Lab-1-Report

Experiment 1

The better distance measure from the experiments performed is the cosine distance. The testing accuracy measured using cosine distance was about 10% better than Euclidean distance at 96.3% and 86.5% respectively. The standard deviation was also smaller with the cosine distance compared to the Euclidean distance.

EXPLAIN WHY COSINE IS BETTER HERE

Experiment 2

- discuss the difference between training and testing accuracies within the experiment, and why this is the case
- analyse the effect of k in the experiment (comment on how the *bias* and *variance* changes depending on k)

The error in the model where we test the data on the same data used for training has a logarithmic shape to it over values of k, where at k = 1 we have 0 error. This is because the training data is simply mapped 1 to 1 to the same data it was trained on. As k increases, data points may not evaluate to the same class as it was trained on as other near data points may change the class assigned, increasing the error and bias.

Because of this reason the variance tends to increase as k increases as well.

For the model that was tested using separate data from the training data we have a more linear looking shape when evaluating the error over values of k. overall the error/bias tends to be higher, and the variance is also a bit larger. One can also note

that the variance stays quite consistent throughout.

This is because in the former model, each data point has at least 1 of it's closest neighbours (itself) in the correct class. The latter model has no guarantees. Of course as k increases this fact becomes less relevant, we can see that at k = 50 the error rate between the two is very close.

Experiment 3

- which classes should the new 5 articles belong to, and why? (based on their content) - can the classifier make appropriate class predictions for these 5 articles? - after repeating the testing for the 5 classes 6 times each, comment on the classifiers performances and what the consequences are of limited data for the 'sports' class - self-learn zero-shot/few-shot learning, then link the concepts to the experiment - is the model performing zero or few shot learning? explain why

From reading the 5 new articles, appropriate classifications might look something like:

- sp0: interest
- sp1: interest
- sp2: trade
- sp3: interest
- sp4: interest

  The classifier fails at making an appropriate prediction, assigning seemingly

random classes.