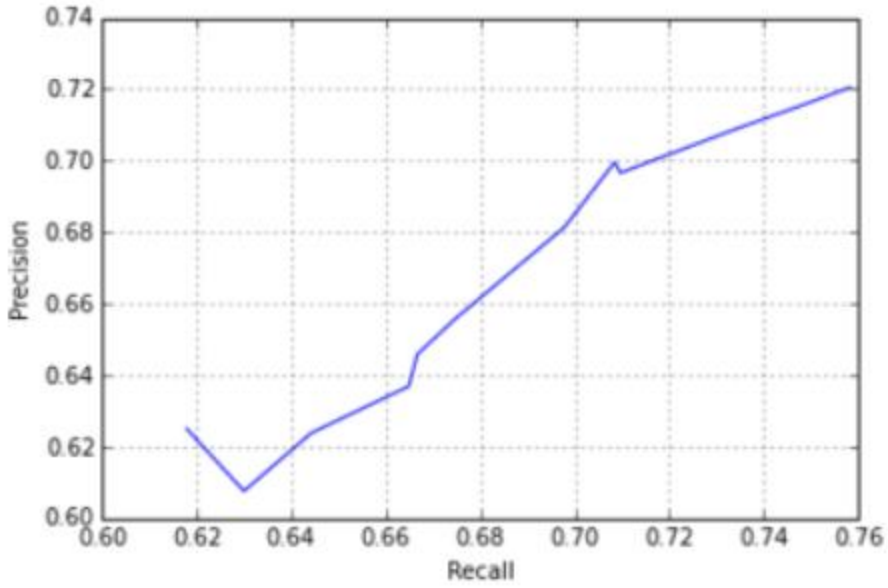


Fig1. Precision Recall Curve – using averages.

3. a.

nD k-Class Gaussian Discriminant Analysis

Compute the mean and variance of samples with y=0 and y= 1 separately on X_Train

$$\mu_j = 1/m_j \sum_{i=1}^{m_j} I(Y^j = i) X^{(i)}$$

$$\sum_j = 1/m_j \sum_{i=1}^{m_j} (X^{(i)} - \mu_j)(X^{(i)} - \mu_j)^T$$

---

b. From the computed model parameter mu and sigma, we can predict the labels for any new X.

$$g_j(X) = -log(|\sum_j|) - 1/2(X - \mu_j)^T \sum^{-1} (X - \mu_j) + log(\alpha_j)$$

c. Performance Evaluation:

Mean Squared Error 1.36082474227

Evaluation Measures from Confusion Matrix for label 0.0 :

Accuracy : 0.701030927835
Recall : 0.702127659574
False Negative : 0.297872340426

Precision : 0.6875
False Positive : 0.3
True Negative : 0.7
F Square : 0.694736842105

Evaluation Measures from Confusion Matrix for label 1.0 :

Accuracy : 0.711340206186
Recall : 0.8125
False Negative : 0.1875
Precision : 0.833333333333
False Positive : 0.764705882353
True Negative : 0.235294117647
F Square : 0.822784810127

Evaluation Measures from Confusion Matrix for label 2.0 :

Accuracy : 0.752577319588
Recall : 0.853658536585
False Negative : 0.146341463415
Precision : 0.853658536585
False Positive : 0.8
True Negative : 0.2
F Square : 0.853658536585

Evaluation Measures from Confusion Matrix for label 3.0 :

Accuracy : 0.835051546392
Recall : 0.894117647059
False Negative : 0.105882352941
Precision : 0.915662650602
False Positive : 0.583333333333
True Negative : 0.416666666667
F Square : 0.904761904762

Evaluation Measures from Confusion Matrix for label 4.0 :

Accuracy : 0.969072164948
Recall : 1.0
False Negative : 0.0
Precision : 0.969072164948
False Positive : 1.0
True Negative : 0.0
F Square : 0.984293193717

