Homework 1 - Data Mining I

John Oehninger
joehninger@student.ethz.ch

1 Exercise 1

1.1 Exercise 1.b

There is an obvious abnormality when looking at the comparisons of documents of the same group vs. documents from different groups. The distances of the documents compared in the same group are not always the smallest distances compared to the documents compared from different groups.

comp.graphics:comp.graphics and comp.sys.mac.hardware:comp.sys.mac.hardware have the lowest distance of all Manhattan distances. In the Hamming distances, only comp.sys.mac.hardware:comp.sys.mac.hardware has the smallest distance.

For the Minkowski distances, comparing the results of d=3 and d=4, the spread is equal to 0.10 in both cases, which is interesting.

1.2 Exercise 1.c

It seems that the Hamming distance provides the best separation between groups. I have concluded this from my output data, as these values are the most "spread out" between groups.

1.3 Exercise 1.d

$$s(x, y) := \frac{x \cdot y}{||x|| \cdot ||y||} \tag{1}$$

Formula 1 actually equals cos(θ), as it is the same formula for calculating the angle between two vectors. Given that x and y are tf-idf vectors the range lies in [0,1]. If on the other hand these are arbitrary vectors, the range lies in [-1,1].

As dimensionality increases, s is bound to get smaller or converge towards zero. If we increase the dimensions in the vectors but hold each value at 1, s will be at its maximum value, which is zero. But once we do this with vectors with arbitrary values, s is bound to get smaller. This is due to the fact that the denominator will become larger than the enumerator in equation 1.

1.4 Exercise 1.e

The Manhattan distance, as shown in 2D space in figure 1, adds up the individual differences in length of each dimension. So it does not compute the direct path, as the crow flies, but takes a path where each step is taken parallel to the respective dimensional axis.

We can understand the Euclidean distance as the exact length of a line connecting two points. It is based on the Pythagorean theorem, which I have shown in figure 2.

As dimensionality increases the Euclidean distance should increase less compared to the Manhattan distance.

Looking at my output data I can see larger differences in the Manhattan distances compared to the Euclidean distances.

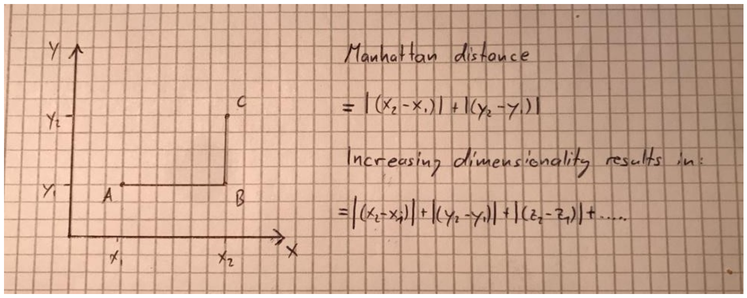

Figure 1: Manhattan distance (L1 norm).

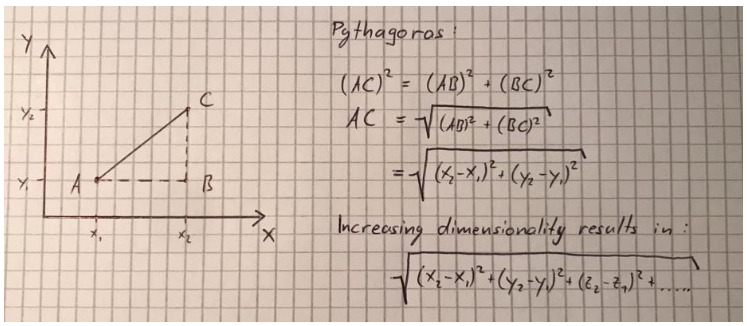

Figure 2: Euclidean distance (L2 norm).

2 Exercise 2

2.1 Exercise 2.a

The equations below represent the four conditions that must be fulfilled for a function to be a metric.

$$1) d(x_1, x_2) \geq 0$$
$$2) d(x_1, x_2) = 0 \text{ iff } x_1 = x_2$$
$$3) d(x_1, x_2) = d(x_2, x_1)$$
$$4) d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$$

The equations in figure 3 represent the functions to be determined if they are a metric or not.

