Machine Learning Assignment-1

Documentation

Submitted by - Maurya Grover, Milind Anand, Akul Singhal

2. Naïve Bayes

The Naïve Bayes algorithm is a classification technique that is based on the Bayes' Theorem, with the assumption of pairwise independence between the features.

Implementation

In the preprocessing stage, a list of stopwords (common words that do not convey any specific meaning to the sentences of the dataset) was created. The stopwords were then removed from each of the lines of the dataset. Furthermore, punctuation symbols were also removed from the dataset.

A set was created containing all of the words that had an instance in the preprocessed data, termed as the vocabulary. Using this, an equivalent vector was constructed for each sentence.

For each class k, the conditional probability of it containing a word from the vocabulary was calculated, after the addition of a constant α (for Laplace Smoothing, which helps to avoid the problem of zero probability), according to the following expression:

$$P(word=1 \mid class = k) = (n_k + \alpha)/(f_k + \alpha * s),$$

Where, n_k = number of times a word is present when the sentiment is k

f_k = number of times sentiment is k

α = Laplace Smoothing constant

s = size of vocabulary

The posterior probabilities for each sentence were calculated by the application of the Bayes' Theorem:

$$P(k=1 \mid word) = P(word \mid k=1) / (P(word \mid k=1) + P(word \mid k=0))$$
$$And\ P(k=0 \mid word) = P(word \mid k=0) / (P(word \mid k=1) + P(word \mid k=0))$$

If the probability of the sentence being spam is greater than it being not spam, then it is classified as being spam, else it is not spam. 7-fold cross validation was performed by splitting the dataset into 7 equal partitions and using 6 parts to classify the sentences of the 7th part. The predictions were compared with the known values to obtain the accuracy for each fold.

Results

7-fold Cross Validation Results:

Test fold 1 : Accuracy = 77.46
Test fold 2 : Accuracy = 82.52
Test fold 3 : Accuracy = 81.12
Test fold 4 : Accuracy = 77.62
Test fold 5 : Accuracy = 86.71
Test fold 6 : Accuracy = 76.92
Test fold 7 : Accuracy = 84.62
Mean of accuracies = 80.99646832041198
Std Dev. of accuracies = 3.5568204700850803

Major Limitation of Naïve Bayes

The Naïve Bayes classifier works on the assumption that the features are pairwise independent. This means that it will fail when it is fed a dataset that violates this assumption, which is often the case with real-world data.