# T-61.3050 : Email Classification as Spam or Ham using Naive Bayes Classifier

Santosh Tirunagari : 245577

January 20, 2011

**Abstract**

This term project gives a solution how to classify an email as spam or ham using the Naive bayes classifier. With the growth of electronic mails, classification can be seen as the most challenging yet, most important task to avoid spam. In this regard, it is important to identify and implement the most effective approach for email classification. I have used Naive Bayes Classifier algorithm for this task. Results obtained show that Naive Bayes Classifier algorithm has good accuracy and is thus comparable to any other methods for email classification.

# Contents

# List of Figures

# List of Tables

## 0.1 Introduction

Since the Internet had been in wide spread, the emails have been growing enormously. The content of an email consist of email header consisting of control information and email body consisting of message. An Email is said to be UBE Unsolicited Bulk Email or spam mail or junk mail or UCE Unsolicited Commercial Email, if its message consists of unwanted content or commercial content in large quantities to an indiscriminate set of recipients [1][2][3]. When internet was publicly opened, spamming becomes prevalent, utilising the popularity of email communication, spammers flood the mail boxes with the unwanted advertisement emails with the commercial purpose. Spam emails are 80% of total emails in the world [4].so there is need of development of a system which can automatically filter the spam mails. In recent years many techniques have been developed for the classification of emails. In general the spam mails are of some specified format, patterns of words and vocabulary found in the header or body of an email [1][3].

To counter the spam problem there are two approaches of spam filtering: knowledge engineering and machine learning.The former method requires to propose an explicit set of rules. But machine learning methods like bayesian classification, K-NN, ANNs, SVMs do not require any rules. Instead they need an algorithm and training data [4].

I have used Naive Bayes Classifier method to classify the messages in the given dataset as either spam or ham. Naive Bayes Classifier is a simple method based on probability theory. Here in this report I discuss the Naive Bayes Classifier algorithm, my experimental methodology, results and conclusions.

## 0.2 Method

### 0.2.1 Naive Bayes Classifier

A classifier is a function f that maps input feature vectors $x \in X$ to output class labels $y \in 1, ..., C$. where $X$ is the feature space [5,6].

#### Why Naive Bayes ?

There are many different types of classifiers. I have chosen Naive Bayesian classifier for my work because it is very easy to understand conceptually. But after a detailed contemplation I understood that Naive Bayesian classifier is very popular among many email clients and is very effective in spam filtering [5, 7, 9]. It is important to know the merits and demerits of a method before opting. I learnt that Naive Bayesian classifier is fast and space efficient, robust to isolated noise points, able to handle quantitative and discrete data, not sensitive to irrelevant features. But it assumes independence of features and fails if conditional probability is zero. Although it is seldom, false positives do happen [5, 7, 9].

### 0.2.2 Error Estimation

The error made by model can be estimated using certain measures of error estimation like precision, recall and accuracy. Precision and recall can be seen

as extended versions of accuracy. In the context of classification, true positives, true negatives, false positives, false negatives are used to classify with desired correctness of classification [15,16].

| | | Correct Classified | |
|---|---|---|---|
| | | E1 | E2 |
| Observed result | E1 | (tp) True positives | (fp) False positives |
| Observed result | E2 | (fn) False negatives | (tn) True negatives |

Table 1: Allocation of true positives, true negatives, false positives, false negatives.

### 0.2.3 Precision

precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$Precision = \frac{tp}{tp + fp}. \tag{1}$$

### 0.2.4 Recall

Recall is given by

$$Recall = \frac{tp}{tp + fn}. \tag{2}$$

### 0.2.5 Accuracy

The accuracy is the proportion of true results (both true positives and true negatives). It is a parameter of the test

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

### 0.2.6 F-measure

F-measure is the harmnic mean of Precision and Recall

$$F - measure = \frac{2.P \times R}{P + R} \tag{4}$$

### 0.2.7 Dimensionality Reduction

I denote by $\mathbf{X}$ the data matrix, which contains the data vectors of the set $X$ as its columns.

