Anca-Florina Radulian
Student ID: 9911494

Date:27/11/2017

Exercise 3 – Naïve Bayes Classifier for Spam Filtering
COMP24111

In the learning phase we are utilising attributeSet and LabelSet in order to extract the conditional probabilities and initial class probabilities which are used in the next phase.

In the testing phase we are calculating the probability that a certain feature vector belongs to class. This is done using this formula:

$$p(C_i|\textbf{x}) = [p(x_1|C_i)\, p(x_2|C_i)\, p(x_3|C_i)...p(x_n|C_i)] * p(C_i)$$

After that, the accuracy is measured based on the number of correct classifications made by the Bayes classifier.

The "parameters" used in discrete naïve Bayes are the following:
- noOfClasses: contains the number of classes used. It is determined using the maximum element found in the labelSet array. It is used as a parameter for the 3-dimensional matrix.
- priorProbabilities: an array containing the initial probabilities of the classes. It is constructed by counting the appearances of each class in the label set.
- probabilityMatrix: the 3-dimensional matrix in which the conditional probabilities are stored for each attribute of the feature vector. The matrix is created similarly to a frequency array. We are increasing each value that we find in the AttributeSet matrix in the new 3D matrix.

The accuracies obtained in the first part of the laboratory were around 88%.

```
Enter a filename to load data for training/testing: av3_c2
*************************************************
Overall Accuracy on Dataset av3_c2: 0.893525
Confusion matrix for part 1:

ans =

      1298        106
       139        758



*************************************************
```

```
Enter a filename to load data for training/testing: av2_c2
*************************************************
Overall Accuracy on Dataset av2_c2: 0.890917
Confusion matrix for part 1:

ans =

      1298        106
       145        752



*************************************************
```

```
Enter a filename to load data for training/testing: av7_c3
*********************************************
Overall Accuracy on Dataset av7_c3: 0.862609
Confusion matrix for part 1:

ans =

      1195          0         90
         3        631        135
        54         34        158



*********************************************
```

As we can see from the confusion matrices, most of the "normal" mails are classified right, whereas for the "spam" mails, there are more misclassified feature vectors as "normal".

To conclude, the first part of this exercise is quite precise, the accuracy being high for a discrete-valued naïve Bayes Classifier.