Distance functions on vectors

Andreas Dahlberg

October 7, 2018

Exercise 1

Exercise 1a)

See the file *distance_fn.py*.

Exercise 1b) and 1c)

By looking at *output_distances.txt* we want to determine which metric that provides the best result. By the best result we mean that the distance between the same type of groups should be small and the distance between different types of groups large.

In *output_distances.txt* we see that values in the same column can differ significantly. For instance, the first column (Manhattan) ranges between 10 and 180 while the fourth column (Chebyshev) only ranges between 0.98 and 1.00. In order to compare the different metrics we want to normalize our data, i.e so that all our data fall in to the interval $[0, 1]$. The python file *normalize.py* reads the output file and creates a new file called *normalized_output.txt* where the distances are in the interval $[0, 1]$. The python file reads each column and finds the maximum and minimum and then replaces the value x with

$$\frac{x - min}{max - min}.$$

Note that in order to run the file *normalize.py* we need to remove the first column in *output_distances.txt* consisting of the strings of group names. This is because it is much easier in python to read a data file consisting only of numbers. The result we get is represented in the table below.

comp.graphics:comp.graphics	0.00	0.06	0.10	1.00	0.13	0.16
comp.graphics:comp.sys.mac.hardware	0.25	0.02	0.03	0.00	0.03	0.03
comp.graphics:rec.autos	0.27	0.48	0.58	1.00	0.61	0.63
comp.graphics:talk.politics.guns	0.26	0.15	0.19	1.00	0.21	0.24
comp.graphics:talk.religion.misc	0.26	0.54	0.56	0.50	0.56	0.55
comp.sys.mac.hardware:comp.sys.mac.hardware	0.41	0.00	0.00	0.00	0.00	0.00
comp.sys.mac.hardware:rec.autos	0.68	0.45	0.54	1.00	0.57	0.60
comp.sys.mac.hardware:talk.politics.guns	0.48	0.12	0.15	1.00	0.17	0.19
comp.sys.mac.hardware:talk.religion.misc	0.72	0.53	0.53	0.00	0.51	0.50
rec.autos:rec.autos	0.91	0.85	0.97	1.00	1.00	1.00
rec.autos:talk.politics.guns	0.75	0.57	0.67	1.00	0.70	0.71
rec.autos:talk.religion.misc	0.96	0.94	0.99	1.00	1.00	0.98
talk.politics.guns:talk.politics.guns	0.56	0.25	0.31	1.00	0.34	0.35
talk.politics.guns:talk.religion.misc	0.79	0.63	0.66	1.00	0.66	0.66
talk.religion.misc:talk.religion.misc	1.00	1.00	1.00	1.00	0.99	0.98

The first column shows the two groups we are comparing, the second the Manhattan distance, the third the Hamming distance, the fourth the Euclidean distance, the fifth the Chebyshev distance, the sixth the Minkowski distance with $p = 3$ and the fourth the Minkowski distance with $p = 4$.

We see that all distance functions except Chebyshev gives a low value when comparing comp.graphics with itself. However, when compaing rec.autos with itself and talk.religion.misc with itself all distance functions gives a high value which isn't a good result. This could be because all documents in talk.religion.misc discuss different topics in religion and hence use completely different words. Another plausible reason is simply that 10 documents in each group is not enough data to draw any conclusions. We do, however, have some good results. For instance, all distance functions give a high value when comparing rec.autos with talk.religion.misc, which they should since these are two completely different topics. Finally, we can see that when compaing comp.graphics with comp.sys.mac.hardware almost all distance functions give a low result. This could be because these two different topics are closely related: they both concern computers.