#### Principle Component Analysis

PCA finds a linear projection subspace so that the preservation of variance of the data is maximal. The subspace is spanned by eigenvectors of the covariance matrix of the data, which can be computed through the eigendecomposition

$$cov(\mathbf{X}) = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^{\mathsf{T}}. \tag{5}$$

The dimension reduction is performed by projecting the data linearly on the subspace spanned by the column vectors of $V$ corresponding to the largest eigenvalues contained on the diagonal of the matrix $\mathbf{\Lambda}$. By choosing the $D$ most relevant eigenvectors as the columns of the matrix $\mathbf{W}_D$, the principal component projection is given by

$$\mathbf{X}_{\text{PCA}} = \mathbf{W}_D^{\mathsf{T}}\mathbf{X}. \tag{6}$$

## 0.3 Experiment

The dataset provided in the course webpage has been used for this term project. The first 1000 samples in the dataset were already labeled. I considered these 1000 samples initially as trained data. Remaining 9000 samples in the dataset which are to be classified as either spam or ham were treated as test data. I have labels in the second column for the trained data. So the information from this column has been used to classify the class of the samples. For classification using naive Bayesian algorithm, I use trained data, test data and class information from trained data. All the 448 features were used for classification.

I have implemented the Naive Bayesian classification algorithm on Matlab. Matlab is a technical computing language which can be used in algorithm development, numeric computation, data analysis and visualization. Matlab has a wide range of applications including signal and image processing, computational biology, communications. For solving particular problems there are special toolboxes in Matlab. For example it has ready toolboxes for many statistical problems and neural networks in the context of my course. [17]. So I chose Matlab for my work.

I have applied the dimensionality reduction method PCA as shown in the Table 2 and found that the accuracy and F-measure values were not good enough as shown in Figure 1. Hence I have used all the dimensions of the data.

| Dimensions | Accuracy | F-measure | Precision | recall |
|---|---|---|---|---|
| 50 | 0.51 | 0.092592593 | 0.625 | 0.05 |
| 75 | 0.515 | 0.110091743 | 0.666667 | 0.06 |
| 100 | 0.51 | 0.092592593 | 0.625 | 0.05 |
| 125 | 0.51 | 0.092592593 | 0.625 | 0.05 |
| 150 | 0.51 | 0.092592593 | 0.625 | 0.05 |
| 175 | 0.51 | 0.092592593 | 0.625 | 0.05 |
| 200 | 0.515 | 0.110091743 | 0.666667 | 0.06 |
| 225 | 0.515 | 0.110091743 | 0.666667 | 0.06 |
| 250 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 275 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 300 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 325 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 350 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 375 | 0.52 | 0.127272727 | 0.7 | 0.07 |
| 400 | 0.52 | 0.127273 | 0.7 | 0.7 |

Table 2: Impact of dimensionality reduction using PCA

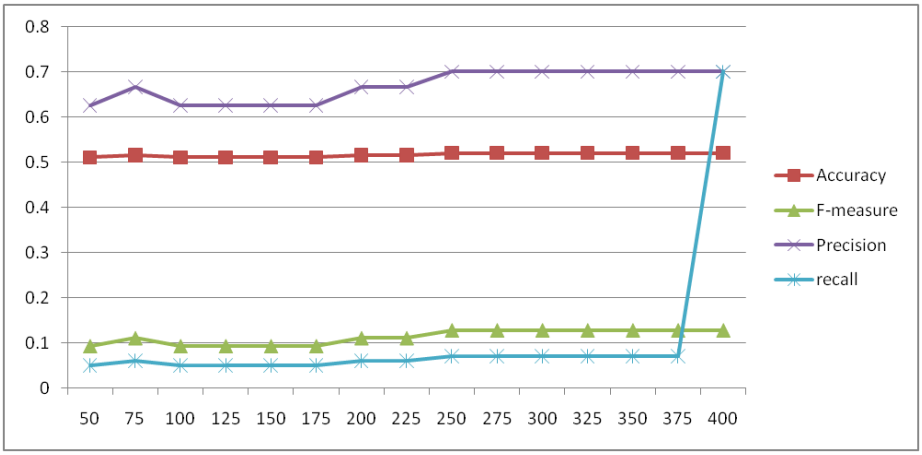


Figure 1: Impact of PCA

### 0.3.1 Validation

Cross validation is a method to evaluate the behavior of the model based on my training data on future data or unknown data. Assume that selection of a particular model depends upon n data points. Then the basic idea of cross validation is to split the n data points into two parts. The first part nc is used for fitting the model and the second part n v which is n-nc is used to validate the predictive ability of the model [18, 19]. I have done this cross validation 5 times with different subsets of data. First I have considered the samples as sets from the training data. The Table 3 shows the training and validation sets.

