Generative Learning

Karthik Bangalore Mani
Department Of Computer Science
Illinois Institute of Technology

March 21, 2016

Abstract

The goal of this assignment is to fit the perform Gaussian Discriminant Analysis on multi-variate , multi-class datasets and implement Naïve Bayes with Bernoulli and Binomial features.

Problem Statement

- **Gaussian Discriminant Analysis** : Given a n Dimensional, k-class Datset, estimate the model parameters and compute discriminant function based on the distribution in each class. The examples must be classified and error should be measured. The confusion matrix should be constructed and the performance measures such as precision,recall,F-Measure and accuracy should be determined from the confusion matrix. The precision-recall curve should be plotted.

- **Naïve Bayes :** Given a 2-class datset with nD features, Implement Naïve Bayes with *Binomial* and *Bernoulli* features. The examples must be classified and error should be measured. The confusion matrix should be constructed and the performance measures such as precision,recall,F-Measure and accuracy should be determined from the confusion matrix.

Proposed Solution

- **Gaussian Discriminant Analysis :** Compute the mean vector for every different class in the data matrix. Also, compute the Co-variance matrix all the different class labels in the training dataset. Compute the membership value for class j of a given feature vector X. The class with the highest membership value will the predicted class. Perform the above mentioned steps for 1D – 2 Class, nD – 2 Class and nD – k Class Datasets.

- **Naïve Bayes with Bernoulli Features :** Compute the prior probabilities for different classes in the dataset, and the parameter alpha for all the features and all the different classes. Compute the membership value for class j of a given feature vector X. The class with the highest membership value will the predicted class.

Implementation Details

1. **Program Design Issues**
   Even the program gives accurate results, it takes lot of time to classify all the examples in the nD – 2 Class Dataset, as the number of training examples is large, around 248,050 examples.

2. **Problems faced**
   Problems were faced in getting good accuracy in the Naïve Bayes Bernoulli features. This was because, the influence prior probability was not included while computing the membership function. Once, the code was fixed, it gave good accuracy.

3. **Instructions to use the program**
   Open the Gen_Learning.ipynb file in iPython notebook, and execute each cell to get the desired output. The datasets must be placed in the same folder as that of the Gen_Learning.ipynb file.

Results and discussion

1. **GDA on 1 Dimensional , 2 Class Dataset :**

   The Skin_NonSkin.txt dataset was used from the UCI website. The original dataset contained 3 features, out of which 2 were removed for this scenario.

   Confusion matrix :
   ```
   [[      0.       0.]
   [  50859.  194198.]]
   ```

   Accuracy:
   ```
   0.792460529591
   ```

Class	Precision	Recall	F-Measure
1	0.0	0.0	0.0
2	0.792460529591	1.0	0.88421531912

   As seen above, the accuracy of predictions is 79.24%, which is quite low. This is because, the original dataset contained 3 features, out which 2 were discarded and only 1 was chosen.

2. **GDA on n Dimensional, 2 Class Dataset :**

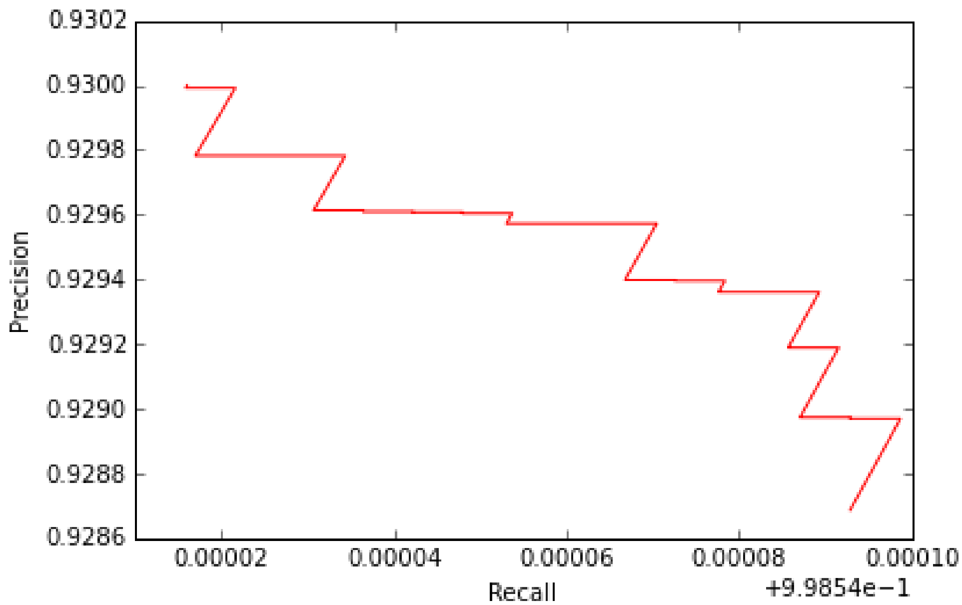
   The Skin_NonSkin.txt dataset was used from the UCI website. In this scenario, no features were removed and the original dataset itself was used. Hence accuracy was found to increase from 79.24% to 94.5%.

