CS 6375 – Mini Project 1 Report
Shubham Shekhar Jha (sxj220028)

Accuracy, Precision, Recall and F1 score

Discrete Naive-Bayes stats for enron1 dataset:
Accuracy = 74.58 %
Precision = 74.35 %
Recall = 76.31 %
F1 score = 75.32 %

Discrete Naive-Bayes stats for enron2 dataset:
Accuracy = 74.53 %
Precision = 61.37 %
Recall = 76.31 %
F1 score = 68.03 %

Discrete Naive-Bayes stats for enron4 dataset:
Accuracy = 84.85 %
Precision = 57.14 %
Recall = 76.31 %
F1 score = 65.35 %

Overall Discrete Naive-Bayes stats:
Accuracy = 91.34 %
Precision = 89.81 %
Recall = 95.04 %
F1 score = 92.35 %

Multinomial Naive-Bayes stats for enron1 dataset:
Accuracy = 88.62 %
Precision = 87.82 %
Recall = 90.13 %
F1 score = 88.96 %

Multinomial Naive-Bayes stats for enron2 dataset:
Accuracy = 88.78 %
Precision = 80.58 %
Recall = 90.13 %
F1 score = 85.09 %

Multinomial Naive-Bayes stats for enron4 dataset:
Accuracy = 93.22 %
Precision = 77.40 %
Recall = 90.13 %
F1 score = 83.28 %

Overall Multinomial Naive-Bayes stats:
Accuracy = 94.68 %
Precision = 95.05 %
Recall = 95.29 %
F1 score = 95.17 %

SGD Classifier stats for Bernoulli representation of enron1 dataset:
Accuracy = 91.40 %
Precision = 90.85 %
Recall = 97.06 %
F1 score = 93.85 %

SGD Classifier stats for Bernoulli representation of enron2 dataset:
Accuracy = 95.17 %
Precision = 98.50 %
Recall = 94.82 %
F1 score = 96.63 %

SGD Classifier stats for Bernoulli representation of enron4 dataset:
Accuracy = 98.13 %
Precision = 99.30 %
Recall = 94.07 %
F1 score = 96.62 %

Overall SGD Classifier stats for Bernoulli representation:
Accuracy = 95.09 %
Precision = 95.53 %
Recall = 95.53 %
F1 score = 95.53 %

SGD Classifier stats for Bag-of-Words representation of enron1 dataset:
Accuracy = 91.40 %
Precision = 90.85 %
Recall = 97.06 %
F1 score = 93.85 %

SGD Classifier stats for Bag-of-Words representation of enron2 dataset:
Accuracy = 95.17 %
Precision = 98.50 %
Recall = 94.82 %
F1 score = 96.63 %

SGD Classifier stats for Bag-of-Words representation of enron4 dataset:
Accuracy = 98.13 %
Precision = 99.30 %
Recall = 94.07 %
F1 score = 96.62 %

Overall SGD Classifier stats for Bag-of-Words representation:

Accuracy = 95.09 %
Precision = 95.53 %
Recall = 95.53 %
F1 score = 95.53 %

Questions:

1. SGD Classifier on either representation performs better than Discrete and Multinomial Naive-Bayes, because SGD Classifier is a linear classifier that does not have strict assumptions about the features. Thus, the model trained is more generalized and can perform better on new test data. I wasn't able to implement Logistic Regression, but I believe Logistic regression would've had a performance close to what the SGD classifier has.

2. No, SGD Classifier outperforms Multinomial Naive-Bayes in the bag-of-words representation, mostly due to the same reason that it has naive assumptions about the features, which may not always be true. However the MultinomialNB implementation lacks in performance only by a small amount. Without the perspective of any linear classifier, we can observe that MultinomialNB performs way better than DiscreteNB, mainly because it takes the count of words into account while learning parameters and also while training.

3. No, The performance of Discrete Naive-Bayes is the lowest compared to all the implementations in this project. SGD Classifier has a much better performance on the Bernoulli representation model, due to the same reason.

4. Cannot comment since I couldn't successfully implement MCAP Logistic Regression.

Citation

NLTK: Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.