COMP24111

Exercise 2 Spam Filtering

REPORT

Learning parameters

In the training phase we need to compute the probabilities for each value of each feature in each class. Therefore, I counted each of these values from the give data set and then divided each of these values by the number of values for each class. This gives me the lookup table needed for the testing.

Storing the parameters

For the first part I stored the probabilities in a 2D matrix in the following way :

LookUpTable =

Class = n	Class = n	…	Class = 1	…	Class = 0
Feature = m	Feature = m-1	…	Feature = m	…	Feature = 0
0.3384	0.6616	…	0.1481	…	0.8519
0.6037	0.3963	…	0.2767	…	0.7233
0.0229	0.9771	…	0.0022	…	0.9978

Also I calculated P(class = n) for each class;

For the second part I copied all the values for each class in a new table and then I calculated the mean and standard deviation for each feature and I got two vectors for each class, one that has the mean and one that has the standard deviation.

Test Results

For the tests I have created a matrix that holds the probabilities for the new data set and outputs a new vector with the expected labels for each instance.

My results are

For av2_c2 : 89.048240

For av3_c2 : 89.308996

For av7_c3 : 86.347826

For avc_c2 : 77.873694

Cross validation

Working with a 10-fold cross validation we need to split the 4601 rows table 10 times so that each time we would get a 460 training data set and the rest of the 4141 rows a testing set.

Implications

Confusion matrix =

av2_c2		av3_c2		av7_c3			avc_c2	
1296	108	1296	108	1195	0	90	969	419
144	753	138	759	0	635	134	68	745
----------	----------	----------	----------	54	36	156	----------	----------

As we can see, the more complex the data gets the less accurate the model is. Therefore, I predict that the more training we do for the model the better will be the results and the accuracy that we will get as a general rule. So, for more complex models we will need a better training phase with more and more data set instances.