We can decide which is "the best" metric by for each metric summing up the distances between different groups and then subtracting the distances between the same groups. The higher the value, the better the metric is. This is because if the value is high, then we have a large distance between different groups and small values between the same type of groups. By doing this we get:

Manhattan: 0.25+0.27+0.26+0.26+0.68+0.48+0.72+0.75+0.96+0.79-(0+0.41+0.91+0.56+1) = 2.54
Hamming: 0.02+0.48+0.15+0.54+0.45+0.12+0.53+0.57+0.94+0.63-(0.06+0+0.85+0.25+1) = 2.27
Euclidean: 0.03+0.58+0.19+0.56+0.54+0.15+0.53+0.67+0.99+0.66-(0.1+0+0.97+0.31+1) = 2.52
Chebyshev: 0+1+1+0.5+1+1+0+1+1+1-(1+1+1+1+1) = 2.5
Minkowski-3: 0.03+0.61+0.21+0.56+0.57+0.17+0.51+0.7+1+0.66-(0.13+0+1+0.34+0.99) = 2.56
Minkowski-4: 0.03+0.63+0.24+0.55+0.6+0.19+0.50+0.71+0.98+0.55-(0.16+0+1+0.35+0.98) = 2.49

From this we can conclude that Minkowski-3 is the best metric while Manhattan is the worst metric. Of course, which one is the best metric can and will probably change if we get more data (documents).

Exercise 1d) (and some more stuff)

From *output_distances.txt* we can conclude that the distance functions can be ordered as (skipping the Hamming distance): Manhattan, Euclidean, Minkowski-3, Minkowski-4 and then Chebyshev where Manhattan usually gives the highest values for distances while Chebyshev usually gives the smallest values. We note that Manhattan is Minkowski with $p = 1$, Euclidean is Minkowski for $p = 2$ and Chebyshev is Minkowski for $p = \infty$. It seems as if the distance functions get smaller values for larger p. Let's try to motivate this with an example: Conside the point $x = (0, 0, ..., 0)$ and $y = (k, k, ..., k)$ in n dimensions. Using the Minkowski-p distance we get

$$d(x, y) = (k^p + \ldots + k^p)^{1/p} = kn^{1/p}$$

which decreases with increasing p.

Now if $p = 1$ (Manhattan) then the above expression is kn and if $p = 2$ (Euclidean) we get $k \cdot n^{1/2}$. Both these expression will go to infinity as n increases and the Euclidean will go to infinity slower than the Manhattan. This means that as the number of words increases, the dimensionality of the vector space that each document belongs to increases and the distance (L1 and L2) increases (this is of course not true if the number of words in the two documents stay the same and the "new words" only fall in to the other documents).

We can compare the L1 and L2 norm by analysing this example: Suppose we have our 5 groups (A,B,C,D and E) containing 10 documents each. Now we get a new document and want to figure out which of the 5 groups this document belongs to. Suppose our document belongs to group A, which we assume is comp.graphics. We decide which group the document belongs to by calculating the average distance to each group and then taking the group that the document is closest to. The word "Algorithm" is probably an important word for this new document but not for the documents in all the other groups except for comp.graphics. So, for instance, when calculating the distance between this new document and the documents in talk.religion.misc the Euclidean distance will "punish" this group much more than the Manhattan distance since the Euclidean distance will square the error and hence the word "Algorithm" will have a larger impact on the distance. When the number of words in each document increases (or the number of documents in each group), it is likely that the word "Algorithm" will increase in importance for this new document and will therefore punish all the other groups but comp.graphics even more. This example motivates that for large dimensions it is better to choose the Euclidean distance when deciding which group a new document belongs to.

We can find this kind of behavior in our dataset. When comparing comp.graphics with all the other groups we see that for the Euclidean distance it produces low values when comparing it with itself and comp.sys.mac.hardware and high values when comparing it to the other groups. This is because comp.graphics probably have completely different important words than groups like rec.autos and when squaring the error this difference has a larger impact. However, the difference between comp.graphics and comp.sys.mac.hardware is not so big for the Euclidean distance and this could be because they have almost the same important words.

Exercise 2

Exercise 2a)

In this exercise we determine which of the below functions that are metrics on the given set.

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

   We see immediately that $d(x,y) \geq 0$ and 0 if and only if $x = y$. We also have that $d(x,y) = d(y,x)$ so it remains to check whether or not the triangle inequality holds. We prove that the triangle inequality does not hold by a counter example. Let $n = 2$, then we have that

   $$d((0,0),(2,2)) = 8$$

   and

   $$d((0,0),(1,1)) = d((1,1),(2,2)) = 2$$

   So we do not have that

   $$d((0,0),(2,2)) \leq d((0,0),(1,1)) + d((1,1),(2,2))$$

   and hence this is not a metric.

