# **Machine Learning**

# **Programming Assignment 2**

# **Report**

**Name:** Apurva Mithal

**NetID:** axm174531

1. **Algorithm:** Naïve Bayes

**Accuracy of the Naïve Bayes Classifier on the test data set:**

|  |  |
| --- | --- |
| With Stop Words | Without Stop words |
| 94.9791% | 94.3515% |

1. **Algorithm**: Logistic Regression

**Accuracy of the Logistic Regression Classifier on the test data set:**

**Learning Rate = 0.001**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lambda**  **(Regularization**  **Constant)** | **Number of Iterations (With Stop Words)** | **Accuracy (With Stop Words)** | **Number of Iterations (Without Stop Words)** | **Accuracy (Without Stop Words)** |
| **0.01** | 2 | 86.4017% | 2 | 85.5649% |
|  | 5 | 89.1213% | 5 | 87.4477% |
|  | 6 | 89.1213% | 6 | 88.4937% |
|  | 7 | 89.5397% | 7 | 88.2845% |
|  | **14** | **89.9582%** | **14** | **89.9582%** |
| **0.02** | 2 | 86.4017% | 2 | 85.5649% |
|  | 5 | 89.1213% | 5 | 87.4477% |
|  | 6 | 89.1213% | 6 | 88.4937% |
|  | 7 | 89.5397% | 7 | 88.2845% |
|  | **14** | **89.9582%** | **14** | **90.5858%** |
| **0.5** | 2 | 86.4017% | 2 | 85.5649% |
|  | 5 | 89.1213% | 5 | 87.4477% |
|  | 6 | 89.1213% | 6 | 88.2845% |
|  | 7 | 89.5397% | 7 | 88.2845% |
|  | **14** | **89.9582%** | **14** | **89.9582%** |
| **0.75** | 2 | 86.4017% | 2 | 85.5649% |
|  | 5 | 89.1213% | 5 | 88.4937% |
|  | 6 | 89.1213% | 6 | 88.7029% |
|  | 7 | 89.5397% | 7 | 89.5397% |
|  | **14** | **89.9582%** | **14** | **90.1674%** |
| **1** | 2 | 86.4017% | 2 | 85.5649% |
|  | 5 | 89.1213% | 5 | 87.4477% |
|  | 6 | 89.1213% | 6 | 88.7029% |
|  | 7 | 89.5397% | 7 | 89.5397% |
|  | **14** | **89.9582%** | **14** | **90.1674%** |

**Result:**

Naïve Bayes Classifier produces the accuracy of **94.9791%** for spam filter before removing the stop words and accuracy of **94.3515%** after removing the stop words from the mail data set.

In case of Logistic Regression, if the algorithm is iterated for fewer number of iterations (2, 5, 6, 7), the accuracy for spam filter before and after removing the stop words from the mail data set is almost the same. However, if the algorithm is run for greater number of iterations, the accuracy for spam filter after removing the stop words from the mail data set increases. Observed accuracy for 14 iterations after removing stop words is around **90.1674%** and before removing stop words is around **89.9582%.**

**Reason:**

The stop words are commonly used English words that are present in spam as well as ham mails. Therefore, they are not good choice for the feature selection. Removing stop words leave us with the useful feature set. As a result, we obtain higher accuracy for spam filter classification after removing the stop words.