**Homework 1**

**Execution:** Execution steps are in README.txt

**Performance Evaluation:**

**Discrete Naïve Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| HW Set | 94.97907 | 88.46153 | 92.74193 | 90.55118 |
| Enron1 | 94.51754 | 89.26174 | 93.66197 | 91.40893 |
| Enron4 | 95.02762 | 96.67519 | 96.42857 | 96.55172 |

**Multinomial Naïve Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| HW Set | 94.14225 | 85.38461 | 92.5 | 88.8 |
| Enron1 | 94.07894 | 86.57718 | 94.85294 | 90.52631 |
| Enron4 | 94.29097 | 96.93094 | 95.22613 | 96. 07097 |

**Logistic Regression Bernoulli**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score | Lambda |
| HW Set | 89.53974 | 92.55531 | 66.92307 | 77.67857 | 0.001 |
| Enron1 | 92.76315 | 94.61538 | 82.55033 | 87.63250 | 0.001 |
| Enron4 | 93.37016 | 91.56908 | 100 | 95.59902 | 0.3 |

**Logistic Regression Bag of Words**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score | Lambda |
| HW Set | 90.16736 | 86.72566 | 75.38461 | 80.65843 | 0.001 |
| Enron1 | 92.32456 | 92.53731 | 83.22147 | 87.63250 | 0.01 |
| Enron4 | 94.29097 | 92.85714 | 99.74424 | 96.17755 | 0.001 |

**SGD Bernoulli**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| HW Set | 94.97907 | 90.15151 | 91.53846 | 90.83969 |
| Enron1 | 95.39473 | 90 | 96.64429 | 93.20388 |
| Enron4 | 95.76427 | 94.44444 | 100 | 97.14285 |

**SGD Bag of Words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| HW Set | 94.97907 | 90.76923 | 90.76923 | 90.76923 |
| Enron1 | 94.95614 | 88.41463 | 97.31543 | 92.65175 |
| Enron4 | 95.58011 | 94.21686 | 100 | 97.02233 |

**Average over 3 Datasets for Comparison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| DNB | 94.84141 | 91.46615 | 94.27749 | 92.83728 |
| MNB | 94.17072 | 89.63091 | 94.19302 | 91.79909 |
| LRB | 91.89102 | 92.91326 | 83.1578 | 86.97003 |
| LRBOW | 92.26096 | 90.7067 | 86.11677 | 88.15616 |
| SGDB | 95.37936 | 91.53198 | 96.06092 | 93.72881 |
| SGDBOW | 95.17177 | 91.13357 | 96.02822 | 93.4811 |

**Hyper Parameter Tuning:**

For logistic regression, I wrote a cross validation function which randomly splits the dataset in a 70:30 ratio. It trains on different values of lambda. The lambda providing the highest accuracy is chosen and then fed into the model for the whole dataset. For number of iterations, I plotted a graph for iterations vs accuracy and looked for the point where change in accuracy vs change in iterations was low.

**Questions and Answers:**

Q1) The best performance is given by the SGD classifier with Bernoulli Representation. It outperforms all other algorithms in Accuracy, Recall and F1 Score and has ranks 3rd in the precision values.

Q2) No. Multinomial Naïve Bayes as compared to BOW Logistic classifier outperforms it in accuracy, recall and f1 score but lacks behind in precision. But compared to SGD BOW, SGD outperforms Multinomial in every metric.

Q3) No. By examining the values Discrete Naïve Bayes outperforms Logistic Regression in accuracy, recall and F1 score but lacks in precision score by a little bit. But SGD using the Bernoulli representation outperforms every algorithm on every metric.

Q4) No. SGD classifiers on both representations outperform logistic regression in nearly every metric. The difference between the performance of SGD on Bernoulli is vastly superior to Logistic regression but SGD on Bag of Words has a comparable but higher performance.