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

   Take $p = (1,0)$ and $q = (-1,1)$, then we have that $d(p,q) = -4$ and since a metric always has to be positive, this is not a metric.

3.

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

We see that $d(x,y) \geq 0$ and 0 if and only if $x = y$, and also that $d(x,y) = d(y,x)$. It remains to prove that the triangle inequality holds. So take $x, y, z \in \mathbb{R}^n$. Then by the triangle inequality for real numbers we have that

$$|x_i - y_i| = |x_i - z_i + z_i - y_i| \leq |x_i - z_i| + |z_i - y_i|.$$

Hence, for $w_i > 0$ we have

$$w_i|x_i - y_i| \leq w_i|x_i - z_i| + w_i|z_i - y_i|$$

and summing over i we get that

$$\sum_{i=1}^{n} w_i|x_i - y_i| \leq \sum_{i=1}^{n} w_i|x_i - z_i| + \sum_{i=1}^{n} w_i|z_i - y_i|$$

which means that $d(x,y) \leq d(x,z) + d(z,y)$ so this is a metric.

4.

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

By looking at it, it would be very surprising if d was symmetric. We will show that this is not the case with a counter example for $n = 2$. So take $p = (1/3, 2/3)$ and $q = (1/2, 1/2)$. We get that

$$d(p,q) = \frac{1}{3}\log(\frac{2}{3}) + \frac{2}{3}\log(\frac{4}{3}) = \frac{1}{3}\log(2) + \frac{4}{3}\log(2) - \frac{1}{3}\log(3) - \frac{2}{3}\log(3) = \frac{5}{3}\log(2) - \log(3)$$

and

$$d(q,p) = \frac{1}{2}\log(\frac{3}{2}) + \frac{1}{2}\log(\frac{3}{4}) = \frac{1}{2}\log(3) + \frac{1}{2}\log(3) - \frac{1}{2}\log(2) - \log(2) = \log(3) - \frac{3}{2}\log(2)$$

and putting this in to the calculator we see that $d(p,q) \neq d(q,p)$ neither in base e nor in base 10, so this is not a metric.

5.

$$x, y \in \{z \in \mathbb{R}^n : \sum_{i=1}^{n} z_i = 1, \ z_i > 0\} \quad d(x,y) = \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$$

We see that $d(x,y) = d(y,x)$ and $d(x,y) \geq 0$ and 0 if and only if $x = y$. However, we will show that the triangle inequality does not hold with a counter example for $n = 2$. So take $p = (1/2, 1/2)$, $q = (1/3, 2/3)$ and $r = (1/4, 3/4)$. We get that

$$d(p,r) = \frac{1}{2}\left(\frac{(1/2 - 1/4)^2}{1/2 + 1/4} + \frac{(1/2 - 3/4)^2}{1/2 + 3/4}\right) = 1/2\left(\frac{1/16}{3/4} + \frac{1/16}{5/4}\right) = \frac{1}{2}\left(\frac{1}{12} + \frac{1}{20}\right)$$

$$d(p,q) = \frac{1}{2}\left(\frac{(1/2 - 1/3)^2}{1/2 + 1/3} + \frac{(1/2 - 2/3)^2}{1/2 + 2/3}\right) = \frac{1}{2}\left(\frac{1/36}{5/6} + \frac{1/36}{7/6}\right) = \frac{1}{2}\left(\frac{1}{30} + \frac{1}{42}\right)$$

$$d(q,r) = \frac{1}{2}\left(\frac{(1/3 - 1/4)^2}{1/3 + 1/4} + \frac{(2/3 - 3/4)^2}{2/3 + 3/4}\right) = \frac{1}{2}\left(\frac{1/144}{7/12} + \frac{1/144}{17/12}\right) = \frac{1}{2}\left(\frac{1}{7 \cdot 12} + \frac{1}{17 \cdot 12}\right)$$

By plugging this into the calculator we get that $d(p,r) \approx 0.066$ and $d(p,q) + d(q,r) \approx 0.037$ so we do not have that $d(p,r) \leq d(p,q) + d(q,r)$ and therefore this is not a metric.

