IMPORTANCE OF NORMALIZATION

SYRACUSE UNIVERSITY
School of Information Studies

IMPORTANCE OF NORMALIZATION 1

Different variables might use different scales

Age: [0,120]

Income: [0,2M]

If averaging the difference on these two variables, income would weigh much more than age.

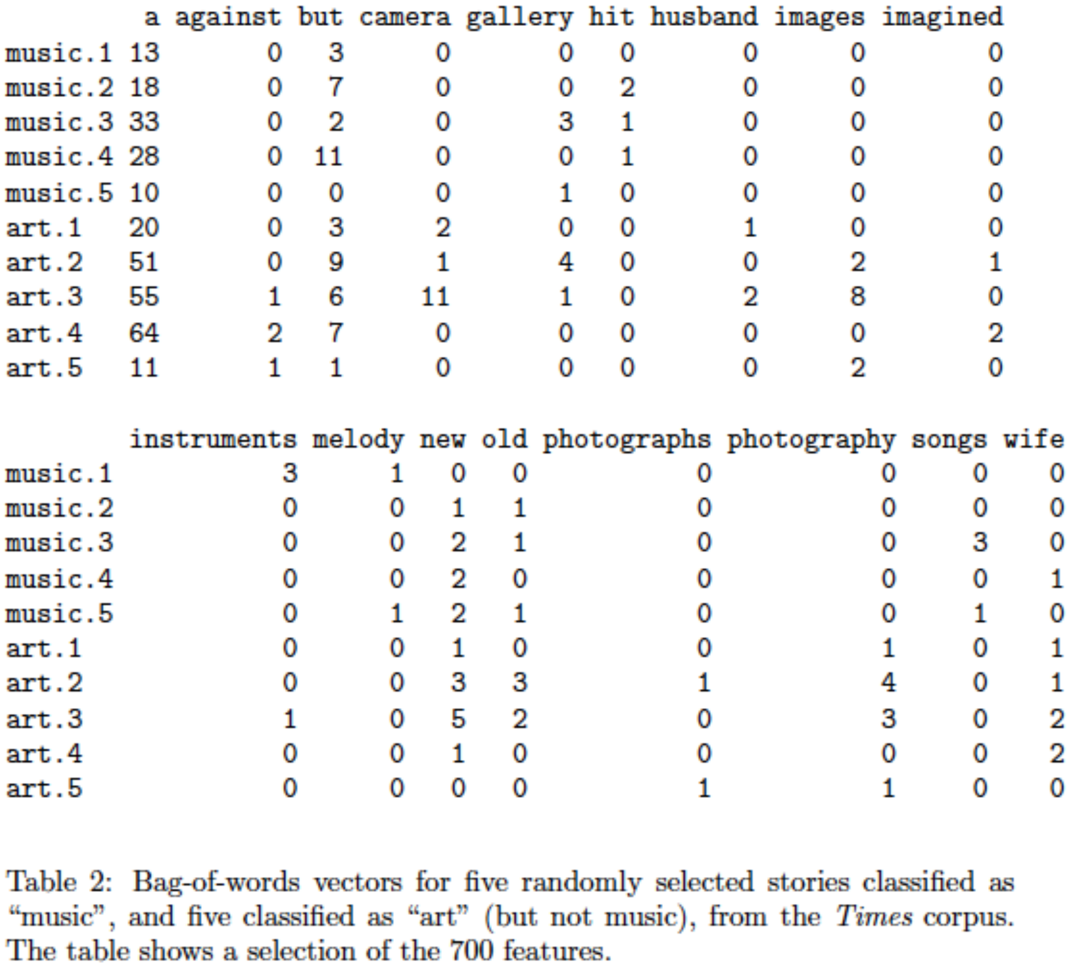
Solution: Normalize both variables to the same scale, e.g., [0,1].

IMPORTANCE OF NORMALIZATION 2

Different examples or vectors might differ greatly in length.

E.g., for text documents, the vector lengths of long documents are much greater than for short documents.

