CS 584: Machine Learning
Assignment-2
Generative Learning

Name: Priyank P Shah Date:03/14/2016
CWID: A20344797 Term:Spring'16

1. Problem Statement :

Objective to develop this project is to classify the datasets using the generative learning approach. Project consist of 5 sub-problems, which are as belows.

1) 1D 2 Class Gaussian Discriminant Analysis:
- Find the required model parameters and compute the membership function.
- Test & train data with the cross-validation(K-Fold)
- Compute the confusion matrix and find Accuracy, Precision, Recall and F-measure.

2) nD 2 Class Gaussian Discriminant Analysis:
- Find the required model parameters and compute the membership function.
- Test & train data with the cross-validation(K-Fold)
- Compute the confusion matrix and find Accuracy, Precision, Recall and F-measure.
- Plot precision, recall curve and measure the area under the curve.

3) nD K Class Gaussian Discriminant Analysis:
- Find the required model parameters and compute the membership function.
- Test data with the cross-validation(K-Fold)
- Compute the confusion matrix and find Accuracy, Precision, Recall and F-measure.

4) Naive Bayes with Bernoulli features :
- Find the required model parameters and compute the membership function on binary nD features where dataset needs to be derived from text document.
- Test data with the cross-validation(K-Fold)
- Compute the confusion matrix and find Accuracy, Precision, Recall and F-measure.

5) Naive Bayes with Binomial features :
- Using Maximum Likelihood derive the model parameters estimate equation.
- Compute the membership function on binary nD features where dataset needs to be derived from text document.
- Test data with the cross-validation(K-Fold).
- Compute the confusion matrix and find Accuracy, Precision, Recall and F-measure.

2. Proposed Solution:

Algorithm:

1D 2-Class Gaussian Discriminant Analysis:

- from the dataset, extract data and place accordingly in array.
- For each Cross validation State find model parameters such as mean, sigma and alpha for dataset of each class. Mean will be calculated by doing sum of each feature for each class and then dividing it by the total number of cases.
- For Sigma calculation take difference of each feature value and mean of particular feature . Take square function of this difference and take sum for all cases, divide it by total number of cases.
- Once sigma is calculated, Count membership function with the formula which consist of several features as log of sigma, alpha. In this case we will only get 1 value of mean and sigma for each class.
- Once membership function calculate value of each function, take difference of two class's case value, if the difference is positive then it indicated that it belongs to class 1 else it will be from class 2, that is discriminant function.
- Confusion matrix is created from actual value of class and predicted value of class from that if actual value and predicted value both matches than it True Positive else it is True Negative, If Algorithm predict value as false and actual value is false then it is True False it it False Negative.
- This Confusion matrix helps to determine the value of precision, recall. From precision and recall we can compute F1 measure and Accuracy.

nD 2-Class Gaussian Discriminant Analysis:

- from the dataset, extract data and place accordingly in array.
- For each Cross validation State find model parameters such as mean, sigma and alpha for dataset of each class. Mean will be calculated by doing sum of each feature for each class and then dividing it by the total number of cases.
- For Sigma calculation take difference of each feature value and mean of particular feature . Take square function of this difference and take sum for all cases, divide it by total number of cases.
- Once sigma is calculated, Count membership function with the formula which consist of several features as log of sigma, alpha. If we will get mean of (1,4) matrix for each class, and sigma matrix of (4,4) shape.
- Once membership function calculate value of each function, take difference of two class's case value, if the difference is positive then it indicated that it belongs to class 1 else it will be from class 2, that is discriminant function.
- Confusion matrix is created from actual value of class and predicted value of class from that if actual value and predicted value both matches than it True Positive else it is True Negative, If Algorithm predict value as false and actual

value is false then it is True False it it False Negative.

- This Confusion matrix helps to determine the value of precision, recall. From precision and recall we can compute F1 measure and Accuracy.

nD K-Class Gaussian Discriminant Analysis:

- Extract the data from the dataset into their respective class.
- For each cross-validation fold compute the mean and sigma value. In K class and n Dimensions, there are K different mean having dimension of (1,n) and sigma having dimension of (n,n); where n denotes the number of features.
- Once mean and sigma are calculated find the value of membership function from the test data.
- Check for the bigger value using discriminant function to identify the proper class of the data.
- Compare the data obtain from discriminant function with the given value of class, once predicted value of class if obtained compute confusion matrix.
- From the confusion matrix we will get the value of precision, recall, F1 measures and accuracy.

