CS 584

MACHINE LEARNING

Assignment 2:

Generative Learning

Nivedita Kalele

A20329966

Generative Learning:

Introduction:

Generative classifier model uses Bayes' rule to calculate P(x|y) for x inputs and y labels and picks most likely label. Generative model will generate new samples using the given joint probability. This algorithm transfers P(x,y) into P(x|y) and uses it for classification.

Iris Dataset (continuous dataset)

This is the multivariate dataset with 4 features and 150 samples in which 3 classes are there. Each class has 50 samples. The classification using generative model is done for these 3 classes. This dataset is used for first 3 questions of the assignment.

1] 1D 2 class Gaussian Discriminant Analysis

a) Loading the dataset:

The dataset is taken from the sklearn datasets. The first feature is take in the x as input and the labels are taken in the y. As 2 class dataset was required and iris is 3 class dataset so, only first 100 samples are considered as input.

b) Estimating the model parameters and computing the discriminant function:

Mean and the variance are the parameters we need to estimate for calculating the membership function for each class.

Membership function:

$$g(x) = \frac{1}{\sqrt{2}\pi\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \log(prior\_probability)$$

This function tells about the membership of input x to class y.

So, by using discriminant function:

d(x)= g1(x)-g2(x)

We can tell which class the input x belongs to. Maximum value from the two membership function values will be the respective class value. Hence, if d(x) is positive, class will be 1 as g1 is maximum otherwise 2.

In this way classification is done using discriminant function.

c) Calculating Confusion matrix, precision, recall, F-measure and accuracy:

Performance with 75% training size
Confusion Matrix: [[12 0]
[2 11]]
Accuracy: 0.92
Precision: [1.0, 0.8461538461538461]
Recall: [0.8571428571428571, 1.0]
F-measure: [0.923076923076923, 0.9166666666666666]

Performance with 10 fold (For Some folds)
Mean Squared Error: 0.122222222222

Accuracy: 0.877777777778
Precision: [0.8536585365853658, 0.8979591836734694]
Recall: [0.875, 0.88]
F-measure: [0.8641975308641976, 0.888888888888889]

Accuracy: 0.888888888889
Precision: [0.875, 0.9]
Recall: [0.875, 0.9]
F-measure: [0.875, 0.9]

Accuracy: 0.9
Precision: [0.8604651162790697, 0.9361702127659575]
Recall: [0.925, 0.88]
F-measure: [0.891566265060241, 0.9072164948453608]

Accuracy: 0.9
Precision: [0.9183673469387755, 0.8780487804878049]
Recall: [0.9, 0.9]
F-measure: [0.9090909090909091, 0.888888888888889]

Accuracy: 0.9
Precision: [0.9183673469387755, 0.8780487804878049]
Recall: [0.9, 0.9]
F-measure: [0.9090909090909091, 0.888888888888889]

The confusion matrix helps in understanding the performance of the classification. Each row of the matrix represents the predicted labels and the columns represents the actual labels. Hence, we can get know easily how many inputs are misclassified.

Similarly, precision, recall, F-measure and accuracy are the parameters that tells us about the performance of algorithm. By looking at the accuracies above we can say that the data was classified better with training size 75% and also for some of the folds in 10 fold cross validation.

2] <u>nD 2 class Gaussian Discriminant Analysis</u>

a) Loading the dataset: (4 features)

The dataset is taken from the sklearn datasets. All the features are taken in the x as input and the labels are taken in the y. As 2 class dataset was required and iris is 3 class dataset so, only first 100 samples are considered as input.

b) Estimating the model parameters and computing the discriminant function:

Mean and the variance are the parameters we need to estimate for calculating the membership function for each class.

Membership function:

$$g(x) = \frac{1}{(2\pi)^n/2\ *(|cov|)\text{^}1/2} \exp\left(-\frac{(x-\mu)^T *inv(cov)*(x-\mu)}{2}\right)$$

This function tells about the membership of input x to class y.

