Homework 1

Lucas Cosier

Data Mining I

ETH zürich

October 29, 2022

Type	Manhattan	Hamming	Euclidean	Chebyshev	Minkowski d=3	Minkowski d=4
Mean						
intra	11.665	102.0975	1.3	0.3675	0.6875	0.5275
inter	12.609	135.779	1.277	0.372	0.665	0.509
overall	12.272	124.070	1.286	0.372	0.674	0.517
Variance						
intra	2.21	481.8	0.0054	0.0018	0.00168	0.00141
inter	1.94	922.5	0.0013	0.0007	0.0004	0.0004
overall	2.11	1020.5	0.00256	0.001	0.0009	0.0008

Table 1: Reported distance means and variances for each metric. Intra means within the same news group, e.g. `comp.graphics:comp.graphics`, and inter means between different groups.

Problem 1

a) **Solution.** Table 1 summarizes the statistics gotten by running the provided scripts with the necessary implementation required to solve task 1.a).

b) **Solution.** Per Table 1, lowest intra (and overall) mean reported for *all* distances was for `comp.sys.mac.hardware`, and highest using Chebyshev and Minkowski (both) for `rec.autos`. This can mean that on average, hardware articles have similar topics, while papers about autos tend to be more varied, or that the magnitude between texts in the automotive area is bigger (bigger differences in text length). One abnormality is that not all distance measures agree that intra means should be lower than their inter counterparts. It can be seen that L_2, L_3, L_4 – norms (from here on out we will refer to Minkowski with $d = 3$ and $d = 4$ as L_3, L_4 and the usual Manhattan and Euclidean as L_1, L_2) report higher intra means than both the inter and total samples. This can be the case if there exist outliers.

c) **Solution.** One would be inclined to choose the Hamming distance, since it has the highest variance between groups, of 9.225e+02 (Table 1), with a standard deviation (not reported here) of 3.03e+01. Intuitively, the higher the variance, the more spread out are the clusters of documents, and hence based on variance alone one could reason the best separation is given by this metric. However, for the Hamming distance, it could also be the case that large variance can be attributed to the fact that document lengths are not equal. Therefore, one other candidate would be the L_1 metric.

d) **Solution.** For tf-idf vectors, the range of the cosine distance is between 0 and 1, since the term frequency cannot be negative. Usually it will range between -1 and 1, since the highest (or lowest) similarity is achieved by vectors which are identical (or point in different directions). Since we know a dot product between orthogonal vectors is 0, this will represent decorrelation. The intuition is the same for increasing dimensionality, it effectively represents the angle between

two vectors. This metric is more robust in higher dimensions since it doesn't take the magnitude into account

e) **Solution.** The L_1 norm is more robust to outliers since it doesn't square the differences. This means that the outliers will have less of a contribution when fitting a model to the data. On the other hand, data becomes more sparse as the number of dimensions increases (as the volume it occupies increases with each dimension), and the average distance between vectors increases, as the ratio between the nearest and farthest points approaches 1. Because all points are uniformly distant from each other, the notion of similarity is meaningless in these metrics. Again, because of the lack of the square, L_1 is more robust against this phenomenon than the L_2 – norm. The reported variance for L_2 is of several orders of magnitude smaller than L_1, supporting the claim that the dataset is high dimensional. In this case, the L_2 – norm would provide poorer clusters (separation) than L_1.

Problem 2

a) (a) **Solution.** is a metric.

(b) **Solution.** not a metric. $x = [-1,1]^\top, y = [0,1]^\top \Rightarrow \sum_{i=1}^{2} x_i y_i (x_i - y_i)^2 = -1 < 0$

(c) **Solution.** is a metric.

(d) **Solution.** not a metric. Take $x, y \in \left\{ \left[\frac{1}{3}, \frac{2}{3}\right]^\top, \left[\frac{2}{3}, \frac{1}{3}\right]^\top \right\} \Rightarrow \sum_i x_i \log(\frac{x_i}{y_i}) = \frac{1}{3}\log(\frac{1}{2}) + \frac{2}{3}\log(\frac{1}{2}) = \log(\frac{1}{2}) < 0$

(e) **Solution.** is a metric.

b) (a) **Solution.** we have that the Minkowski distance is $d(x,y) = \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{p}}$. Therefore, for $a \in \mathbb{R}$, $d(ax, ay) = \left(\sum_i |ax_i - ay_i|^p\right)^{\frac{1}{p}} = \left(\sum_i |a(x_i - y_i)|^p\right)^{\frac{1}{p}} = \left(\sum_i |a|^p |x_i - y_i|^p\right)^{\frac{1}{p}}$ where the last equality follows from multipicativity. Then, to conclude, $d(ax, ay) = |a| \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{p}} = |a| d(x,y)$.

(b) **Solution.** $d(x + z, y + z) = \left(\sum_i |x_i + z_i - (y_i + z_i)|^p\right)^{\frac{1}{p}} = \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{p}} = d(x,y)$

c) **Solution.** Since $d(ax, ay) = \begin{cases} 1 & \text{if } ax = ay \\ 0 & \text{if } ax \neq ay \end{cases}$ the expressions can be further simplified by a and we get the original metric back. Therefore, $d(ax, ay) = d(x,y)$. Homogeneity would therefore only hold for $a = 1$.

d) **Solution.** The metric does not fulfill the property of *translation invariance*. Counterexample: pick two basis vectors $x = [0,1]^\top, y = [1,0]^\top$. Let the translating vector be $z = y$. Then, $d([1,1]^\top, [0,2]^\top) = \frac{2}{\pi} \arccos\left(\frac{1}{\sqrt{2}}\right) = \frac{2}{\pi} \cdot \frac{\pi}{4} = \frac{1}{2}$.

However, note that $d([0,1]^\top, [1,0]^\top) = \frac{2}{\pi}\arccos(0) = 1$, therefore $d(x+z, y+z) \neq d(x,y)$.