**Naïve Bayes**

**Data Exploration Questions:**

1. What is the number of attributes in each Dataset?
   1. Optdigits Dataset: 64 input+1 class attribute
   2. Amazon Reviews Dataset: 1 input(review)+1 class attribute(rating)
2. What is the number of observations?
   1. Optdigits Dataset:
      1. Training = 3823
      2. Test = 1797
   2. Amazon Reviews Dataset:
      1. Training = 146824
      2. Test = 36707
3. What is the mean and standard deviation of each attribute?
   1. Optdigits Dataset:

****

* 1. Amazon Reviews Dataset:
     1. Training(Rating): Mean = 4.121798 SD = 1.284965
     2. Test(Rating): Mean = 4.115046 SD = 1.285226

1. What is the distribution of the different classes in each of the datasets?
   1. Optdigits Dataset:
      1. Class: No of examples in training set
      2. 0: 376
      3. 1: 389
      4. 2: 380
      5. 3: 389
      6. 4: 387
      7. 5: 376
      8. 6: 377
      9. 7: 387
      10. 8: 380
      11. 9: 382
      12. Class: No of examples in testing set
      13. 0: 178
      14. 1: 182
      15. 2: 177
      16. 3: 183
      17. 4: 181
      18. 5: 182
      19. 6: 181
      20. 7: 179
      21. 8: 174
      22. 9: 180
   2. Amazon Reviews Dataset:
      1. Training:
         1. 12146
         2. 9040
         3. 13364
         4. 26509
         5. 85765
      2. Test:
         1. 3037
         2. 2270
         3. 3415
         4. 6696
         5. 21289

**Results – Optdigits Dataset**

* Result for running Optdigits dataset on multiple Naïve Bayes classifiers in Weka.
* The different classifiers that were used in this were
  + Naïve Bayes



* + Naïve Bayes Multinomial



* + Naïve Bayes Multinomial Updateable



* + Naïve Bayes Updateable



* + Bayes Network



**Amazon Reviews Dataset – Results**

* Results for running **Naïve Bayes using Multinomial** by scikit-learn of Python, using Bag of Words with TF-IDF (Short Term Frequency – Inverse Document Frequency) on Amazon Reviews Dataset.
* Number of features from Tf-IDF = 2500 from a dummy dataset of containing Train = 8000 instances, Test = 2000 instances.
* Naïve Bayes Parameters and Results:

Using Multinomial Naïve Bayes:



Using Multinomial Naïve Bayes for Alpha=1.0.



* Using cross-validation on the training set, I was able to obtain the following cross-validation scores for a 10 fold method. Last array gives the cross validation scores.

