Assignment-05
Naive Bayes Classifier

Ravi Kumar
14MCMI12
SCIS, University of Hyderabad

March 23, 2015

1 Introduction

Bayesian reasoning provides a probabilistic approach to Inference.It is based on the assumption that the quantities of interest are governed by probability distribution and that optimal decision can made by reasoning about these probabilites together with observed data.Bayesian learning algorithm that calculate explicit probabilites for hypotheses i.e naive Bayes classifier.

Naive Bayes Classifier:

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem with *Independence* assumptions between predictors and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

The naive Bayes classifier applies value and where each instance x is describe by a *conjunction* of attribute values and where the $Target$ function $f(x)$ can take on any value from finite set V. A set of training examples of the target function is provided, and a new instance is presented, describe by the tuple of attribute values $< a_1, a_2, ...a_n >$.The learner is asked to predict the target value, of classification, for this new instance.

$$v_{MAP} = \arg\max_{j \epsilon V} P(v_j \mid a_1, a_2 ... a_n)$$

By using Bayes theorem

$$v_{MAP} = \arg\max_{j \epsilon V} P(a_1, a_2 ... a_n \mid v_j)$$

and

$$v_{NB} = arg \max_{v_j \epsilon V} P(v_j) \prod_i P(a_i | v_j)$$

The naive Bayes learning method involves a learning step in which the various $P(v_j)$ and $P(a_i \mid v_j)$ terms are estimated, based on their frequencies over the training data. The set of these estimates corresponds to the learned hypothesis. This hypothesis is then used to classify each new instance by applying the

rule the above equation. Whenever the naive Bayes assumption of conditional independence is satisfied, this naive Bayes classification v_{NB} is identical to the MAP classification.

2 Modules

2.1 Preprocessing

The data is read from a file which is in txt format . This module takes the dataset name via command line arguments, the other two arguments represents if there are any column names have given to dataset. If given then 1 else 0. Similarly if there is column specifying the row names then that also should be ignored. The data is divided into two parts first one is Training(75%)data and another Testing(25%) data.

2.2 Matric Calculation

In this module first we are finding how many *lables*(yes,no,..etc) are present on class label From the training dataset and all the statistical values that are used for the calculation of probabilities.On the basis of this lable find the number of different labels for each class values of attributes.At the last of this module we named column the list heading with attribute name and row with class label.

2.3 Probablity Calculation

This is the module that classifies the test instances. The function considers one instance at a time and classifies the instance according to the naive Bayes formula and assigns the proper class label. This is the whole functionality of this module. This is then passed directly to the error calculation module which calculates the accuracy of the classification.

2.4 Error Calculation

This modules actualy contains the classified data labels and test labels. The test data instances label matched with the tainned data label.This mapping is *onetoone* & at the same time count the matched labels and at last with the help of this we provide the accurecy of Algorithm.

2.5 Contribution

My contribution to this Assignment is the implementation of Error Calculation & average accuracy calculation.

2.6 Experimental Result

This code is working completely fine.According to problem definition the training data set & test data set are divided randomly.so the accuracy rate is also depend upon the type of data set that is involved and division. The given data set was around uniformly distributed across different classes.

2.7 Conclusion

We iterate the code 50 times and final accuracy of the code is average accuracy.This is to conclude that the Navie Bayes Classifier has been coded and tested on 1,2 data sets giving average results on performance also. This code can be improved further.

S.No	Accuracy	S.No	Accuracy
1	89.72222	26	90.30864
2	89.62963	27	91.38889
3	89.75309	28	90.46296
4	90.80247	29	90
5	90.70988	30	91.32716
6	90.27778	31	90.06173
7	88.30247	32	90.40123
8	89.62963	33	89.96914
9	90.09259	34	89.72222
10	90.03086	35	89.59877
11	90.61728	36	89.75309
12	89.41358	37	90.18519
13	90.61728	38	90.98765
14	89.87654	39	90.74074
15	90	40	90.64815
16	89.72222	41	90.24691
17	90.74074	42	89.96914
18	89.90741	43	90.52469
19	90.15432	44	90.33951
20	89.84568	45	90.40123
21	90.37037	46	89.25926
22	90.61728	47	89.66049
23	90.33951	48	90.24691
24	90.49383	49	90.46296
25	90.74074	50	90.83333

mean Accuracy : 90.19815