Homework 1: Distance functions on vectors

Exercise 1

Exercise 1.b

When considering the distance between documents in the same group and the average when comparing a document to one from an average group, no consistency can be seen in the results. Taking the particular case of the 'talk.politics.guns', most metrics report a higher distance between this and a document from the same group than the average for the rest of the documents, excepting the Manhattan and the Hamming distance. The same behaviour can be seen in other groups as well, such as 'talk.religion.misc' where only Euclidean (1.24) and Minkowski-4 (1.65) distances seem to report a lower distance within the group the average distance reported outside the group(1.28, 0.67 respectively).

Exercise 1.c

The Hamming distance seems to provide on average the best separation between groups. In 3 ('talk.politics.guns', 'comp.sys.mac.hardware', 'comp.graphics') out of the 5 groups tested, the Hamming distance is lower within the group than the average distance outside the group. The most evident is in the case of the 'comp.sys.mac.hardware' group which is expected since that type of document is expected to have more technical/specific language within (distance: 80) which helps in observing a higher distance compared to the other groups. It has a relatively low Hamming distance when compared to 'comp.graphics'(92), as some of the technical terms could be shared with that group. The Hamming distance is better on average as its binarity amplifies the differences between words present in some groups, such as technical terms, and absent from others.

Exercise 1.d

The similarity measure depicted in this exercise is defined by diving the dot products of two vectors by their Euclidean distance which equals the cosine of the angle between the two given vectors. Given that x and y are td-df vectors which have only non-negative values, s(x,y) can take values from 0 to 1. If the vectors would be arbitrary, the range of values would be the range of values that can be taken by the cosine function, meaning from -1 to 1. Even if dimensionality is increased, the similarity measure remains relevant. Despite possibly seeing a higher dot product in increased dimensionality, the division by the Euclidean distance normalizes the measure relatively to the dimensionality of the vector.

Exercise 1.e

As dimensionality increases, a different behaviour can be observed between L1 and L2 norms. The Euclidean distance, L2 norm, is calculated using many coordinates and in a high dimensional space, the data become sparse. This will then lead to very low values for the L2 norm which would make it less efficient in its use as a metric. As, L1 norm is the sum of the absolute differences of the coordinates, this holds true in higher dimensions as well. This behaviour can already be visible in our dataset, even if not to its full extent. The Euclidean distances are quite small (1.18-1.37), without obvious differences for within groups comparisons vs. outside groups comparisons, while the Manhattan distance performs better on average in distinguishing the two.

Exercise 2

Exercise 2.a

i) It is not a metric as it does not satisfy d(x_1, x_3)<= d(x_1, x_2)+ d(x_2, x_3). Given x_1 = -5, x_2 =1 and x_3 =2, the expression would be 49<=37, which is false.

ii) It is not a metric as it does not satisfy the non-negativity constraint. Given x_1 =-1, x_2 =1, y_1 = 1, y_2= -1, the expression would be -4.

iii) It is a metric.

iv) It does not satisfy the similarity condition. Given x_1 =0.2, x_2 =0.8, y_1 = 0.4, y_2= 0.6, the expression would be 0.2*log(0.2/0.4)+0.8*log(0.8/0.6)=0.4*log(0.4/0.2) + 0.6*log(0.6/0.4). Through calculus, the expression renders to be false.

v) It is a metric.

Exercise 2.b

i) $d(ax, ay) = \left(\sum_{i=1}^{n} |ax_i - ay_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |a * (x_i - y_i)|^p\right)^{\frac{1}{p}} =$

$$= \left(\sum_{i=1}^{n} |a|^p * |x_i - y_i\,|^p\right)^{\frac{1}{p}} =$$

$$= \left(|a|^p \ * \sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} =$$

. $$= |a| * \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} =$$

$= |a| * d(x, y)$ (true)

ii) $d(x + z, y + z) = \left(\sum_{i=1}^{n} |x_i \ + z_i \ - y_i \ - z_i\ |^p\right)^{\frac{1}{p}} = d(x, y)(true)$

Exercise 2.c

Given a=2, x=1, y=0, d(ax, ay) = 1, d(x,y)=1 and |a|* d(x,y)=2*1=2. This counterexample proves that for this function homogeneity does not apply as d(ax,ay) is not equal to |a|* d(x,y).

Exercise 2.d

Given the function

$$x, y \in \mathbb{R}^n_+ : \quad d(x,y) = \frac{2}{\pi} \operatorname{acos} \left(\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \right)$$

and knowing that cos (x,y)= dot product(x,y)/Euclidean distance (x,y), the acos term would be the value of the angle between the two vectors. The translation invariance would apply to the following function if and only if the two angles formed by the pairs of vectors (x,y) and (x+z, y+z) would be the same. One counterexample could be x_1 =0, x_2 =1, y_1 = 1, y_2= 1, z_1 =1, z_2 =2, for which d(x,y)= $\frac{1}{\sqrt{2}}$ and d(x+z,y+z)= $\frac{11}{5\sqrt{6}}$ which are not equal. Therefore, translation invariance does not apply to the following function.