6.
$$x, y \in \mathbb{R}^n_+ \quad d(x,y) = \frac{2}{\pi} \arccos\left(\frac{\sum_{i=1}^n x_i y_i}{(\sum_{i=1}^n x_i^2)^{1/2} (\sum_{i=1}^n y_i^2)^{1/2}} \right).$$
We have $\arccos : [-1,1] \to [0,\pi]$ and $\arccos(x) = 0$ if and only if $x = 1$. This means that $d(x,y) = 0$ if and only if
$$\sum_{i=1}^n x_i y_i = \sum_{i=1}^n (x_i^2)^{1/2} (\sum_{i=1}^n y_i^2)^{1/2}.$$
Suppose $x = (1,1,1,\ldots,1)$ and $y = 2x = (2,2,2,\ldots,2)$. Then $\sum_{i=1}^n x_i y_i = 2n$ and
$$\sum_{i=1}^n (x_i^2)^{1/2} (\sum_{i=1}^n y_i^2)^{1/2} = \sqrt{n}(2^2 n)^{1/2} = 2n$$
so that $d(x,y) = 0$ even though $x \neq y$. This means that $d(x,y)$ is not a metric.

7.
$$x, y \in \mathbb{R}^n \quad d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$$
Clearly $d(x,y) \geq 0,\ d(x,y) = d(y,x)$ and $d(x,y) = 0$ if and only if $x = y$. It remains to prove that the triangle inequality holds. So take some $z \in \mathbb{R}^n$. If $d(x,z) = 0$ then obviously $d(x,z) \leq d(x,y) + d(y,z)$. If $d(x,z) = 1$ then $x \neq z$ and $d(x,y) + d(y,z)$ is 1 or 2 because if it is 0 then $x = y$ and $y = z$ which implies $x = z$ and leads to a contradiction. So in both cases we have $d(x,z) \leq d(x,y) + d(y,z)$ which means that this is a metric.

Exercise 2b)

The Minkowski distance is defined as
$$d(x,y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}, \quad p > 0.$$

1. For $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$ we have
$$d(ax, ay) = (\sum_{i=1}^n |ax_i - ay_i|^p)^{1/p} = (\sum_{i=1}^n |a|^p |x_i - y_i|^p)^{1/p} = |a| (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} = |a| d(x,y)$$
Which means that the Minkowski distance has *homogeneity*.

2. For $x, y, z \in \mathbb{R}^n$ we have
$$d(x+z, y+z) = (\sum_{i=1}^n |(x_i + z_i) - (y_i + z_i)|^p)^{1/p} = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} = d(x,y)$$
Which means that the Minkowski distance is *translation invariant*.

Exercise 2c)

We want to determine whether or not function 7 in exercise 2a) has homogeneity. There are two cases: $a = 0$ and $a \neq 0$. If $a = 0$ then $d(0,0) = 0$ and $0 \cdot d(x,y) = 0$. If $a \neq 0$ then
$$d(ax, ay) = \begin{cases} 0, & ax = ay \\ 1, & ax \neq ay \end{cases}$$
but since $a \neq 0$ then $ax = ay$ is equivalent to $x = y$ so $d(ax, ay) = d(x,y)$. This proves that this function has homogeneity.

Exercise 2d)

We want to determine whether or not function 6 in exercise 2a) is translation invariant. We saw earlier that $d(x, y) = 0$ if and only if

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} (x_i^2)^{1/2} (\sum_{i=1}^{n} y_i^2)^{1/2}.$$

This is equality in Cauchy-Schwarz which means that this equality holds if and only if x and y are linear dependent. This means that $d(x, y) = 0$ if and only if x and y are linear dependent. We take $y = 2x$, then $d(x, 2x) = 0$ but it is not certain that $d(x, +z, 2x + z)$ is 0 since it is not sure that $x + z$ and $2x + z$ are linear dependent. For instance, we can take $x = (1, 1)$ and $z = (1, 0)$. Then $x + z = (2, 1)$ and $2x + z = (3, 2)$ which of course are not linear dependent. So in this case we have $d(x, 2x) = 0$ and $d(x + z, 2x + z) \neq 0$ which means that this function is not translation invariant.