AN EXAMPLE OF NORMALIZATION IN INFORMATION RETRIEVAL



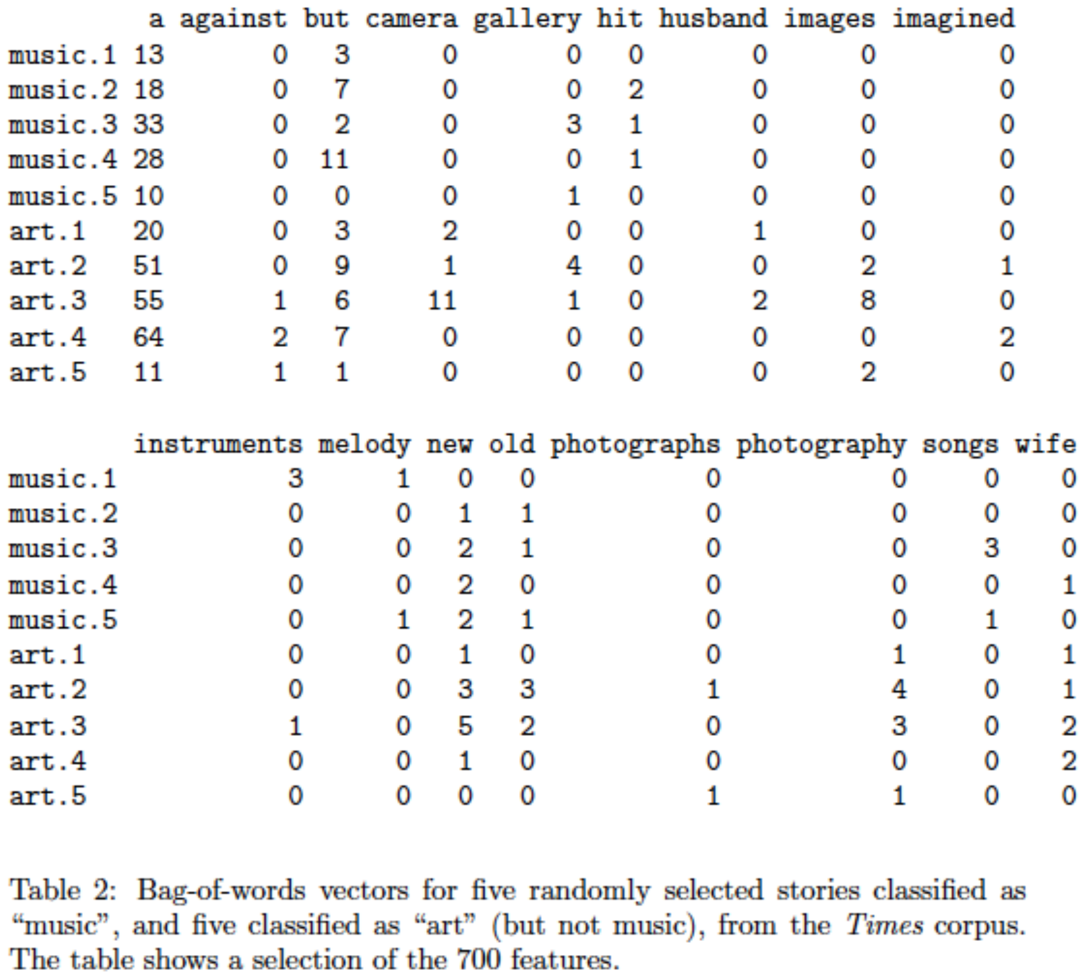
	a	against	but	camera	gallery	hit	husband	images	imagined
music.1	13	0	3	0	0	0	0	0	0
music.2	18	0	7	0	0	2	0	0	0
music.3	33	0	2	0	3	1	0	0	0
music.4	28	0	11	0	0	1	0	0	0
music.5	10	0	0	0	1	0	0	0	0
art.1	20	0	3	2	0	0	1	0	0
art.2	51	0	9	1	4	0	0	2	1
art.3	55	1	6	11	1	0	2	8	0
art.4	64	2	7	0	0	0	0	0	2
art.5	11	1	1	0	0	0	0	2	0

	instruments	melody	new	old	photographs	photography	songs	wife
music.1	3	1	0	0	0	0	0	0
music.2	0	0	1	1	0	0	0	0
music.3	0	0	2	1	0	0	3	0
music.4	0	0	2	0	0	0	0	1
music.5	0	1	2	1	0	0	1	0
art.1	0	0	1	0	0	1	0	1
art.2	0	0	3	3	1	4	0	1
art.3	1	0	5	2	0	3	0	2
art.4	0	0	1	0	0	0	0	2
art.5	0	0	0	0	1	1	0	0