$$\begin{array}{lll}
\text{i)} & x, y \in \mathbb{R}^n, & d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \\
\text{ii)} & x, y \in \mathbb{R}^n, & d(x,y) = \sum_{i=1}^{n} x_i y_i (x_i - y_i)^2 \\
\text{iii)} & x, y \in \mathbb{R}^n, & d(x,y) = \sum_{i=1}^{n} w_i |x_i - y_i|,\ w_i > 0\ \forall i \\
\text{iv)} & x, y \in \{z \in \mathbb{R}^n \mid \sum_{i=1}^{n} z_i = 1, z_i > 0\ \forall i\}, & d(x,y) = \sum_{i=1}^{n} x_i \log\left(\frac{x_i}{y_i}\right) \\
\text{v)} & x, y \in \mathbb{R}^n, & d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}
\end{array}$$

Figure 3: Functions to be determined if they are a metric or not.

2.1.1 i)

Condition	1)	2)	3)	4)
	✓	✓	✓	✓

All conditions are fulfilled, making it a metric.

2.1.2 ii)

Condition	1)	2)	3)	4)
	✓	✗	✓	✓

All conditions are not fulfilled, which doen't make this a metric. Condition 2 is not fulfilled because the entire term can be zero if x or y are Null vectors.

2.1.3 iii)

Condition	1)	2)	3)	4)
	✓	✓	✓	✓

All conditions are fufilled, making it a metric.

2.1.4 iv)

Condition	1)	2)	3)	4)
	✗	✓	✗	✓

All conditions are not fulfilled, which doesn't make this a metric. Condition 1 is not fulfilled because if the fraction in the log lies between 0 and 1, the term immediately becomes negative. Condition 3 is not fulfilled because if we swap x and y in the log the value will not be the same.

2.1.5 v)

Condition	1)	2)	3)	4)
	✓	✓	✓	✓

All conditions are fulfilled, making it a metric.

2.2 Exercise 2.b

Equation 2 is the formula for the Minkowski distance.

$$d(x,y) = (\sum |x-y|^p)^{\frac{1}{p}} \tag{2}$$

2.2.1 i)

Show that $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n, d(ax, ay) = |a|d(x,y)$ for the Minkowski distance. This property is called *homogeneity*.

$$\begin{gathered} \text{d(ax,ay) = |a| d(x,y)} \\ (\sum |ax-ay|^p)^{\frac{1}{p}} \\ (\sum |a(x-y)|^p)^{\frac{1}{p}} \\ (\sum |a|^p|(x-y)|^p)^{\frac{1}{p}} \\ (|a|^p \sum |(x-y)|^p)^{\frac{1}{p}} \\ |a|(\sum |(x-y)|^p)^{\frac{1}{p}} = |a|d(x,y) \end{gathered} \tag{3}$$

We see that the condition of *homogeneity* is fulfilled.

2.2.2 ii)

Show that $x, y, z \in \mathbb{R}^n, d(x+z, y+z) = d(x,y)$ for the Minkowski distance. This property is called *translation invariance*.

$$\begin{gathered} d(x+z, y+z) \\ (\sum |(x+z)-(y+z)|^p)^{\frac{1}{p}} \\ (\sum |(x+z-y-z)|^p)^{\frac{1}{p}} \\ (\sum |(x-y)|^p)^{\frac{1}{p}} = d(x,y) \end{gathered} \tag{4}$$

We see that the condition of *translation invariance* is fulfilled.

2.3 Exercise 2.c

Determine if *homogeneity* applies to function 5

$$x, y \in \mathbb{R}^n, \quad d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \tag{5}$$

If I think through the calculation it appears that function 5 does not fulfill the homogeneity condition.

2.4 Exercise 2.d

Determine if *translation invariance* applies to the function 6.

$$x, y, z \in \mathbb{R}^n_+ : d(x,y) = \frac{2}{\pi} \arccos \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2}} \tag{6}$$

Function 6 is in fact the formula for the angle between two vectors, see equation 7.

$$\theta = \arccos \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{7}$$

Figure 4 shows an example that does not fulfill the condition of *translation invariance*. The angle is not conserved.

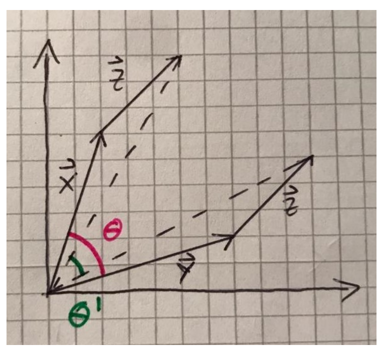

Figure 4: Functions to be determined if they are a metric or not.