Naive Bayes with Bernoulli features:

- This algorithm is used to classify document into the discrete category from the given set of different work appearance.
- We calculate the word frequency and identify the class of the document.
- Compute the value of Alpha from the training dataset, that is the word count for a word appeared in all of the documents given. We will get Alpha with the dimension of (1,n) where n denotes the given different word set.
- We will also calculate the prior that the odd of word appearance against the total number of documents.
- Compute the membership function value g(x) from the testing dataset.
- Using the discriminant function we will identify the likelihood of a document, from which category it belongs, which is the predicted classification of the document.
- Compare this predicted value with the actual value set and find the confusion matrix. Obtain value of precision, recall, F1 measures and accuracy.

Naive Bayes with Binomial features:

- Instead of classify document just the appearance of the word in document, count of the word appear in the document also effects a decision a lot, so instead of using discrete value of word appearance, we will check the word count which classify the document in more than two classes.
- Count model parameter Alpha and prior from the training dataset, which is frequency of particular word appeared in document.

- Pass model parameter to compute membership function and discriminant function which results into predicted class value of member.
- Compare predicted class value and actual class value to determine confusion matrix.
- Compute Precision, Recall, Accuracy and F1 Measures from the achieved confusion matrix.
-

3. Implementation Details:

Instructions:

- Before running program file, put the datasets into the same folder where the program file resides.
- I have developed date-set oriented codes so It may require some modification to run different dataset, while core functionality will remain same for any dataset.
- At the end of the program user will be shown output as Confusion matrix, Accuracy, Precision, Recall and F-1 Measures.
- For nD 2-class precision recall curve graph will be automatically generated..

Design Issues:

- While creating confusion matrix , I had issues with comparing values especially in K-class implantation, in order to find the value of class which has the maximum value for discriminant function.
- During the development of Naive Bayes – Binomial features, due to higher dimension and big values, it took time to find optimized way to compute membership function. I have used gmpy library to find combination(ncr).

4. Results and Discussion:

1. 1D 2-Class Gaussian Discriminant Analysis:

Database: `data_banknote_authentication.txt`

Result:

	Precision	Recall	F1 Measures
Class 0	0.7	0.778	0.737
Class 1	0.5	0.4	0.444

Accuracy: 64.28 %
Confusion Matrix: [7 2]
[3 2]

I observed that due to one dimensional feature, accuracy achieved is lower than n dimensional feature

set.

2. nD 2-Class Gaussian Discriminant Analysis:

