Brendan Martel
Bxm240013
Project 1 Final Report

1. Scoring reporting

1.1 Enron1

```
accuracy = 0.8048245614035088
precision = 0.9838709677419355
recall = 0.40939597315436244
F1 score = 0.5781990521327015
```

Multinational Bayes

```
accuracy = 0.9429824561403509
precision = 0.8819875776397516
recall = 0.9530201342281879
F1 score = 0.9161290322580646
```

Bernoulli Naive Bayes

```
accuracy = 0.8552631578947368
precision = 0.8267716535433071
recall = 0.7046979865771812
F1 score = 0.7608695652173914
```

Logistical Regression
BOW

```
accuracy = 0.8267543859649122
precision = 0.8888888888888888
recall = 0.5369127516778524
F1 score = 0.6694560669456066
```

Logistical Regression
Bernoullis

1.2 Enron2

```
accuracy = 0.8472803347280334
precision = 1.0
recall = 0.43846153846153846
F1 score = 0.6096256684491979
```

Multinational Bayes

```
accuracy = 0.9100418410041841
precision = 0.7604790419161677
recall = 0.9769230769230769
F1 score = 0.8552188552188552
```

Bernoulli Naive Bayes

```
accuracy = 0.8556485355648535
precision = 0.7961165048543689
recall = 0.6307692307692307
F1 score = 0.703862660944206
```

Logistical Regression
BOW

```
accuracy = 0.8284518828451883
precision = 0.8428571428571429
recall = 0.45384615384615384
F1 score = 0.59
```

Logistical Regression
Bernoullis

1.3 Enron4

```
-------------------------------------
accuracy = 0.8895027624309392
precision = 0.8669623059866962
recall = 1.0
F1 score = 0.9287410926365796
-------------------------------------
```

Multinational Bayes

```
-------------------------------------
accuracy = 0.9631675874769797
precision = 0.9649122807017544
recall = 0.9846547314578005
F1 score = 0.9746835443037974
-------------------------------------
```

Bernoulli Naive Bayes

```
-------------------------------------
accuracy = 0.8858195211786372
precision = 0.931758530183727
recall = 0.907928388746803
F1 score = 0.9196891191709844
-------------------------------------
```

Logistical Regression
BOW

```
-------------------------------------
accuracy = 0.8471454880294659
precision = 0.9350282485875706
recall = 0.8465473145780051
F1 score = 0.8885906040268458
-------------------------------------
```

Logistical Regression
Bernoullis

2. Hyper parameter tuning

For the tuning of the hyper parameters I started with an educated guess on my parameters. I started with .1 for learning rate and regularization constant. Testing with the 70/30 split I found that while the overall stats increased with iterations. I ran from 5 all the way to 100 and while it seemed to increase as I approached 100 iterations that 20 seemed to a reasonable number both for the amount of time that it took to run and the performance that I was getting out of the model. After finding the iterations I tuned the learning rate and regularization constant in a similar fashion starting at my .1 original guess and going up and down per iteration until arriving at .01 working best for both parameters.

3. Questions

3.1

Overall Naive Bayes performed better more specifically Bernoulli Naive Bays performed consistently the best overall. First of all I think the NB was more suited to the smaller dataset that we had. Something that you may observer is that Enron 4 which had a vocabulary of about double Enron 2 or 3 performed the best with LR. I also think that BNB performed the best overall because testing for absence or presence was the best approach for determining spam vs not spam as once the stop words were removed there seemed to be less overlap.

3.2

Enron 4 BNB was the best performer of all of the tests ran. I think this has to do with the two factors mentioned in the above answer. However, I also think that there being a larger dataset for enron4 overall also helped to make this combination the highest performers overall.

3.3

Yes it did perform better overall on BOW representations of the data with the exception of Enron 4 where it was barely beat by the NB version. I think that part of it is that NB struggles with continuous features whereas LR is more suited for them. So the way that the BOW approach encodes data is more suited to LR then NB.

3.4

Yes it absolutely did. As discussed above I believe that the Bernoulli representation of the date plays directly into the strengths of NB as a classifier. This is because of the binary representation of the data. While this is great for the NB it doesn't work as well with LR leading to the large discrepancy.