ARJUN KOLUMAM RAMAKRISHNAN

AKOLUMA

ECE 8560

TAKEHOME #1

Introduction

This report explains in detail about the design and implementation of a Bayesian classifier. There are two datasets given - training data and test data, each with 15000 feature vectors having 4 features each. The training data has been used to develop the Bayesian classifier and then it is tested on the test data. The training data we are given is labeled and the first 5000 feature vectors correspond to the first class, the next 5000 feature vectors correspond to the second class and so on. Overall, there are three classes – class 1, class 2 and class 3.

Design of Classifier

Determining the distribution of the data set

A histogram plot of the training data and test data is generated in MATLAB using the function **histogram** and it can be inferred that it is Gaussian distribution.

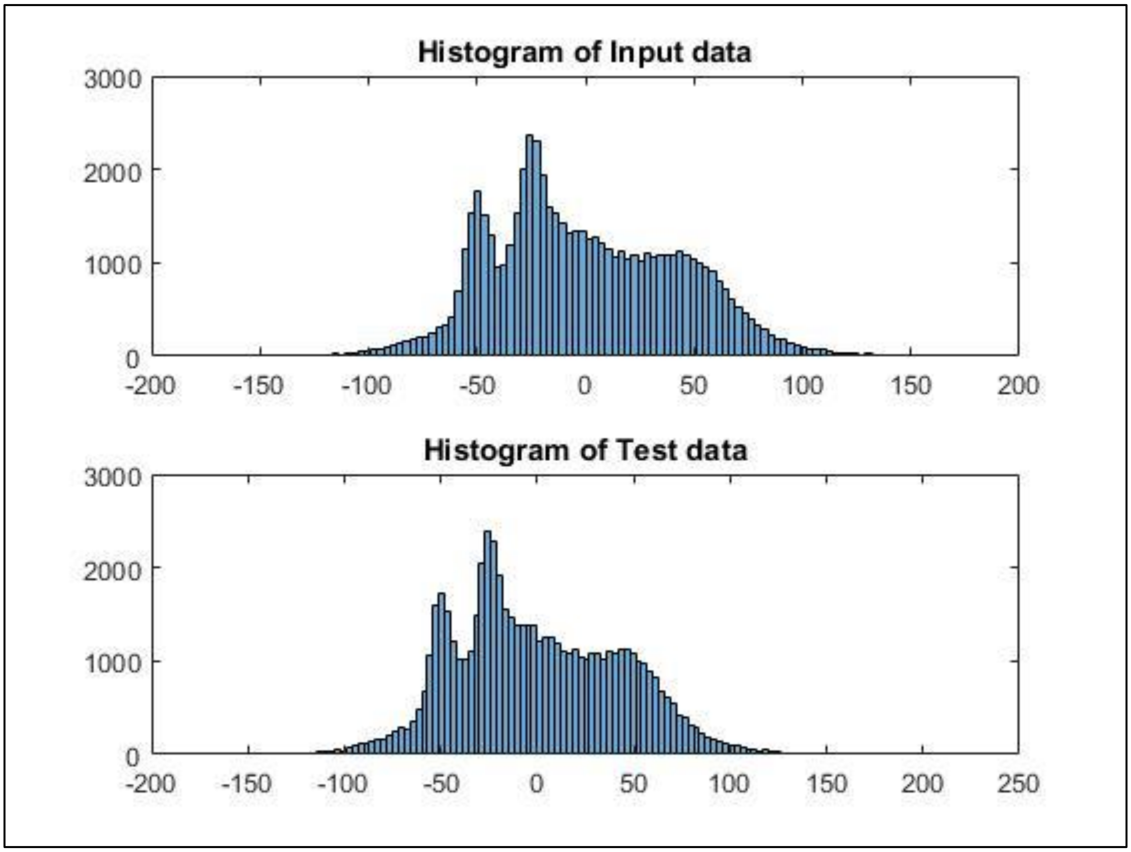

Fig 1.Histogram generated in MATLAB

Also Jarque-Bera test which specifies the number of feature vectors belonging to normal distribution was conducted and it showed that 14257 vectors out of 15000 vectors belong to normal distribution. Hence this also proves that the distribution is Gaussian.