So, by using discriminant function:

d(x)= g1(x)-g2(x)

We can tell which class the input x belongs to. Maximum value from the two membership function values will be the respective class value. Hence, if d(x) is positive, class will be 1 as g1 is maximum otherwise 2.

In this way classification is done using discriminant function.

c) Calculating Confusion matrix, precision, recall, F-measure and accuracy:

<u>Performance with 75% training size</u>
Accuracy: 1.0

Confusion Matrix: [[14 0]
[0 11]]
Precision: [1.0, 1.0]
Recall: [1.0, 1.0]
F-measure: [1.0, 1.0]

Performance with 10 fold (For one of the folds)
Accuracy: 1.0
Confusion Matrix: [[40 0]
[0 50]]
Precision: [1.0, 1.0]
Recall: [1.0, 1.0]
F-measure: [1.0, 1.0]

The confusion matrix helps in understanding the performance of the classification. Each row of the matrix represents the predicted labels and the columns represents the actual labels. Hence, we can get know easily how many inputs are misclassified.

Similarly, precision, recall, F-measure and accuracy are the parameters that tells us about the performance of algorithm. By looking at the accuracies above we can say that the data was classified better with both for training size 75% and also for all the folds in 10 fold cross validation.

3] nD n class Gaussian Discriminant Analysis

a) Loading the dataset: (4 features,3 classes)

The dataset is taken from the sklearn datasets. All the features are taken in the x as input and the labels are taken in the y. All the samples are considered as input.

b) Estimating the model parameters and computing the discriminant function:

Mean and the variance are the parameters we need to estimate for calculating the membership function for each class.

Membership function:

$$g(x) = \frac{1}{(2\pi)^n/2 \ * (|cov|)\char`^1/2} \exp\left(-\frac{(x-\mu)^T *inv(cov)*(x-\mu)}{2}\right)$$

This function tells about the membership of input x to class y.

So, by using discriminant function:

$d(x)= (g1(x)-g2(x))+(g2(x)-g3(x))+(g1(x)-g3(x))$

We can tell which class the input x belongs to. Maximum value from the three membership function values will be the respective class value. Hence, if g1 is maximum, class 1 will be selected, if g2 is maximum class 2 will be selected otherwise 3 .In this way classification is done using discriminant function.

c) Calculating Confusion matrix, precision, recall, F-measure and accuracy:

Performance with 75% training size

Confusion Matrix: [[15 0 0]
[0 11 0]
[0 0 12]]

Precision: [1.0, 1.0, 1.0]

Recall: [1.0, 1.0, 1.0]

F-measure: [1.0, 1.0, 1.0]

Accuracy: 1.0

Performance with 10 fold (For 2 of the folds)

Mean Squared Error: 0.0148148148148

Confusion Matrix: [[35 0 0]
[0 49 1]
[0 1 49]]

Precision: [1.0, 0.98, 0.98]

Recall: [1.0, 0.98, 0.98]

F-measure: [1.0, 0.98, 0.98]

Accuracy: 0.985185185185

Confusion Matrix: [[50 0 0]
[0 49 0]
[0 1 35]]

Precision: [1.0, 1.0, 0.9722222222222222]

Recall: [1.0, 0.98, 1.0]

F-measure: [1.0, 0.98989898989899, 0.9859154929577464]

Accuracy: 0.992592592593

The confusion matrix helps in understanding the performance of the classification. Each row of the matrix represents the predicted labels and the columns represents the actual labels. Hence, we can get know easily how many inputs are misclassified.

Similarly, precision, recall, F-measure and accuracy are the parameters that tells us about the performance of algorithm. By looking at the accuracies above we can say that the data was classified better for training size 75%. Also with 10 folds, performance was almost accurate.

Spambase dataset (Discrete dataset):

This database contains 4601 samples and 2 classes- email is spam or not.

This dataset is considered for the last 2 questions of this assignment.

4] <u>Naïve Bayes with Bernoulli features</u>

a) Loading the dataset: (n features, 2 classes)

The dataset is derived from the text document. All the features are taken in the x as input and the labels are taken in the y. All the samples are considered as input.

b) Estimating the model parameters and computing the discriminant function:

Alpha is the parameter we need to estimate for calculating the membership function for each class.