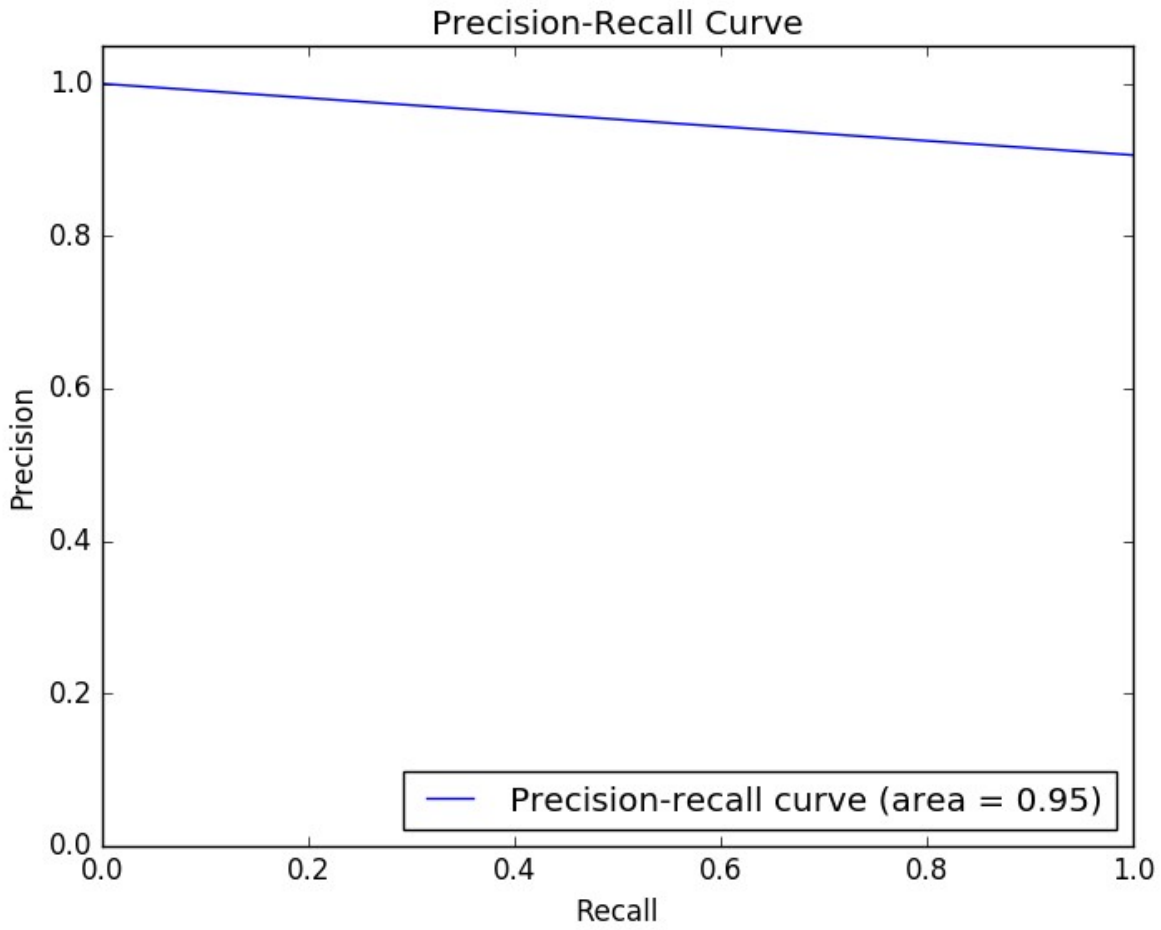
Database: `data_banknote_authentication.txt`

Result:

	Precision	Recall	F1 Measures
Class 0	1.0	0.9173	0.9568
Class 1	0.9063	1.0	0.9508

Accuracy: 95.40 %
Confusion Matrix: [699 63]
[0 710]

by using multiple dimension, I have achieved higher accuracy. Precision vs Recall graph is as below:



3. nD K-Class Gaussian Discriminant Analysis:

Database: `iris.data.txt`

Result:

	Precision	Recall	F1 Measures
Class 0	1.0	1.0	1.0
Class 1	1.0	0.9166	0.9565
Class 2	0.9230	1.0	0.96

Accuracy: 96.66 %
Confusion Matrix: [6 0 0]
[0 11 1]
[0 0 12]

From this result we can conclude that, Accuracy I have achieved is equivalent to the accuracy achieved by inbuilt function for this datasets.

4. Naive Bayes with Bernoulli features:

Database: `spambase.data.txt`

Result:

	Precision	Recall	F1 Measures
Class 0	0.90	0.868	0.884
Class 1	0.804	0.849	0.826

Accuracy: 86.1%
Confusion Matrix: [244 37]
[27 152]

In Bernoulli features implementation, we are converting features into binary features, thus we are ignoring the frequency of any particular word(feature), that is important to classify the class in appropriate category. Hence we are achieving less accuracy compare to Binomial approach.

5. Naive Bayes with Binomial features:

Database: `spambase.data.txt`

Result:

	Precision	Recall	F1 Measures
Class 0	0.74	0.899	0.8111
Class 1	0.579	0.306	0.4

Accuracy: 71.3%
Confusion Matrix: [142 16]
[50 22]

In Binomial features , we are considering actual count of words for predicting its class definition. Dataset contains the word frequency instead of actual count, hence I have assume document length as 100, which is not accurate. To achieve higher frequency, one need to take Document length at least of 10,000.
Document length is trade off for ncr computation, that is Accuracy vs. NCR combination calculation time. Hence I have achieved less accuracy compare to accuracy that we can get with big document length.

5. Reference:

- http://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=&area=&numAtt=&numIns=&type=text&sort=nameUp&view=table
- http://www.text-analytics101.com/2014/10/computing-precision-and-recall-for.html
- http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note07-2up.pdf
- https://pypi.python.org/pypi/gmpy2