CS F464 | Machine Learning
Aditya Agarwal | Manisha Katariya | Jalaj Bansal

Assignment - 1B: Naïve – Bayes Classifier

1. Description:

Our model takes an input file with email contents and labels signifying whether they are spam or not and classifies testing samples using a 7-fold cross validation scheme.

First, we clean the data to create a dataset in the desired format. We start by removing punctuation marks, numbers and converting the text to lower case since these features do not determine whether an email is spam or not. We then remove stopwords such as an, out, be, which do not add significant meaning to the email contents. We then create a vocabulary with all the words of our dataset.

Then, we perform 7-fold cross validation to train and test the model. After splitting the dataset into 7 parts, we call a fit function, which calculates all the prior probabilities, and for each word in the vocabulary, calculates the probability of that word being present in a spam message or a non-spam (ham) message.

P(spam) = No. of spam messages / Total number of messages
P(ham) = No. of ham messages / Total number of messages

For each word w_i in the vocabulary,

$$\mathbf{P(w_i|Spam)} = \frac{\mathbf{N_{w_i|Spam}} + \boldsymbol{\alpha}}{\mathbf{N_{Spam}} + \boldsymbol{\alpha} \cdot \mathbf{N_{Vocabulary}}}$$

$$\mathbf{P(w_i|Ham)} = \frac{\mathbf{N_{w_i|Ham}} + \boldsymbol{\alpha}}{\mathbf{N_{Ham}} + \boldsymbol{\alpha} \cdot \mathbf{N_{Vocabulary}}}$$

α - Laplace Smoothing Parameter
In our model we have chosen, $\alpha = 1$

In the prediction phase, the condition used for classification is,

if $P(Ham|w_1, w_2, ..., w_n) > P(Spam|w_1, w_2, ..., w_n)$, the message is classified as ham, and vice-versa.

The accuracy for each fold is calculated using the formula,

$$accuracy = \frac{number\ of\ correctly\ classified\ testing\ samples}{total\ number\ of\ training\ samples}$$

2. Accuracy Results:

Fold	Accuracy
1	0.784
2	0.84
3	0.742
4	0.721
5	0.77
6	0.763
7	0.784

Final Accuracy = 0.7719

3. Major Limitations of the Naïve – Bayes Classifier:

- It assumes that all features (in our case words of the email) are independent of each other, but this is rarely true.
- It has a ‘zero-frequency’ problem where we may get a division by zero error and for a word not present in the vocabulary, we get a zero probability due to multiplication by 0, but we have taken care of that with Laplace Smoothing