Membership function:

$$g(x) = \sum_{j=1}^{n} x * \log\big(\alpha(j|y=i)\big) + (1-x) * \log\big(1-\alpha(j|y=i)\big)$$

This function tells about the membership of input x to class y.

So, by using discriminant function:

d(x)= g1(x)-g2(x)

We can tell which class the input x belongs to. Maximum value from the two membership function values will be the respective class value. Hence, if d(x) is positive, class will be 1 as g1 is maximum otherwise 2.

In this way classification is done using discriminant function.

c) Calculating Confusion matrix, precision, recall, F-measure and accuracy:

Performance with 75% training size
Confusion matrix: [[570 744]
[7 60]]
Accuracy: 0.456191165822
Precision: [0.4337899543378995, 0.8955223880597015]
Recall: [0.9878682842287695, 0.07462686567164178]
F-measure: [0.6028556319407721, 0.1377726750861079]

The confusion matrix helps in understanding the performance of the classification. Each row of the matrix represents the predicted labels and the columns represents the actual labels. Hence, we can get know easily how many inputs are misclassified.

Similarly, precision, recall, F-measure and accuracy are the parameters that tells us about the performance of algorithm. We can see that the accuracy is very less for this classification.

5] <u>Naïve Bayes with Binomial features</u>

b) Loading the dataset: (n features, 2 classes)

The dataset is derived from the text document. All the features are taken in the x as input and the labels are taken in the y. All the samples are considered as input.

c) Estimating the model parameters and computing the discriminant function:

Alpha is the parameter we need to estimate for calculating the membership function for each class.

Membership function:

$$g(x) = \sum_{j=1}^{n} \log \text{choose}(\text{P x})\big(\alpha(j|y=i)\big) \wedge x * \log\big(1-\alpha(j|y=i)\big)\wedge(P-x)$$

This function tells about the membership of input x to class y.

So, by using discriminant function:

d(x)= g1(x)-g2(x)

We can tell which class the input x belongs to. Maximum value from the two membership function values will be the respective class value. Hence, if d(x) is positive, class will be 1 as g1 is maximum otherwise 2.

In this way classification is done using discriminant function.

