Report on Assignment-2

Text Classifier

1105021

1105025

T-test Result:

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Mean | Standard Deviation | Total Runs |
| Naïve Bayes | 73.4% | 0.5385% | 10 |
| KNN | 81.95% | 1.2539% | 10 |

Let,

= Mean Accuracy of Naïve Bayes

= Population average of accuracy of Naïve Bayes = 73.4

= = Population Variance of Naïve Bayes = = 0.29

=Number of runs of Naïve Bayes = 10

= Mean Accuracy of KNN

= Population average of accuracy of KNN = 81.95

= Population Variance of KNN = = 1.5723

=Number of runs of KNN = 10

Let, Null Hypothesis : - >=0

And the Alternative Hypothesis: : - <0

Now,

Putting the corresponding values , t = -19.8126

Now, degree of freedom =

Putting the corresponding values, d.o.f = 12

Now, the chart with t-statistic at significance levels of 0.005, 0.01 and 0.05 is given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Level of significance, alpha |  | Critical Value , t(dof,1-alpha) | Comment |
| 0.005 | 19.8126 | t(12,0.995) = 4.318 | Reject . KNN is better |
| 0.01 | 19.8126 | t(12,0.99) = 2.681 | Reject . KNN is better |
| 0.05 | 19.8126 | t(12,0.95) = 1.782 | Reject . KNN is better |

Question #2:

In the distance measuring of K-NN algorithm among 3 versions , cosine similarity gives better performance. Because the Hamiltonian and Euclidean way of measuring the distance gives the distance of two points in the vector space. Then we took the lowest distance to predict in which class our target document Dt falls.

In the above case it does not depend whether the relative distance with respect to origin or a certain point is proportion to one another. That means one point in the space may have same proportion of dimensional axis value to other but have the maximum distance in sense Euclidean or Hamiltonian .

But in case of cosine similarity the angular distances are measured. That's why the above mentioned problem does not arise. The point in the space which have same ratio of axes value is selected as most probable result.

Question #3:

NB gives better performance than K-NN algorithm when we have more data to train. Because K-NN calculates distances of the target document from known trained set of data. Again calculating probability gives efficient prediction than calculating distance if there is no conditional probability or co relation among words.

But for small data to train NB gives poorer result.