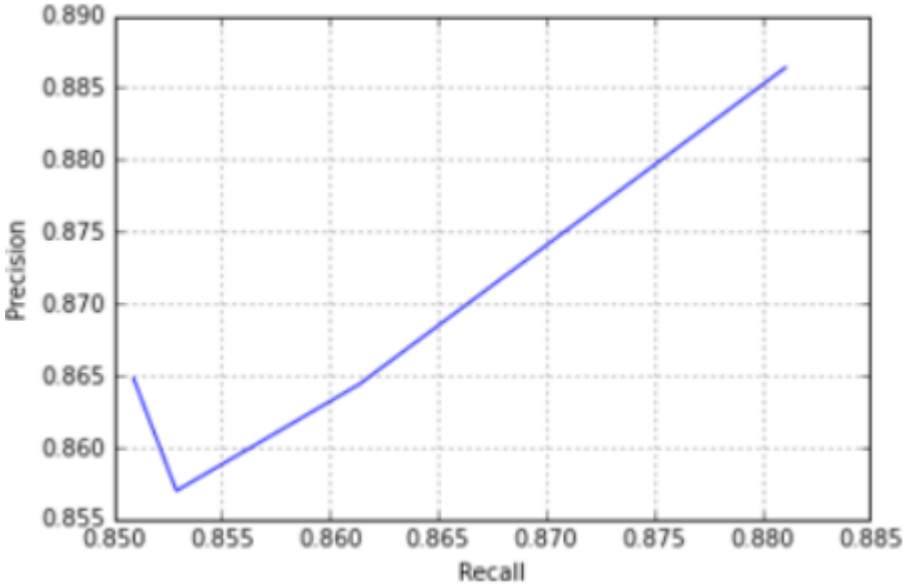


Fig2. Precision – Recall curve.

4. Naive Bayes with Bernoulli Features.
   a. The credit_Score data set of UCI, has been used in this experiment. It has two classes (ie. Good Credit: 1 Bad Credit: 0). The dataset contains text categorical values as 0s and 1s.

   b. Model Parameter:

$$\alpha_{p|y=1} = 1/m_j \sum_{i=1}^{m_j} X_p^i$$

$$Prior_i = 1/m_j \sum_{i=1}^{m_j} I(Y^j = i)$$

   c.

Membership Function

$$g_l(X) = \sum_{j=1}^{n} log(\binom{P}{X_j} \alpha_{j/y=l}^{X_j} (1 - \alpha_{j/y=l})^{p - X_j}) + log(\alpha_l)$$

Classification: $\hat{y} = \arg\max_l g_l(X)$

   d. Performance Evaluation:
   Mean Squared Error 0.399209486166

Evaluation Measures from Confusion Matrix for label 0. :

Accuracy : 0.600790513834
Recall : 0.202380952381
False Negative : 0.797619047619
Precision : 0.333333333333
False Positive : 0.201183431953
True Negative : 0.798816568047
F Square : 0.251851851852

Evaluation Measures from Confusion Matrix for label 1.0:

Accuracy : 0.600790513834
Recall : 0.798816568047
False Negative : 0.201183431953
Precision : 0.668316831683
False Positive : 0.797619047619
True Negative : 0.202380952381
F Square : 0.727762803235

5. Naive Bayes – Binomial Features.
   a. Used a new Text based Dataset – Spmbase.csv , it contains the frequency of the words , as a part of Data Cleaning , the frequencies are converted into whole numbers and the features. The two classes are 0 (not spam) and 1(spam).

   b. Model Parameter

$$\alpha_{j/y=l} = (\sum_{i=1}^{m} I(Y^i = l) X_j^i) / (\sum_{i=1}^{m} I(Y^i = l) P^i)$$

Membership Function

$$g_l(X) = \sum_{j=1}^{n} log(\binom{P}{X_j} \alpha_{j/y=l}^{X_j} (1 - \alpha_{j/y=l})^{P - X_j}) + log(\alpha_l)$$

Classification: $\hat{y} = \arg\max_l g_l(X)$

   c.
   d. Performance Evaluation:

Mean Squared Error 0.15

Evaluation Measures from Confusion Matrix for label 0.0:

Accuracy : 0.85
Recall : 0.626666666667
False Negative : 0.373333333333
Precision : 0.959183673469
False Positive : 0.016
True Negative : 0.984
F Square : 0.758064516129

Evaluation Measures from Confusion Matrix for label 1.0:

Accuracy : 0.85
Recall : 0.984
False Negative : 0.016
Precision : 0.814569536424
False Positive : 0.373333333333
True Negative : 0.626666666667
F Square : 0.891304347826