**K-Nearest Neighbours**

**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 IBk on WEKA, with K=10, linear search algorithm.

(Open files with sublime text)



* Result for running Optdigits dataset on IBk on WEKA, with K=21, linear search algorithm.



* Result for running Optdigits dataset on IBk on WEKA, with K=25, linear search algorithm and using cross validation



* Result for running Optdigits dataset on IBk on WEKA, with K=30, linear search algorithm and using Distance Weighting.



* Result for running Optdigits dataset on IBk on WEKA, with k=30 and using Euclidean distance and KDtree as algorithm.



* Result for running Optdigits dataset on IBk on WEKA, with k=30 and using Filtered Neighbours Search and using Weighted Distance.



**Amazon Reviews Dataset – Results**

* K=20

0.625125628141

[[ 26 22 9 2 104]

[ 14 11 11 6 78]

[ 5 8 14 18 129]

[ 1 2 8 45 288]

[ 11 6 2 22 1148]]

* K=25

0.621608040201

[[ 39 10 9 2 103]

[ 17 6 8 6 83]

[ 9 4 11 14 136]

[ 3 0 8 42 291]

[ 14 3 7 26 1139]]

* K=30

0.624120603015

[[ 34 20 11 2 96]

[ 15 14 10 10 71]

[ 4 9 22 19 120]

[ 3 5 19 45 272]

[ 8 9 13 32 1127]]

* K=35

0.610050251256

[[ 36 28 5 4 90]

[ 16 15 13 12 64]

[ 7 16 16 26 109]

[ 5 7 18 58 256]

[ 16 11 20 53 1089]]

* K=40

0.633567839196

[[ 39 25 6 2 91]

[ 20 16 10 9 65]

[ 9 15 18 18 114]

[ 7 7 15 57 258]

[ 17 14 18 49 1091]]

* K=45

0.647638190955

[[ 31 15 8 1 108]

[ 18 7 6 6 83]

[ 5 7 12 11 139]

[ 4 1 5 39 295]

[ 7 4 0 18 1160]]

* Using Hamming Distance

0.624120603015

[[ 34 20 11 2 96]

[ 15 14 10 10 71]

[ 4 9 22 19 120]

[ 3 5 19 45 272]

[ 8 9 13 32 1127]]

* Using Euclidean Distance

0.645125628141

[[ 26 22 9 2 104]

[ 14 11 11 6 78]

[ 5 8 14 18 129]

[ 1 2 8 45 288]

[ 11 6 2 22 1148]]

* Using Minowski Distance

0.624120603015

[[ 34 20 11 2 96]

[ 15 14 10 10 71]

[ 4 9 22 19 120]

[ 3 5 19 45 272]

[ 8 9 13 32 1127]]

* Using Manhattan Distance

0.633567839196

[[ 39 25 6 2 91]

[ 20 16 10 9 65]

[ 9 15 18 18 114]

[ 7 7 15 57 258]

[ 17 14 18 49 1091]]