# Report on Assignment-2

Submitted By:

Shamik Roy(1105053)

Md Mazharul Islam(1105013)

## Data:

**Naïve Bayes:**

Mean Accuracy: 93% (approximately)

Standard Deviation: 1.7% (approximately)

**K-NN:**

Mean Accuracy: 82% (approximately)

Standard Deviation: 3.6% (approximately)

## t-statistics:

**Null Hypothesis:** Mean Accuracy(NB) – Mean Accuracy(KNN) >= 12

**Alternative Hypothesis:** Mean Accuracy(NB) – Mean Accuracy(KNN) < 12

Result: t(absolute) = 1.4747, d.o.f = 71

At significance level of 0.05: t(critical) = 1.66715

So, t(absolute) < t(critical)

The Null Hypothesis cannot be rejected and the Alternative Hypothesis cannot be accepted.

At significance level of 0.01: t(critical) = 2.3812

So, t(absolute) < t(critical)

The Null Hypothesis cannot be rejected and the Alternative Hypothesis cannot be accepted.

At significance level of 0.005: t(critical) = 2.64845

So, t(absolute) < t(critical)

The Null Hypothesis cannot be rejected and the Alternative Hypothesis cannot be accepted.

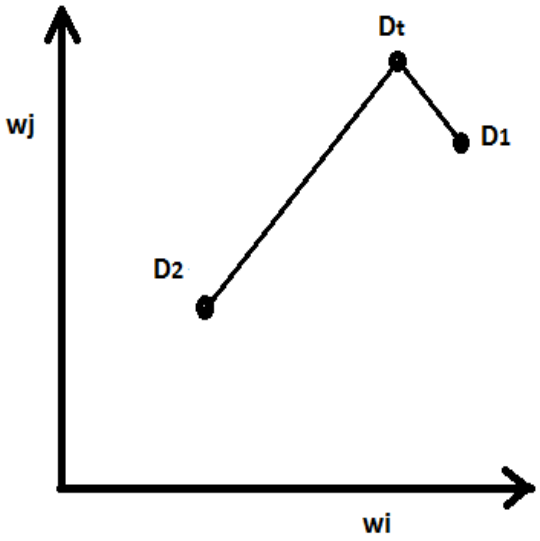
**So, with higher confidence level we can conclude that the actual mean accuracy of Naïve Bayes algorithm is at least 12% more than the actual mean accuracy of K-NN algorithm. Hence, Naïve Bayes is better than K-NN.**

## Questionnaire:

1. Among the three different measures of the K-NN, which one shows maximum accuracy? Why does it work better than other two?

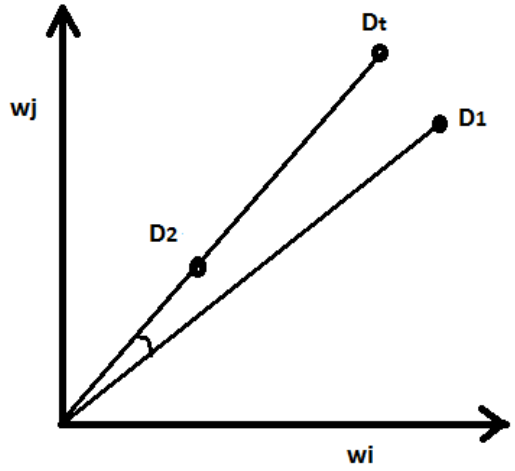
Answer: Cosine similarity shows maximum accuracy.

Explanation: In case of Hamming distance and Euclidian distance technique the pairwise distances between the test point and all other training points are measured and the training point with minimum distance from the test point is taken as result. For example, in the figure below, test point D1 will be selected as it is the nearest neighbor of Dt in Hamming distance and Euclidian distance techniques.




But it can be seen that training point D2 is almost in the same line joining the origin and the point Dt(shown in the figure below). That means training point Dt has the same proportion of value along wi and wj as point Dt. Because Dt may be a document of the same class as D2 but of a larger size. So, despite of not being the nearest neighbor, D2 is the best estimated class for Dt. In this point, the other two techniques lags behind the Cosine Similarity technique. Cosine similarity determines the angular difference between the vectors and takes the vector

which has the least angular difference as the figure below. So D2 will be selected in Cosine similarity technique.




2. Which one between the k-NN (the best measure among three) and the NB works better? Why?

Answer: NB works better than K-NN(Cosine similarity).

Explanation: K-NN takes decision in basis of the distances between points(either linear or angular). But in Naïve Bayes technique, the decision is taken in basis of prior probabilities. The prior probabilities are more reliable as they are clear evidences of being an attribute in a certain class. As a result, it yields better performance.

==End==