Course: CSE 472
Student ID: 1505022
Name: Mahim Mahbub

Report on Machine Learning Offline 2

Accuracies on Validation Set

I. KNN: Accuracies (%) for K = 1, 3 and 5 are shown.

Similarity Measure	K = 1	K = 3	K = 5
Hamming Distance	39.68181818	40.04545455	42.68181818
Euclidean Distance	55.95454545	54.59090909	53.59090909
Cosine Similarity	78.86363636	79.18181818	80.5

II. Naive Bayes: Accuracies (%) for **top 10** smoothing factors (α) are shown. Table is sorted in descending order of accuracy.

Smoothing Factor (α)	Accuracy(%)
0.04	88.5
0.02	88.5
0.03	88.45454545
0.05	88.45454545
0.07	88.45454545
0.01	88.40909091
0.15	88.40909091
0.24	88.40909091
0.18	88.40909091
0.17	88.40909091

Accuracies on Test Set (50 iterations using best models)

- For KNN, best model combination was using cosine-similarity (TF-IDF) with K = 5
- For Naive Bayes, best model combination was using smoothing factor (α) = 0.04

Accuracy(%) using KNN (K = 5)	Accuracy(%) using Naive Bayes (α = 0.04)
86.36363636	88.18181818
75.45454545	84.54545455
76.36363636	90.90909091
85.45454545	90
80.90909091	91.81818182
78.18181818	86.36363636
72.72727273	90.90909091
81.81818182	91.81818182
77.27272727	90.90909091
81.81818182	89.09090909
81.81818182	90
81.81818182	92.72727273
81.81818182	88.18181818
80.90909091	90.90909091
83.63636364	90
80	88.18181818
77.27272727	91.81818182
83.63636364	86.36363636
81.81818182	90.90909091
86.36363636	90.90909091
74.54545455	85.45454545
80	87.27272727
86.36363636	92.72727273
86.36363636	94.54545455
80	83.63636364
78.18181818	93.63636364
82.72727273	94.54545455
80	89.09090909
83.63636364	94.54545455
82.72727273	90.90909091
77.27272727	89.09090909

82.72727273	92.72727273
80.90909091	89.09090909
82.72727273	87.27272727
84.54545455	91.81818182
81.81818182	95.45454545
79.09090909	86.36363636
80	90.90909091
75.45454545	90.90909091
79.09090909	89.09090909
77.27272727	88.18181818
69.09090909	82.72727273
83.63636364	92.72727273
84.54545455	89.09090909
80	88.18181818
77.27272727	89.09090909
86.36363636	90
80.90909091	88.18181818
85.45454545	89.09090909
73.63636364	87.27272727

Summarized results on test-set :

Method	Mean Accuracy(%)	Std-dev	Std Error	Minimum Accuracy(%)	Maximum Accuracy(%)	Number of Iterations
KNN	80.63636364	3.832225256	0.5419584931	69.09090909	86.36363636	50
NB	89.76363636	2.77420229	0.3923314504	82.72727273	95.45454545	50

Statistical Consistency (T-Statistics):

Using dependent t-test for paired samples i.e. ttest_rel(NB, KNN), we obtain the following:

Statistic	p-value
17.570878565892993	8.56639952050839e-23

We know that **ttest_rel** is a two-sided test for the null hypothesis that two **related** or repeated samples have **identical average** (expected) values.

This value of p, satisfies for all significance levels 0.005, 0.01 and 0.05 since p = **8.56639952050839e-23** is smaller than **all** significance values of 0.005, 0.01 and 0.05.

With Naive-Bayes (NB) having a higher mean accuracy than KNN, we can conclude that for this dataset, NB shows a statistically consistent higher accuracy than KNN for **all** significance levels of 0.005, 0.01 and 0.05.

Hence, NB is better (being statistically consistent on all significance levels) with respect to KNN.

Performance Analysis of each methods:

- ❏ Hamming distance similarity method (for KNN) shows the worst performance compared to Euclidean and cosine-similarity because having binary representation for each word does not capture the importance of that word and overall is not a good representation of the text document.

- ❏ Euclidean distance similarity method performs worse compared to TF-IDF cosine-similarity because TF-IDF finds extra information per label/class/topic.
  Both TF and IDF are separately calculated and both of these information are used.
  More frequent words are given less weight in TF-IDF calculation.
  Additionally, IDF information considers all documents in its calculation for each word.
  Euclidean metric on the other hand treats all documents independently.

- ❏ Bayesian method i.e. Naive Bayes (Multinomial Naive Bayes in this case) has the best performance compared to all KNN similarity based methods.
  In KNN, each document is taken and considered independently for neighbor calculation.
  Naive Bayes on the other hand considers all documents' information under one label/class/topic. Additionally, prior probability for each label is considered. If documents are not uniformly distributed, this extra information is also utilized.