RS 321 Pattern Recognition

Programming Assignment

Atma Anand
SC12B156

Date: March 09, 2015 Department: Physical Sciences

Aim: 1. To classify the given database of Iris plants using Naive (simple) Bayes classifier.

Procedure:

1. Identify the data subset to be used and the attributes to be considered.
   Here we have considered all 4 attributes. Ideally the data should have been separated into training and test sets but to limit randomness we have use the whole dataset as training samples and then considered a few samples for testing.
2. Separate the data according to its class and attribute and apply Maximum Likelihood Estimator (MLE) to find its mean and variance assuming a *Gaussian distribution*.
   A histogram of the individual (4) attributes of the 3 classes appear to resemble a Gaussian profile most closely (if anything at all). Hence, we apply MLE to find the parameters and thus the likelihood.
3. Find the posterior probability using prior probability and obtained likelihoods.
   Here the priors are equal hence this step may be skipped without any classification errors whatsoever. We have *assumed each attribute to be independent,* hence a product of the prior and the 4 attributes' likelihood gives the functional form of posterior probability required for classification.
4. Classify the input test samples.
   The attributes for each sample is input individually and the result of classification is displayed. Upon termination the no. of test samples in each class is displayed.
5. Error analysis and finding efficiency of the method.
   From our results we can find the errors of omission and commission and hence the efficiency of the algorithm. This is best represented by a confusion matrix.

Observation:

	Class 1: Iris Setosa		Class 2: Iris Versicolour		Class 3: Iris Virginica	
Attribute	**Mean**	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Sepal length	5.006	0.121764	5.936	0.261104	6.588	0.396256
Sepal Width	3.428	0.140816	2.77	0.0965	2.974	0.101924
Petal length	1.462	0.029556	4.26	0.2164	5.552	0.298496
Petal Width	0.246	0.010884	1.326	0.038324	2.026	0.073924

Program Execution: (in C)

```
Mean    Variance
Class 1:
5.006   0.121764
3.428   0.140816
1.462   0.029556
0.246   0.010884

Class 2:
5.936   0.261104
2.77     0.0965
4.26     0.2164
1.326   0.038324

Class 3:
6.588   0.396256
2.974   0.101924
5.552   0.298496
2.026   0.073924


Enter four feature of plant to be classified:
5.1  3.5  1.4  0.2
Classifying for: 5.1 3.5 1.4 0.2
P1=342.758      P2=4.65413E-016         P3=2.43797E-023
Plant belongs to Iris-setosa

Classify more samples? (1 or 0): 1

Enter four feature of plant to be classified:
7   3.2   4.7  1.4
Classifying for: 7 3.2 4.7 1.4
P1=0            P2=1.80756              P3=0.440544
Plant belongs to Iris-versicolor

Classify more samples? (1 or 0): 1

Enter four feature of plant to be classified:
7.9  3.8  6.4  2
Classifying for: 7.9 3.8 6.4 2
P1=0            P2=1.18957E-011         P3=0.0400893
Plant belongs to Iris-virginica

Classify more samples? (1 or 0): 0

Total Classified: 3
Class 1         Class 2         Class 3
1               1               1

--------------------------------
Process exited after 60.3 seconds with return value 0
```

Result:

We have successfully implemented an algorithm to classify the given database of Iris plants using Naive Bayes classifier. There were *no errors* of omission and commission encountered during the test run. However, not all the data was used for testing hence the exact accuracy of this classification cannot be stated. Under the test conditions we have used the confusion matrix takes the principal diagonal form.

(**Note:** Also check the C program attached in the mail which has been omitted here due to lack of clarity)