CS5691: Pattern Recognition and Machine Learning

Assignment-3

Sarthak Naithani

CS22M078

OBJECTIVE:

To build a spam classifier from scratch

For classification of spam and ham mails I used Naive Bayes Classifier. The dataset has been taken from the following given source http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html.

APPROACH

Data modeling algorithm used-
Naive Bayes is a classification algorithm which is based on generative modeling and works on the principle of Bayes Theorem. Here we make an assumption such that our features are independent of each other.

Cleaning

After extracting the dataset I *preprocessed* my data by

1. Replaced '\n' to " " and changed the words to lowercase alphabets..
2. Considering only the alphanumeric words.

Training

After that I maintained a count of each word/feature in a
dictionary for each spam and ham mail named as spam_words and ham_words.
Then we will do *feature extraction* by selecting the top 2000 most occurred set of words in main dictionary named main_dict

Modeling our data by applying Naive Bayes

Now we have to check the probability of each word given label-y indicating test mail is spam or not spam.

Which can be termed as P(Spam | Mail) and P(Ham | Mail)

Where, P(Mail | Spam) = P(word1 | Spam) * P(word2 | Spam) ... P(word-n | Spam), where word-1, word-2 ... word-n is the set of words in the given test mail

P(Spam) - Prior we calculated Spam mail

P(Mail) - Evidence

The same is calculated for ham mail also.

Testing

Now we have P(Spam | Mail) and P(Ham | Mail)

If P(Spam | Mail) > P(Ham | Mail): Mail will be predicted as Spam

else the mail will be predicted as Ham

It is observed that after training with the below-mentioned, we get an accuracy of 90% approx depending on the dataset given.

•