# Data Mining I: Homework 1
# Distance Functions on Vectors

Toby Law

October 15, 2019

## Exercise 1.b

**Report and discuss any abnormalities in the results. For example, do all distance functions report a lower average distance when comparing documents of the same group vs. documents from different groups?**

None of the distance functions were able to consistently report a lower average distance when comparing documents of the same group vs. documents from different groups. This might be because that distance functions are based on quantitative measures of similarity: in this case it measures the similarity of documents depending on the frequency of the words in the document with respect to the entire dataset. Within the universe of words, some might be special terms that occur in some topics/contexts more often than others, but it might also include words that grammatically common, such as articles, tenses, pronouns etc. The frequency of which these grammatically common words occur is independent from the context of the document, but fluctuations could introduce irrelevant noise. Distance functions are also qualitatively-ineffective: they do not discriminate different pairs of documents well according to their semantic relationships, nor are weights imposed on words that semantically more important and better distinguishes newsgroups in our analysis. Therefore, the incorporation of noisy features in our analysis might be why the distance functions were not able to consistently report lower average distances when comparing documents of the same group vs. different groups.

Out of all the newsgroups, comp.sys.mac.hardware is the only newsgroup where all six metrics gave a lower average distance between documents within the group versus that between them and other different groups. This might be because the documents in the comp.sys.mac.hardware newsgroup contains high frequencies of certain words in the universe that are rare in other groups, hence are exceptionally able to discriminate them from documents in the other groups. Yet we can also see that for the newsgroup talk.politics.guns,

a higher average distance is quite often found for documents within that group compared to the average distance between it and other groups. This suggests that, coincidentally, there are very few to no words in the universe that are able to discriminate documents in the talk.politics.guns newsgroup from the others.

## Exercise 1.c

**Which metric seems to provide, on average, the best separation between groups? Explain why this is the case.**

The Hamming distance seems to provide the best separation between the groups on average. It is able to report the largest distance contrasts between the group comparisons, allowing us to assess to what extent each group comparison differs from another. This better ability to separate groups of documents is because the Hamming distance metric determines the distance between two documents only based on the occurrence of words, but not their relative frequencies. Only if a certain word exists in one document but not another, will this be recorded by the Hamming distance metric. Essentially this filters out most of the noise caused by trivial words which occur in most documents, and brings emphasis to words that are semantically more important to certain newsgroups over others.

## Exercise 1.d

**The Manhattan and Euclidean distances are also known as L1 and L2 norms respectively. In general, what behavior can we expect about the L1 vs. L2 norms as the dimensionality of the data increases? Is this behavior observed in our dataset? If not, why not?**

High dimensional datasets are more likely to contain more diverse features, including those which are irrelevant. In this dataset, irrelevant features correspond to words that are generally frequent but do not really affect the semantic meaning of the documents, like 'a', 'an', 'the', 'is'. Compared to L1 norms, L2 norms, with its additive effect and the sum of squares approach, tends to emphasize larger differences while smaller differences between vectors are diminished. Hence, L2 norms tends to magnify the noise of irrelevant features more than L1, causing the distance contrast between vectors measured by L2 norms to become smaller as dimensionality of the data increases. This behaviour is observed in our dataset, which has quite high dimensionality. We could see that for L2 norms, the average distances for inter-group vs. intra-group comparisons are very similar, only differing by 0.01-0.1, while those for L1 norms differ by at least 0.3-3.