Solution to Homework 1

Dr. Thomas Gumbsch
thomas.gumbsch@bsse.ethz.ch

Prof. Dr. Karsten Borgwardt
karsten.borgwardt@bsse.ethz.ch

Exercise 1

Exercise 1.a, programming component

The source file for `distance_fn.py` can be downloaded from the course website. All functions can be implemented in a single line. Longer solutions are of course also correct!

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?

For any two groups G_1 and G_2, intuition tells us that the average distance between pairs of documents of $G_1 \times G_1$ should be smaller than the average distance between pairs of documents of $G_1 \times G_2$. Yet, the results do not show this behaviour! We observe, for example, that in most cases the average distance to `comp.sys.mac.hardware` is smaller than the average distance to any other group. Figure 1 below shows the erratic behaviour of the metrics when computing the average distances between `talk.politics.guns` and the other groups.

Exercise 1.c

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

Without a clear definition of what constitutes a *good* separation between the groups, we can choose a metric whose values lead to larger variability when computed across groups. Both the Manhattan and the Hamming distance seem to satisfy this requirement and, to a lesser extent, the Euclidean distance. The Chebyshev and the two Minkowski metrics give the same average distance for comparisons of different groups. For example, the average distances of `talk.religion.misc:talk.religion.misc` and `talk.religion.misc:rec.autos` have the same value for the Minkowski metrics with $d = 3$ and $d = 4$.

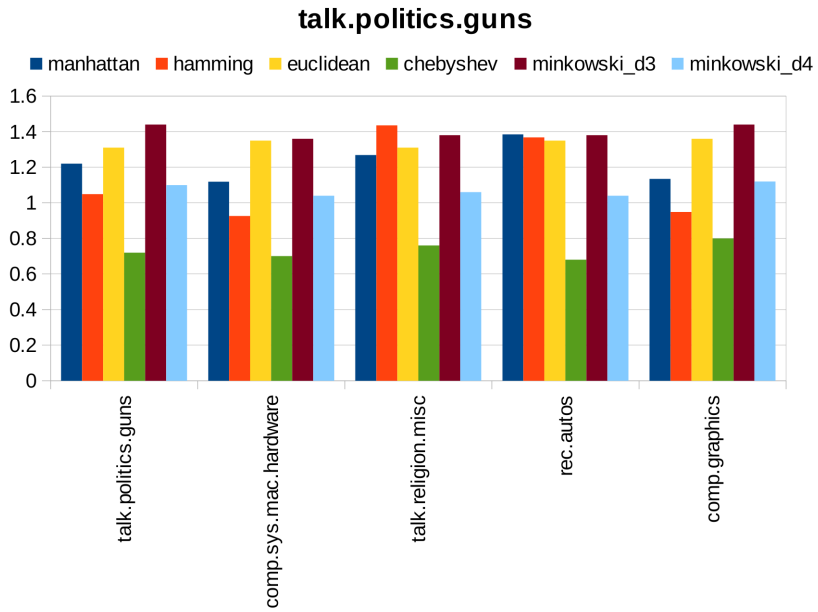

Figure 1: Average distances of group `talk.politics.guns`. The value d of the metrics were scaled for visualization purposes: a) Manhattan = $\frac{d}{10}$, b) Hamming = $\frac{d}{100}$, c) Euclidean = no change, d) Chebyshev = $d * 2$, e) Minkowski = $d * 2$.

Exercise 1.d

Another similarity measure is also often suggested in the literature, viz.

$$s(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|},$$

where $\mathbf{x} \cdot \mathbf{y}$ refers to the dot product of the two vectors, and $\| \cdot \|$ denotes the usual Euclidean distance. Given that $\mathbf{x}$ and $\mathbf{y}$ are tf-idf vectors, what can you say about the range, i.e. the possible values, that s can attain? How does this change when $\mathbf{x}$ and $\mathbf{y}$ are arbitrary vectors? What happens (both in case $\mathbf{x}$ and $\mathbf{y}$ are tf-idf vectors, and for arbitrary vectors) if the dimensionality is increased?

In general, the range of the expression above is $[-1, 1]$. This is a consequence of the range of the cosine function. However, in the case of td-idf vector, we obtain a range of $[0, 1]$, because the vectors are never negative (making angles over 90° impossible).

The metric is, to some extent, impervious to high dimensions: it will always have the same range; this is sometimes preferred to the 'strange' behaviour observed for other distance metrics in high dimensions.

Exercise 1.e

The Manhattan and Euclidean distances are also known as L_1 and L_2 norms, respectively. In general, what behaviour can we expect about the L_1 vs. L_2 norms as the dimensionality (i.e. the number of attributes) in the data increases? Is this behaviour observed in our dataset? If not, why not?

