Homework 1: Distance functions on vectors

Theresa Wakonig
Data Mining I

October 13, 2020

1 Exercise 1

Analyze dataset '20 Newsgroups Data'.

1.a

Generated results for average distance between documents of specified newsgroups:

Pair of newsgroups	Manhattan	Hamming	Euclid.	Chebysh.	Mink. d=3	Mink. 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

1.b

From the generated output one can not observe a general trend, that comparing documents of the same group with each other yields the smallest average distances. Instead, all newsgroups reportedly have the lowest average distance with the *comp.sys.mac.hardware* group (regardless of the metric used). As my understanding of the approach was that documents are more similar (smaller average distance) the more words they have in common, this result does not fulfill my expectations. As the tf-idf method can not truly capture the meaning or topic of the document and only compares words, the outcome might be a result of that scenario happening in overlapping vocab of certain newsgroups. Moreover, the weight/importance of a word increases the more rare it is - this might have also played a role in this analysis. The metrics used are probably not all suitable for handling sparse data as encountered in this example.

1.c

For me a clear separation of groups is not visible. However, the Manhattan and Hamming distances display the biggest and most obvious differences when looking at the different comparisons. As there are always a number of false associations in the output I find it hard to distinguish between groups or choose a specific metric.

1.d

$\vec{x} \cdot \vec{y} = \|x\|\|y\| \cdot cos(\vec{x}, \vec{y}) \quad \rightarrow \quad s(\vec{x}, \vec{y}) = cos(\vec{x}, \vec{y}) \quad , s \in [-1, 1]$

When **x** and **y** are tf-idf vectors all entries are positive by definition which means that the angle between the vectors will be no greater than 90 degrees: $s \in [0, 1]$. Arbitrary vectors may enclose any angle: $s \in [-1, 1]$. Increasing the dimensionality will keep s in the same range for both scenarios, as this metric only cares about the angle between the vectors and not the magnitude. Moreover I believe that it is likely to receive rather sparse tf-idf vectors with many zero entries.

1.e

As we are summing up positive values the distances should generally be getting larger for higher dimensions. Both norms give the same importance to all the dimensions and in high dimensions the metrics seem to report uniform distances despite small degrees of similarity. Especially the results of the L2 norm depict this behaviour as all values lie in a very small range (1.18 - 1.37). It seems to me, as if the "correct" choice of metric is essential for analyzing given data in a meaningful way.

Exercise 2)

a)

i) metric

ii) no metric:

- $d(\vec{x}, \vec{y}) \geq 0$ violated

↪ $\vec{x}$ & $\vec{y}$ may have negative entries.

↪ $\sum_{i=1}^{n} \underbrace{x_i y_i}_{\in \mathbb{R}} \underbrace{(x_i - y_i)^2}_{\geq 0}$ → sum is not guaranteed to be positive.

iii) metric

iv) no metric:

- $d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$ violated (non-symmetric)

↪ $\log\left(\frac{x_i}{y_i}\right) \neq \log\left(\frac{y_i}{x_i}\right)$

v) metric

b)

i) $a \in \mathbb{R}$ and $\vec{x}, \vec{y} \in \mathbb{R}^n$

$d(a\vec{x}, a\vec{y}) \overset{!}{=} |a| \, d(\vec{x}, \vec{y})$

Proof: $d(a\vec{x}, a\vec{y}) = \left(\sum_{i=1}^{n} |a x_i - a y_i|^p\right)^{1/p}$

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

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

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

$= |a| \cdot \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} = |a| \, d(\vec{x}, \vec{y})$ □

ii) $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$

$d(\vec{x}+\vec{z}, \vec{y}+\vec{z}) \overset{!}{=} d(\vec{x}, \vec{y})$

Proof: $d(\vec{x}+\vec{z}, \vec{y}+\vec{z}) = \left(\sum_{i=1}^{n} |(x_i + z_i) - (y_i + z_i)|^p \right)^{1/p}$

$= \left(\sum_{i=1}^{n} |x_i + z_i - y_i - z_i|^p \right)^{1/p}$

$= \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = d(\vec{x}, \vec{y})$ □

c) Homogeneity.

$d(a\vec{x}, a\vec{y}) \overset{!}{=} |a| \, d(\vec{x}, \vec{y})$

$\hookrightarrow d(a\vec{x}, a\vec{y}) = \begin{cases} 0 & \text{if } a\vec{x} = a\vec{y} \\ 1 & \text{if } a\vec{x} \neq a\vec{y} \end{cases}$ → $a\vec{x} - a\vec{y} \Rightarrow \vec{x} = \vec{y}$

$\neq |a| \, d(\vec{x}, \vec{y})$ ↝ does not apply.

↝ multiplying both vectors by a scalar will not make a change to function v), as the scalars would cancel each other out.

d) Translation invariance.

↝ $\sqrt{\sum_{i=1}^{n} x_i^2} = \|\vec{x}\|$

$d(\vec{x}, \vec{y}) \overset{!}{=} d(\vec{x}+\vec{z}, \vec{y}+\vec{z})$

$d(\vec{x}, \vec{y}) = \frac{2}{\pi} \operatorname{acos}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \right)$

$= \frac{2}{\pi} \operatorname{acos}\left(\frac{\|\vec{x}\| \|\vec{y}\| \cos(\vec{x}, \vec{y})}{\|\vec{x}\| \cdot \|\vec{y}\|} \right)$

$= \frac{2}{\pi} \operatorname{acos}\left(\cos(\underbrace{\vec{x}, \vec{y}}_{:= \theta}) \right) = \frac{2}{\pi} \theta$

$$d(\vec{x}+\vec{z}, \vec{y}+\vec{z}) = \frac{2}{\pi} \arccos\left(\frac{\|\vec{x}+\vec{z}\| \, \|\vec{y}+\vec{z}\| \, \cos(\vec{x}+\vec{z}, \vec{y}+\vec{z})}{\|\vec{x}+\vec{z}\| \, \|\vec{y}+\vec{z}\|} \right)$$

$$= \frac{2}{\pi} \arccos\left(\cos\left(\underbrace{\vec{x}+\vec{z}, \vec{y}+\vec{z}}_{=\theta} \right) \right) = \frac{2}{\pi}\theta = \underline{d(\vec{x}, \vec{y})}$$ □

→ angle stays same if we add same vector to both points

↝ Translation invariance applies.