Data Mining Homework 1

Probst Jennifer

14. October 2020

Question 1

This is the table output for exercise 1a.

Pair of new groups	Manhattan	Hamming	Euclidean	Chebyshev	Minkowski d=3	Minkowski d=4
comp.graphics:comp.graphics	10.52	86.36	1.30	0.44	0.72	0.57
comp.graphics:comp.sys.mac.hardware	10.24	82.38	1.23	0.39	0.67	0.52
comp.graphics:rec.autos	13.04	128.16	1.32	0.39	0.69	0.53
comp.graphics:talk.politics.guns	11.34	94.88	1.33	0.40	0.72	0.56
comp.graphics:talk.religion.misc	11.88	134.42	1.27	0.42	0.69	0.54
comp.sys.mac.hardware:comp.sys.mac.hardware	10.09	80.38	1.18	0.33	0.62	0.47
comp.sys.mac.hardware:rec.autos	12.79	125.02	1.27	0.33	0.64	0.48
comp.sys.mac.hardware:talk.politics.guns	11.18	92.60	1.28	0.35	0.68	0.52
comp.sys.mac.hardware:talk.religion.misc	11.71	132.66	1.22	0.37	0.64	0.49
rec.autos:rec.autos	15.28	165.33	1.33	0.33	0.65	0.48
rec.autos:talk.politics.guns	13.85	136.74	1.35	0.34	0.69	0.52
rec.autos:talk.religion.misc	14.32	173.86	1.30	0.37	0.65	0.50
talk.politics.guns:talk.politics.guns	12.20	104.91	1.37	0.36	0.72	0.55
talk.politics.guns:talk.religion.misc	12.68	143.52	1.31	0.38	0.69	0.53
talk.religion.misc:talk.religion.misc	12.97	179.84	1.24	0.39	0.65	0.50

Question 1b

We generally notice that the distance measures give values in different ranges, e.g. the Hamming distance measure has a far bigger variance than the rest. This is the case because the data is made binary, and the frequency values in the data which are all quite small get all set to one. High values in the table indicate high distance between the documents sampled from the two groups; meaning they have different words that are important with respect to the entire dataset.
In general one can't really say if comparing documents of the same group vs. documents from different groups results in lower values. The individual metrics don't really agree with each other and talk.religion.misc:talk.religion.misc for example displays in the Hamming distance even the highest value.
I would have expected some comparisons to result in quite high values as for example talk.religion.misc:rec.autos as they cover really diverse topics with maybe uncommon words or even in different languages. Comp.graphics:comp.graphics

might cover more similar words as everyone programs in the same programming languages. These hypotheses are true in some metrics as Manhattan or Hamming but not for other, making proper statements hard.

Question 1c

Chebyshev, Minkowski and Euclidean distance measures don't really have a big variance in output values, making the separation between groups hard. Manhattan and Hamming distance measures disagree only in some cases. If I have to pick one, I'd say the Hamming distances separates the groups most clearly, as the binary conversion of the vectors makes increases distances.

Question 1d

We know that:

$$s(x, y) = \frac{x * y}{|x| * |y|} = cos(\theta)$$

and:

$$cos(\theta) \in [-1, 1]$$

where θ represents the angle between vectors x and y.

In the case of tf-idf vectors, the distance between two vectors can never be negative, as there are no negative values in tf-idf vectors. It should also not be 0 if $x, y \neq 0$. For the scalarproduct to be 0, x and y would have to be orthogonal. As tf-idf vectors don't contain negative values, at each position of the vector either x or y would have to be 0, meaning they have absolutely no word in common. I is likely that every text should at least contain common words like 'and', 'or' or 'no'. But with arbitrary vectors the distance can be 0 even if $x \neq y$, in the case of x and y being orthogonal. So distances between tf-idf vectors can take values $(0, 1]$ and $[-1, 1]$ for arbitrary vectors.

There is no change if dimensionality is increased, as always $cos(\theta) \in [-1, 1]$, which is independent of the dimensionality of vectors.

Question 1e

Generally one can observe that the distances between individual vectors decreases as the dimensionality in the data increases, which is known as the 'curse of dimentionality problem'. This is even more a problem for the L2 norm (already lower var in low-dimensional space) than the L1 norm. We would need more samples to equally well detect structure higher dimensional space. In case

of tf-idf vectors we have frequencies represented, so the individual values are normalized. So generally I'd say with longer vectors the output values should not really change, otherwise the comparison between different document distances (e.g. in the table above) would not make any sense.

Question 2a

i) This is not a metric as the triangle inequality does not hold. Take $x = (0, -4)$, $y = (1, 0)$ and $z = (1, 1)$. Then $d(x, y) + d(y, z) = 1 + 4 * 4 + 1 = 18 < 26 = 1 + 5 * 5 = d(x, z)$, which contradicts the triangle inequality.

ii) This is not a metric as the second proposition on slide 37 does not hold. Take $x = (0, 0)$ so the distance to any y would be 0. $d(x, y)$ is therefore not 0 only if x=0 and y=0.

iii) This is the Manhattan Distance multiplied with a scalarvector, which does not have any impact on it being a metric.

iv) This is not a metric as it is not symmetric. Take $x = (0.25, 0.5, 0.25)$ and $y = (0.5, 0.1, 0.4)$, then $d(x, y) = 0.51393 \neq 0.37363 = d(y, x)$.

v) This is the discrete metric.

Question 2b

These proofs use the Minkowski metric:
i) For $a \in R, x, y \in R^n$ we show that $d(ax, ay) = |a| * d(x, y)$:

$$d(ax, ay) = \sum_{i=1}^{d}((|ax-ay|)^p)^{1/p} = \sum_{i=1}^{d}((|a(x-y)|)^p)^{1/p} = \sum_{i=1}^{d}|a|*|(((x-y)|)^p)^{1/p} = |a|*d(x, y)$$

ii) For $x, y, z \in R^n$ we show that $d(x + z, y + z) = d(x, y)$:

$$d(x + z, y + z) = \sum_{i=1}^{d}((|x + z - (y + z)|)^p)^{1/p} = \sum_{i=1}^{d}((|x - y|)^p)^{1/p} = d(x, y)$$

Question 2c

The property does not hold. Let a be any value $\in R \setminus \{1\}$ and $x \neq y$. Now $d(ax, ay) = 1 \neq |a| * 1 = |a| * d(x, y)$.

Question 2d

As defined in Exercise 1d:

$$s(x, y) = \frac{x * y}{|x| * |y|} = cos(\theta)$$

.
We furthermore know that $acos(cos(\theta))$ simplifies to θ, therefore:

$$d(x, y) = \frac{2}{\pi} * acos(cos(\theta)) = \frac{2}{\pi} * \theta$$

This θ is dependent on the choice of x and y and if we translate those by some value z, it will change. We find that angles are not translation invariant. Proof: take x=(0,1) and y=(1,0), where $\theta = 0$, because x and y are orthogonal to each other - so $d(x, y) = 0$. If we now take z=(1,0) as translation vector, we result in x'=(2,0) and y'=(1,1) which are not orthogonal to each other anymore, so $\theta \neq 0$ and $d(x + z, y + z) = d(x', y') \neq 0 = d(x, y)$.