Table 2: Bag-of-words vectors for five randomly selected stories classified as "music", and five classified as "art" (but not music), from the *Times* corpus. The table shows a selection of the 700 features.

http://www.stat.cmu.edu/~cshalizi/350/lectures/01/lecture-01.pdf

LONGER DOCS TEND TO BE FAR AWAY FROM SHORT ONES BASED ON RAW EUCLIDEAN DISTANCE

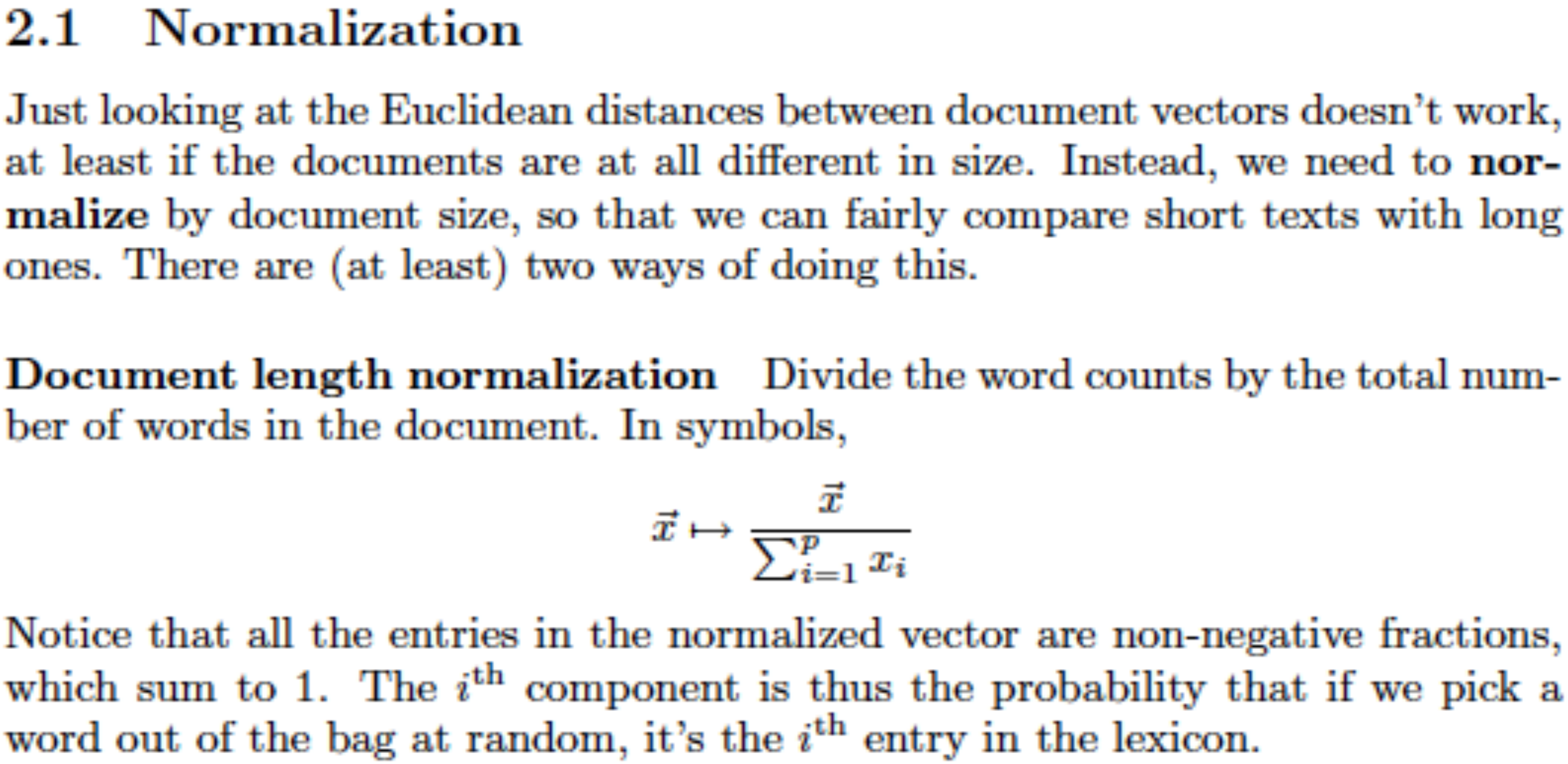


	a	against	but	camera	gallery	hit	husband	images	imagined	instruments	melody	new	old	photographs	photography	songs	wife
music.1	13	0	3	0	0	0	0	0	0	3	1	0	0	0	0	0	0
music.2	18	0	7	0	0	2	0	0	0	0	0	1	1	0	0	0	0
music.3	33	0	2	0	3	1	0	0	0	0	0	2	1	0	0	3	0
music.4	28	0	11	0	0	1	0	0	0	0	0	2	0	0	0	0	1
music.5	10	0	0	0	1	0	0	0	0	0	1	2	1	0	0	1	0
art.1	20	0	3	2	0	0	1	0	0	0	0	1	0	0	1	0	1
art.2	51	0	9	1	4	0	0	2	1	0	0	3	3	1	4	0	1
art.3	55	1	6	11	1	0	2	8	0	1	0	5	2	0	3	0	2
art.4	64	2	7	0	0	0	0	0	2	0	0	1	0	0	0	0	2
art.5	11	1	1	0	0	0	0	2	0	0	0	0	0	1	1	0	0

Table 2: Bag-of-words vectors for five randomly selected stories classified as "music", and five classified as "art" (but not music), from the *Times* corpus. The table shows a selection of the 700 features.

http://www.stat.cmu.edu/~cshalizi/350/lectures/01/lecture-01.pdf

NORMALIZATION BY DOC LENGTH (L-1)



