CSE 472 Report Submission

Submitted By:
Tanjim Bin Faruk
Stduent ID: 1505082

October 13, 2020

Bangladesh University of Engineering and Technology
(BUET)

1 Validation Dataset

1.1 K Nearest Neighbor

Heurisitic	k = 1	k = 3	k = 5
Hamming Distance	42.136%	39.409%	38.182%
Euclidean Distance	57.909%	55.955%	55.591%
Cosine Similarity	81.136%	83.227%	83.773%

Table 1: KNN Accuracy on Validation Dataset

1.2 Naive Bayes

Iteration	Smoothing Factor	Accuracy
1	0.1	89.182%
2	0.2	89.000%
3	0.3	88.909%
4	0.4	88.727%
5	0.5	88.818%
6	0.6	88.591%
7	0.7	88.318%
8	0.8	88.045%
9	0.9	87.864%
10	1.0	87.682%

Table 2: Naive Bayes Accuracy on Validation Dataset

2 Test Dataset

Iteration	KNN (k = 5, C.S)	NB (S.F = 0.1)
1	83.636%	90.000%
2	90.000%	95.455%
3	86.364%	88.181%
4	84.545%	85.455%
5	88.182%	92.727%
6	88.182%	89.091%
7	88.182%	92.727%
8	90.000%	92.727%
9	85.455%	85.455%
10	87.273%	81.818%
11	86.364%	90.000%
12	89.091%	90.909%
13	80.909%	82.727%
14	84.545%	86.364%
15	83.636%	89.091%
16	81.818%	87.273%
17	83.636%	89.091%
18	81.818%	90.000%

19	84.545%	86.363%
20	80.909%	83.636%
21	86.364%	84.545%
22	85.455%	90.000%
23	88.182%	91.818%
24	83.636%	90.909%
25	86.363%	83.636%
26	84.545%	89.091%
27	86.364%	90.000%
28	79.091%	83.636%
29	86.364%	87.273%
30	80.909%	88.182%
31	87.273%	90.000%
32	86.364%	91.812%
33	87.273%	91.812%
34	90.000%	94.545%
35	82.727%	87.273%
36	87.273%	87.273%
37	76.364%	87.273%
38	80.000%	89.091%

39	86.364%	89.091%
40	89.091%	92.727%
41	82.727%	89.091%
42	89.091%	90.909%
43	85.455%	85.455%
44	88.182%	91.818%
45	85.455%	92.727%
46	90.909%	90.000%
47	82.727%	92.727%
48	85.455%	82.727%
49	87.273%	85.455%
50	89.091%	90.000%
N = 50	KNN Avg. = 84.872%	NB Avg. = 88.000%

Table 3: Accuracy on Test Dataset

3 Computation of T-statistics

To perform the T-statistics, firstly we need to describe the null hypothesis and the alternative hypothesis.

- Null Hypothesis (H_0): There isn't any significant difference between the performance / accuracy between the K Nearest Neighbor and Naive Bayes algorithm.
- Alternative Hypothesis (H_1): There is a significant difference between the performance / accuracy between the two algorithms.

To reject/fail to reject H_0, we have to compute the t-statistics value or the p-value. Using the scipy's *ttest_rel* function, we can compute the t-stat and p-value from the 50 observed samples.

Scipy ttest_rel function

```
stat, p_value =
    ttest_rel(knn_accuracy_values, nb_accuracy_values)
```

The calculated valuess are (stat, p_value) = $(-7.463, 1.275 \times 10^{-9})$. Now at given significance levels:

- Significance level $\alpha = 0.05$: As p-value $< \alpha$, we reject the null hypothesis and as average accuracy difference is negative (mean KNN $<$ mean NB), we decide with a confidence of $(1 - p) = 99.99\%$ that NB outperforms KNN.
- Significance level $\alpha = 0.01$: As p-value $< \alpha$, we reject the null hypothesis and as average accuracy difference is negative (mean KNN $<$ mean NB), we decide with a confidence of $(1 - p) = 99.99\%$ that NB outperforms KNN.

- Significance level $\alpha = 0.005$: As p-value $< \alpha$, we reject the null hypothesis and as average accuracy difference is negative (mean KNN $<$ mean NB), we decide with a confidence of $(1 - p) = 99.99\%$ that NB outperforms KNN.