For high-dimensional data, a distance function such as the L_p norm—the Minkowski distance for generic p—degrades much more rapidly for larger values of p. Therefore,

when comparing the Manhattan (L_1) and Euclidean (L_2) distances, one will expect the Manhattan distance to perform better in high dimensional spaces. This behaviour is not clearly seen in our data due to not only to the sparsity of the data but also to the fact that the vectors have a dimension of $d = 1850$, which may not be considered to be very large; the effects are more apparent in extremely high dimensions, or in the limit of $d \to \infty$.

Exercise 2

Based on the conditions that must be met for a function to be a metric, answer the following questions:

Exercise 2.a

Which of the functions listed below are not metrics? Indicate what condition is not satisfied, if any. It is sufficient to provide a simple counterexample to prove that something is *not* a metric. By contrast, you do not have to prove that something *is* a metric, you merely have to state it.

i) $x, y \in \mathbb{R}^n, \quad d(x,y) = \sum_{i=1}^n (x_i - y_i)^2$

Not a metric, as it fails property 4 (subadditivity). A simple counterexample in $n = 1$ dimensions is given by setting $x = 1$, $y = 2$ and $z = 3$, which satisfy $4 = d(x,z) > d(x,y) + d(y,z) = 2$.

ii) $x, y \in \mathbb{R}^n, \quad d(x,y) = \sum_{i=1}^n x_i y_i (x_i - y_i)^2$

Not a metric, as it fails property 2 (identity of indiscernibles). A simple counterexample in $n = 1$ dimensions is given by setting $x = 0$ and $y \neq 0$.

iii) $x, y \in \mathbb{R}^n, \quad d(x,y) = \sum_{i=1}^n w_i |x_i - y_i|, \; w_i > 0 \; \forall i$

This is a metric. Since $w_i|x_i - y_i| = |w_i x_i - w_i y_i|$ for $w_i > 0$, it follows that $d(x,y)$ is equivalent to the Manhattan distance between vectors $\tilde{x} = w \odot x$ and $\tilde{y} = w \odot y$, where $w = (w_1, w_2, \cdots, w_n)$ and $\odot$ denotes the Hadamard (i.e. elementwise) product. This also illustrates a more generic principles of metrics: it is possible to scale them (with positive values) and obtain a new metric. Try to prove this as an additional exercise.

iv) $x, y \in \{z \in \mathbb{R}^n \mid \sum_{i=1}^n z_i = 1, z_i > 0 \; \forall i\}, \quad d(x,y) = \sum_{i=1}^n x_i \log\left(\frac{x_i}{y_i}\right)$

Not a metric, as it fails property 3 (symmetry).

v) $x, y \in \mathbb{R}^n, \quad d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$

It is a metric; in some sense, it is the simplest metric, leading to a very 'coarse' distance.

Exercise 2.b

For the Minkowski distance, show that:

i) $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$, $\quad d(ax, ay) = |a| d(x, y)$

$$\begin{aligned} d(ax, ay) &= \left(\sum_{i=1}^{n} |ax_i - ay_i|^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |a|^p |x_i - y_i|^p \right)^{\frac{1}{p}} = \left(|a|^p \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \\ &= (|a|^p)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = |a| \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = |a| d(x, y). \end{aligned}$$

ii) $x, y, z \in \mathbb{R}^n$, $\quad d(x + z, y + z) = d(x, y)$

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

Exercise 2.c

Property i) in Exercise 2.b is called *homogeneity*. Determine if homogeneity applies to function v) in Exercise 2.a

Homogeneity does not apply to function v) in Exercise 2.a. As a counterexample, assume $x \neq y$ and $a = 2$ then $ax \neq ay$. Therefore, $d(x, y) = d(ax, ay) = 1 \neq 2 = |a| d(x, y)$

Exercise 2.d

Property ii) in Exercise 2.b is called *translation invariance*. Determine if translation invariance applies to the following function:

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

Translation invariance does not apply to this function. We can rewrite the equation to show that $d(x, y) = \frac{2}{\pi} \theta(x, y)$ where $\theta(x, y) \in \left[0, \frac{\pi}{2}\right]$ is the angle between vectors x and y in the positive orthant of Euclidean space. Since the angle $\theta(x, y)$ of two points with respect to a fixed point in the coordinate system is *not* translation invariant, the resulting distance function $d(x, y)$ is not translation invariant either. For a specific counterexample, consider $x = (1, 1)$, $y = (1, 2)$ and $z = (1, 1)$.