   Confusion Matrix :
   ```
   [[  37846.     295.]
    [  13013.  193903.]]
   ```

   Accuracy :
   ```
   0.945694267048
   ```

Class	Precision	Recall	F-Measure
1	0.992265541019	0.744135747852	0.850471910112
2	0.937109745017	0.998480931832	0.966822399617

Precision – Recall Curve:



3. **GDA on n Dimensional, k Class Dataset :**
   The iris.data dataset was used from the UCI website.

   Confusion Matrix :
   ```
   [[ 50.   0.   0.]
    [  0.  46.   1.]
    [  0.   4.  49.]]
   ```

   Accuracy :
   0.966666666667

Class	Precision	Recall	F-Measure
1	1.0	1.0	1.0
2	0.937109745017	0.92	0.948453608247
3	0.924528301887	0.98	0.95145631068

4. **Naïve Bayes with Bernoulli Features :**
   The spambase.data from UCI website was used. Since the features were not binary, it was binarized.

Confusion Matrix :

```
[[ 1478.    190.]
 [  335.   2598.]]
```

Accuracy :

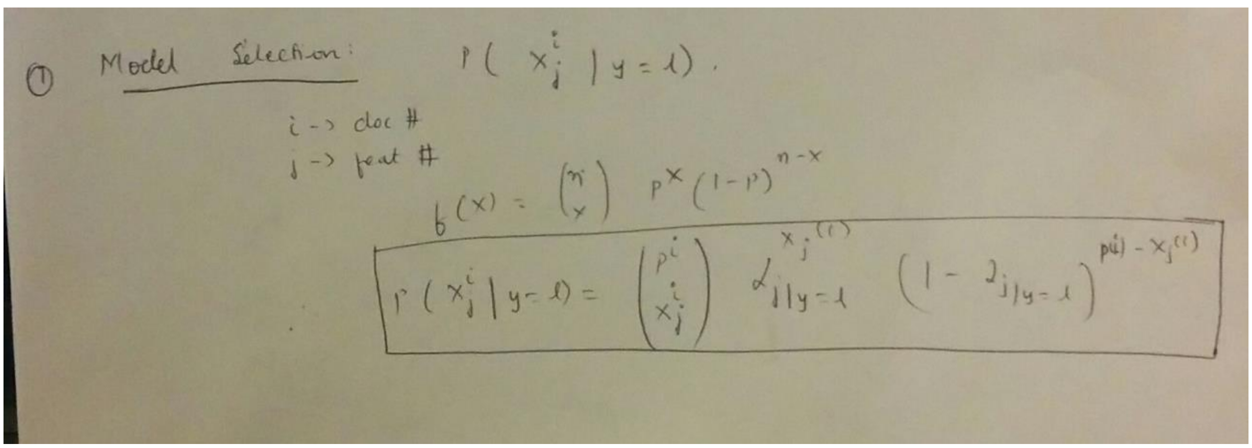
```
0.885894370789
```

Class	Precision	Recall	F-Measure
1	0.886091127098	0.815223386652	0.84918126975
2	0.885782475281	0.931850789096	0.908232826429

The accuracy was so less i.e 88.58 % because, the features in the SPAM dataset were binarized during Bernoulli NB.

5. Naïve Bayes with Binomial Features :

Parameter estimates derivation :



Compute parameters (M-L):

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

I I D.

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

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

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

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

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

$$\frac{dl}{d\theta} = 0$$

$$\theta = \left[\alpha_{1|y=1} \cdots \alpha_{n|y=1}, \cdots, \alpha_{1|y=k}, \cdots \alpha_{n|y=k}, \alpha_1, \cdots, \alpha_k \right].$$

$$\frac{\partial l}{\partial \alpha_i|_{y=j}} = 0 \quad ; \quad \frac{\partial l}{\partial \alpha_j} = 0$$

$$\boxed{P(y=l) \equiv \alpha_l = \frac{\sum_{i=1}^{m} 1(y^i = l)}{m}} \quad \boxed{l = 1, \ldots, k}$$

$$\boxed{\alpha_{j|y=l} = \frac{\sum_{i=1}^{m} 1(y^i = l) \cdot x_j^{(i)} + \epsilon}{\sum_{i=1}^{m} 1(y^i = l) \cdot p^{(i)} + 2\epsilon}}$$

(3) <u>Membership</u>

$$g_l(x) = \log P(y=l|x) \propto \log\left(P(x|y=l) \cdot P(y=l)\right)$$

l ↓ class label

$$= \log\left(P(x|y=l)\right) + \log\left(P(y=l)\right)$$

$$= \log \prod_{j=1}^{n} P(x_j|y=l) + \log\left(P(y=l)\right)$$

$$= \sum_{j=1}^{n} \log P(x_j|y=l) + \log\left(P(y=l)\right)$$

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

total no of words in doc.

The spambase.data from UCI website was used. Since the length of documents was not present in the dataset, the length of the documents were assigned to be 10.

Confusion Matrix :

```
[[ 1736.  2182.]
 [   77.   606.]]
```

Accuracy :

```
 0.509019778309
```

Class	Precision	Recall	F-Measure
1	0.443083205717	0.957528957529	0.605827953237
2	0.887262079063	0.217360114778	0.349178910977

Weakness : No suitable dataset with Bernoulli features was found, and hence the lengths of documents were randomly assigned. Because of which, it resulted in a low accuracy for Naïve Bayes with Bernoulli Features.

References

1. **Stackoverflow.com**
2. **Wikipedia.org**