| Sets | Training sets | Validation sets |
|---|---|---|
| Set 1 | 1-400 & 501 - 900 | 401-500 & 901 - 1000 |
| set 2 | 1-400 & 601 - 1000 | 401-500 & 501 - 600 |
| set 3 | 101-500 & 501 - 900 | 1-100 & 901 - 1000 |
| set 4 | 101-500 & 601 - 1000 | 1-100 & 501 - 600 |
| set 5 | 1-800 | 801 - 1000 |

Table 3: Collection of Training and Validation sets

After performing Naive Bayes Classifier algorithm on the these training sets and validating it, I found that set 1 has good results as accuracy 96% and F-measure 96% as shown in Table 4 and Figure 2.

| Training Sets | Accuracy | F-measure | Precision | recall |
|---|---|---|---|---|
| Set 1 | 0.965 | 0.9641 | 0.9895 | 0.94 |
| set 2 | 0.925 | 0.9198 | 0.9885 | 0.86 |
| set 3 | 0.96 | 0.9592 | 0.9792 | 0.94 |
| set 4 | 0.92 | 0.9149 | 0.9773 | 0.86 |
| set 5 | 0.875 | 0.9333 | 0.9773 | 0.875 |

Table 4: Impact of Validation

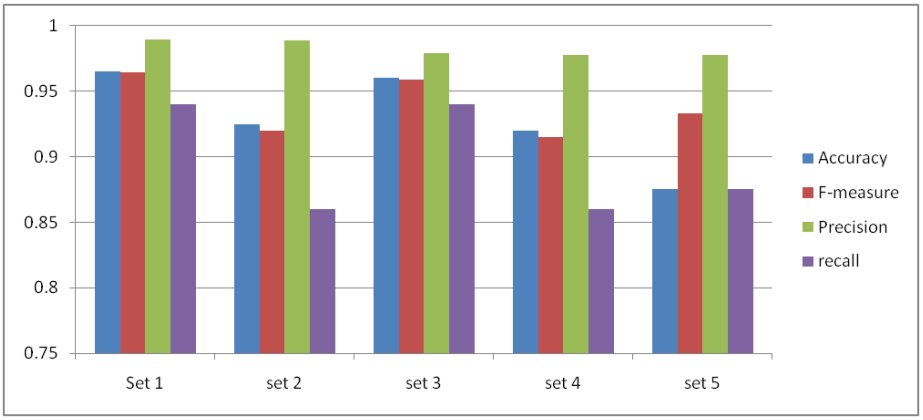


Figure 2: Impact of Validation

## 0.4 Results

Considering the accuracy apparently all the 448 features were essential. So all the 448 features from the test data were subjected to classification without any application of dimensionality reduction method and the model was found to be 93.5% accurate with a precision of 98.7% and a recall of 91.6% and has a F-measure of 95.05%. These results indicate that this model can be used to classify emails with high accuracy.The Final result for the test-data was as shown in Table 5.

| Accuracy | F-measure | Precision | recall |
|---|---|---|---|
| 0.935 | 0.9505 | 0.9872 | 0.9165 |

Table 5: Result on Test Data

## 0.5 Conclusion

The classification algorithm that I used is equally good and comparable to other methods of classification. I found that the model is 93.5% accurate and has F-measure of 95.05%and a precision of 98.72%. Which substantiates my conclusion.

## 0.6 Acknowledgements

I thank course instructor and coordinator for giving me this oppurtunity for conducting the experiments and writing a good report.

## 0.7 References

1. McAfee/ICF, Carbon footprint of Email spam report.
2. Scott Hazen Mueller, What is spam?

3. Gordon cormack, Email spam filtering – A systematic review.
4. Konstantin tretyakov, Machine learning techniques in spam filtering.
5. Kevin Murphy, Naive Bayes Classifiers.
6. Naive Bayes classification tutorial – Zhang Hongxin
7. Tutorialcode project - http://www.codeproject.com/KB/recipes/Naive_Bayes.aspx
8. Lecture slides – Machine learning and data mining – Radford neal – University of Toronto
9. I.Rish, An empirical study of naive Bayes classifier.
10. Mehran Sahami, A bayesian approach to filter Junk email.
11. Naive Bayesian classification of structured data – Peter flach
12. A comparative study for email classification – Seongwook youn
13. Naive-Bayes vs Rule-Learning in classification of email – Jefferson provost
14. Ethem Alpaydin, Introduction to machine learning.
15. John makhoul, Performance measures for information criterion.
16. Wikipedia, Precision and Recall.
17. Matlab, The language of technical computing- url www.mathworks.com/products/matlab
18. Jun shao, Linear model selection by cross validation.
19. Ron kohavi, A study of cross validation and bootstrap for accuracy estimation and model selection.