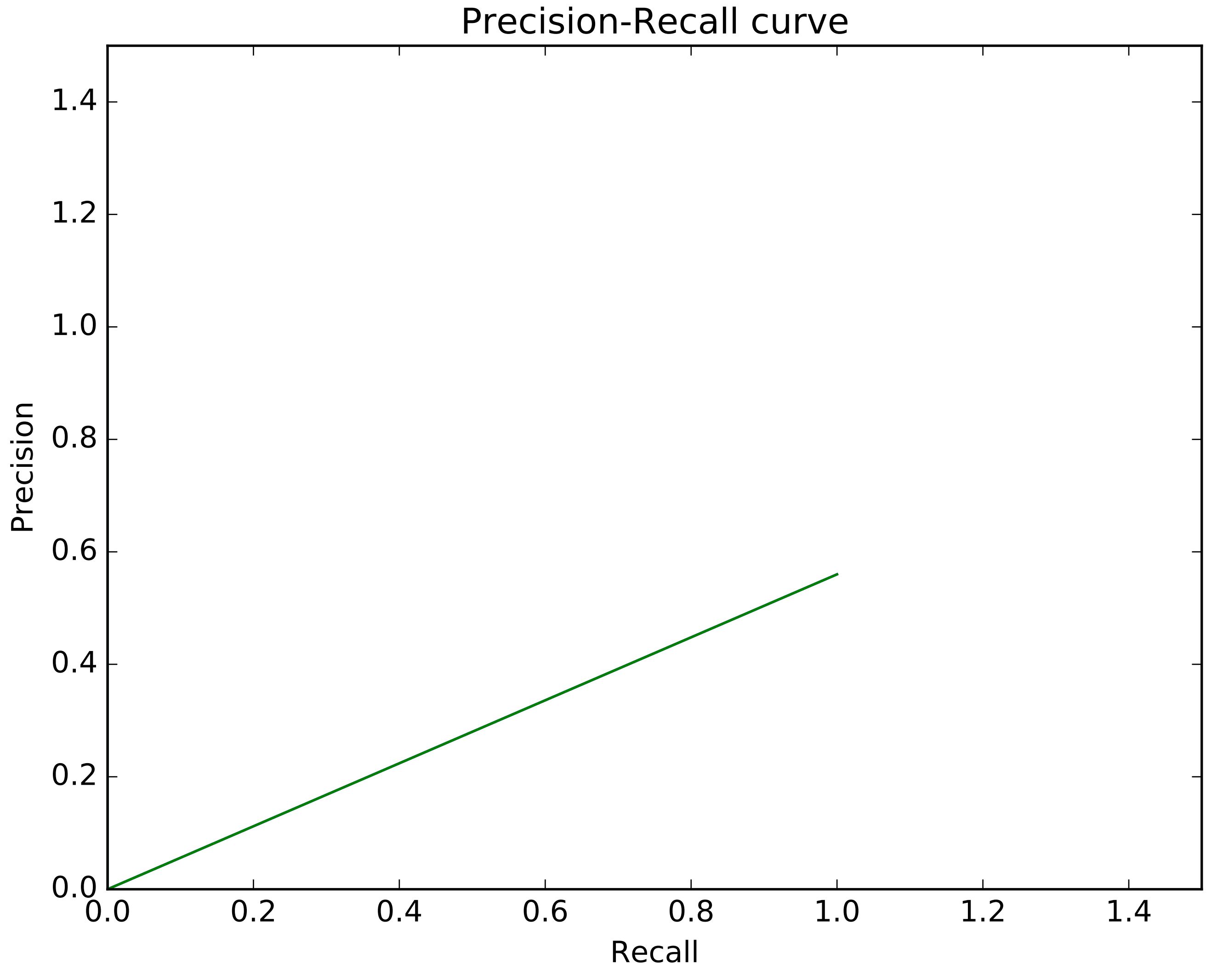
d) Calculating Confusion matrix, precision, recall, F-measure and accuracy:

Performance with 75% training size
Confusion matrix: [[396 556]
[181 248]]
Accuracy: 0.466328747285
Precision: [0.0, 1.0]
Recall: [1.0, 0.0]
F-measure: [0.0, 0.0]

The confusion matrix helps in understanding the performance of the classification. Each row of the matrix represents the predicted labels and the columns represents the actual labels. Hence, we can get know easily how many inputs are misclassified.

Similarly, precision, recall, F-measure and accuracy are the parameters that tells us about the performance of algorithm. We can see that the accuracy is very less for this classification.

Precision Recall curve for 2 Question. 2d)


Precision-Recall curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Q.5) Naive Bayes with Binomial features:

a) Maximum likelihood.

$$l(\theta) = \log \prod_{i=1}^{m} P(x^{(i)} | y^{(i)}; \theta) P(y^{(i)})$$

$$= \log \prod_{i=1}^{m} \left[\prod_{j=1}^{n} P(x_j^{(i)} | y^{(i)}; \theta) \right] P(y^{(i)})$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \log P(x_j^{(i)} | y^{(i)}; \theta) + \sum_{i=1}^{m} P(y^{(i)})$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \log \binom{p^{(i)}}{x_j^{(i)}} \alpha_{j|y=y^{(i)}}^{x_j^{(i)}} \left(1 - \alpha_{j|y=y^{(i)}}\right)^{p^{(i)} - x_j^{(i)}}$$

$$+ \sum_{i=1}^{m} \log P(y^{(i)})$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ l(\theta)$$

$$\frac{\partial l}{\partial \theta} = 0 \quad \text{i.e.} \quad \frac{\partial l}{\partial \alpha_{j|y=l}} = 0$$

$$\alpha_{j|y=l} = \frac{\sum_{i=1}^{m} 1(y^{(i)} = l)\, x_j^{(i)}}{\sum_{i=1}^{m} 1(y^{(i)} = l)\, p^{(i)}}$$