Mean and Covariance

The discriminant function of the Gaussian model requires mean and covariance of the input data which are calculated using **mean** and **cov** functions in MATLAB. These functions use the formula,

$$\mu = \sum_{i=0}^{n} \frac{x_i}{n}$$

$$\sum \quad = E[(X - E[X])(X - E[X])^T]$$

For mean and covariance respectively. Since the input data is transposed, the mean is transposed to obtain the proper form.

Apriori Probability of each class

The apriori probability of each class is calculated as follows,

$$P(w_i) = \frac{no. of\ feature\ vectors\ in\ class\ i}{Total\ number\ of\ features\ in\ input}$$

Using the above formula we get probability of class 1, class 2 and class 3 as,

$$P(w_1) = P(w_2) = P(w_3) = \frac{1}{3}$$

Since the apriori probabilities are equal, it is not factored in the estimation of discriminant function.

Implementation of Classifier

Once the Bayesian classifier is trained using training data, it is then used to classify test data. The classification is based on the value of discriminant function and is explained below in detail. The classified test data is then written into a “.txt” file using **fopen** and **fclose** functions in MATLAB.

Discriminant Function

The mean and covariance for the feature vectors in training data is calculated for each class separately and then plugged into the formula for discriminant function which is given by,

$$g_i(x) = -\frac{1}{2}\left(\underline{x}-\underline{\mu}_i\right)^T \textstyle\sum_i{}^{-1}\left(\underline{x}-\underline{\mu}_i\right) - \left(\frac{d}{2}\right)\log(2\pi) - \frac{1}{2}log\left|\textstyle\sum_i\right| + \log\{P(w_i\,)\}$$

Where,

μ= Mean of each class

∑= Covariance of each class

d= Number of features i.e. 4 in this case

$\mathrm{P(w_i)} = \mathrm{probability\ of\ class\ i}$

$\underline{x}$=feature vector

i=1, 2, 3

Since the probability of all three classes are the same i.e. 0.33(calculated below) it is not included in the discriminant function and also the (d/2) bias is removed while calculating the discriminant function. Also, since the input data consists of a feature vector transposed per line, it is again transposed before plugging into the formula shown above. The discriminant function is calculated for all the three classes and the feature vector is assigned to the class which has the largest value. An error matrix has also been constructed to visually see how well the classifier works.

	Class 1	Class 2	Class 3
Class 1	4539	318	143
Class 2	577	4389	34
Class 3	236	50	4714

Table1. Error Matrix

The above table represents that out of the 5000 feature vectors in class 1,only 4539 have been classified as class 1 and the remaining have been classified erroneously. Similarly in class 2 and class 3, 4389 and 4714 vectors have been classified properly.

Probability of Error

The probability of error is calculated using the following formula,

$$\begin{aligned} P(error) = \quad & P(choose\ class\ 1\ and\ x\ actually\ from\ class\ 2) \\ & + P(choose\ class\ 1\ and\ x\ actually\ from\ class\ 3) \\ & + P(choose\ class\ 2\ and\ x\ actually\ from\ class\ 1) \\ & + P(choose\ class\ 2\ and\ x\ actually\ from\ class\ 3) \\ & + P(choose\ class\ 3\ and\ x\ actually\ from\ class\ 1) \\ & + P(choose\ class\ 3\ and\ x\ actually\ from\ class\ 2) \end{aligned}$$

The probability of error using the above formula was found to be 0.0905 (or) 9.05%

Conclusion

We can see that our implementation of Bayesian Classifier was successful and has a very low probability of error of 0.0905 or 9.05 % and a success rate of 0.9095 or 90.95 %.