Assignment-3 Reports

Random seed for both Naïve Bayes and Logistic Regression are set to 42 that guarantee the result reproducibility.

# Naïve Bayes

Naïve Bayes algorithm is implemented with and without stopword removal to classify “spam” and “ham” emails.

Without stopword removal, the model performance is observed as following:

* Training accuracy: 0.9957
* Validation accuracy: 0.9497

With stopword removal, the model performance is observed as following:

* Training accuracy: 0.9914
* Validation accuracy: 0.9393

# Logistic Regression

Hyperparameters for training Logistic Regression are:

* Iteration = 100
* Learning rate = 0.01
* Lambdas for L2 regularizer ranges from 0.0001 to 0.001 with 0.0001 step.

Without stopword removal, the model performance is observed as following:

* Lambda = 0.0001
  + Training accuracy = 1.0
  + Validation accuracy = 0.884 9
* Lambda = 0.0002
  + Training accuracy = 1.0
  + Validation accuracy = 0.8849
* Lambda = 0.0003
  + Training accuracy = 1.0
  + Validation accuracy = 0.8912
* Lambda = 0.0004
  + Training accuracy = 0.9978
  + Validation accuracy = 0.8912
* Lambda = 0.0005
  + Training accuracy = 0.998
  + Validation accuracy = 0.8933
* Lambda = 0.0006
  + Training accuracy = 1.0
  + Validation accuracy = 0.887
* Lambda = 0.0007
  + Training accuracy = 1.0
  + Validation accuracy = 0.887
* Lambda = 0.0008
  + Training accuracy = 1.0
  + Validation accuracy = 0.887
* Lambda = 0.0009
  + Training accuracy = 1.0
  + Validation accuracy = 0.887
* Lambda = 0.001
  + Training accuracy = 1.0
  + Validation accuracy = 0.8849

With stopword removal, the model performance is observed as following:

* Lambda = 0.0001
  + Training accuracy = 0.870
  + Validation accuracy = 0.824
* Lambda = 0.0002
  + Training accuracy = 0.872
  + Validation accuracy = 0.788
* Lambda = 0.0003
  + Training accuracy = 0.788
  + Validation accuracy = 0.0.757
* Lambda = 0.0004
  + Training accuracy = 0.762
  + Validation accuracy = 0.7426
* Lambda = 0.0005
  + Training accuracy = 0.768
  + Validation accuracy = 0.757
* Lambda = 0.0006
  + Training accuracy = 0.855
  + Validation accuracy = 0.813
* Lambda = 0.0007
  + Training accuracy = 0.859
  + Validation accuracy = 0.0.814
* Lambda = 0.0008
  + Training accuracy = 0.859
  + Validation accuracy = 0.812
* Lambda = 0.0009
  + Training accuracy = 0.771
  + Validation accuracy = 0.749
* Lambda = 0.001
  + Training accuracy = 0.866
  + Validation accuracy = 0.831

Look at the accuracy results above, the Logistic Regression without stopword removal performs better than with stopword removal because:

* The stopwords may contribute to the difference between “spam” and “ham” emails. Before doing this, the data analysis was not done. Hence, removing stopwords may reduce distinct words that separate “spam” and “ham” emails.

Look at the accuracy results above, the Logistic Regression without stopword removal suffers from overfitting more than that with stopword removal:

* Again, this observation could be due to the removal of distinct stopwords that intensively generalize Logistic Regression.