2.1 Normalization

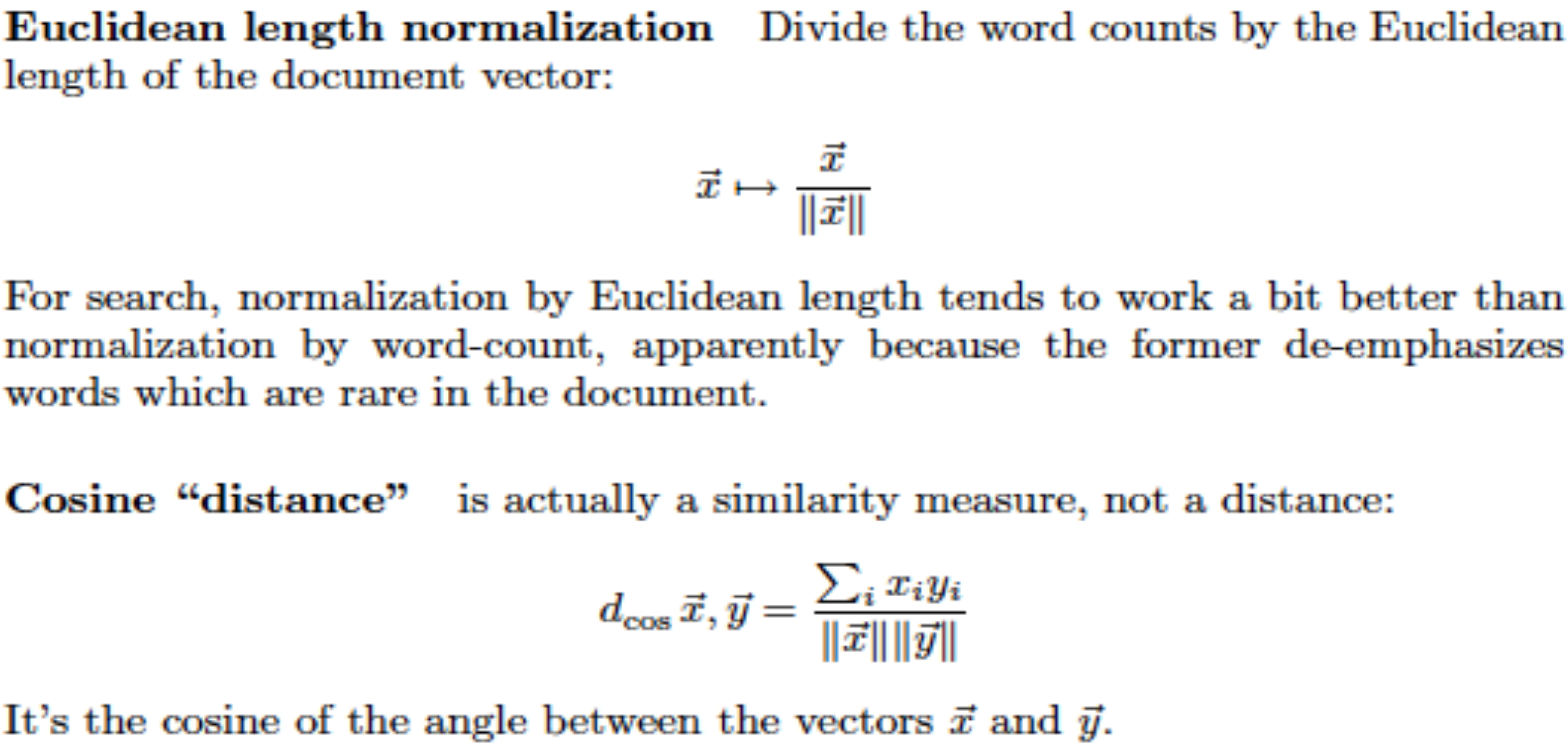
Just looking at the Euclidean distances between document vectors doesn't work, at least if the documents are at all different in size. Instead, we need to **normalize** by document size, so that we can fairly compare short texts with long ones. There are (at least) two ways of doing this.

Document length normalization Divide the word counts by the total number of words in the document. In symbols,

$$\vec{x} \mapsto \frac{\vec{x}}{\sum_{i=1}^{p} x_i}$$

Notice that all the entries in the normalized vector are non-negative fractions, which sum to 1. The i^{th} component is thus the probability that if we pick a word out of the bag at random, it's the i^{th} entry in the lexicon.

NORMALIZATION BY EUCLIDEAN LENGTH (L-2)



Euclidean length normalization Divide the word counts by the Euclidean length of the document vector:

$$\vec{x} \mapsto \frac{\vec{x}}{\|\vec{x}\|}$$

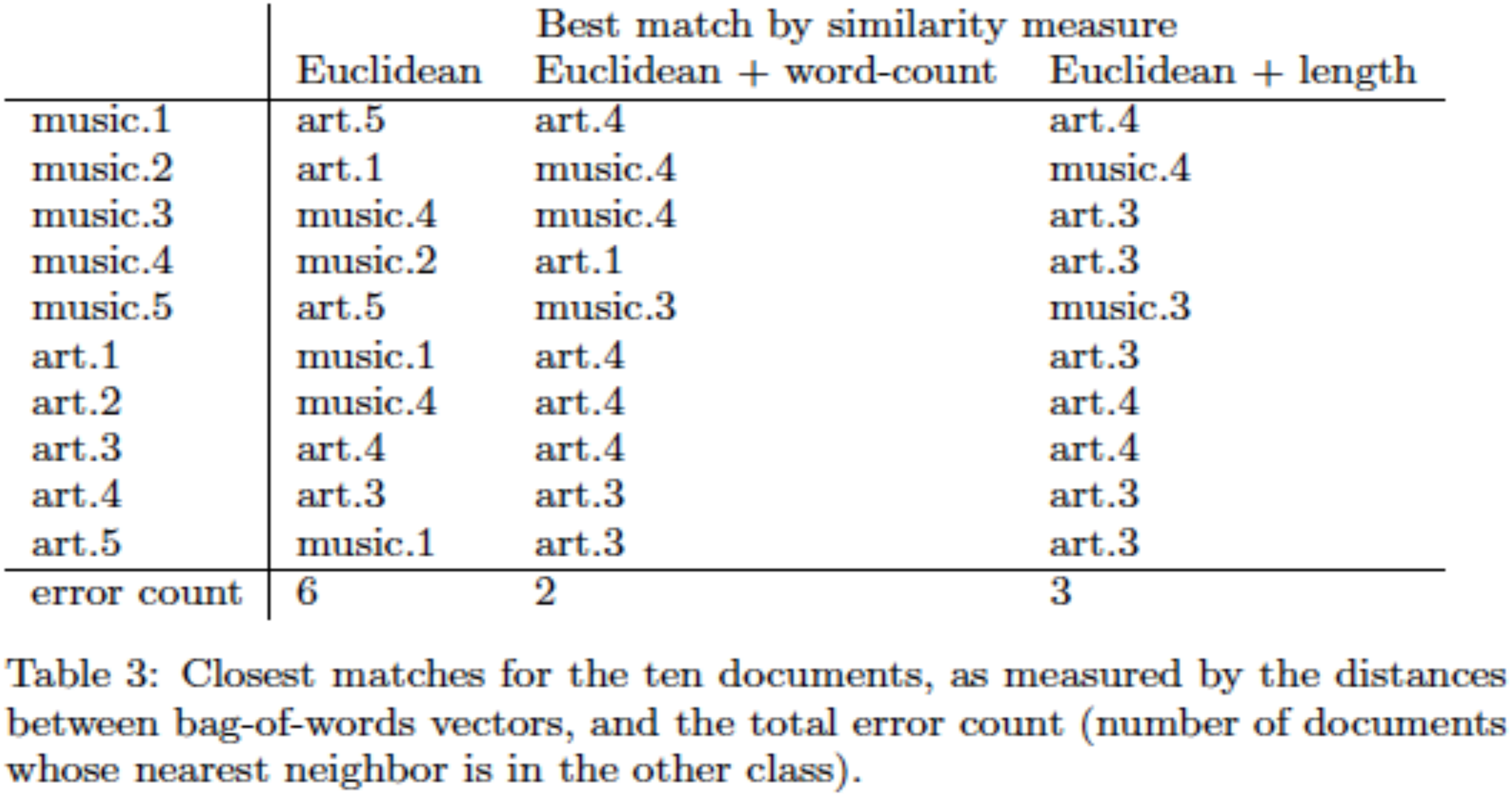
For search, normalization by Euclidean length tends to work a bit better than normalization by word-count, apparently because the former de-emphasizes words which are rare in the document.

Cosine “distance” is actually a similarity measure, not a distance:

$$d_{\cos} \vec{x}, \vec{y} = \frac{\sum_i x_i y_i}{\|\vec{x}\| \|\vec{y}\|}$$

It’s the cosine of the angle between the vectors $\vec{x}$ and $\vec{y}$.

COMPARE RESULTS WITH AND WITHOUT NORMALIZATION



		Best match by similarity measure	
	Euclidean	Euclidean + word-count	Euclidean + length
music.1	art.5	art.4	art.4
music.2	art.1	music.4	music.4
music.3	music.4	music.4	art.3
music.4	music.2	art.1	art.3
music.5	art.5	music.3	music.3
art.1	music.1	art.4	art.3
art.2	music.4	art.4	art.4
art.3	art.4	art.4	art.4
art.4	art.3	art.3	art.3
art.5	music.1	art.3	art.3
error count	6	2	3

Table 3: Closest matches for the ten documents, as measured by the distances between bag-of-words vectors, and the total error count (number of documents whose nearest neighbor is in the other class).

WEIGHTED DISTANCE AND SIMILARITY

For some data, some dimensions are more important than others, and thus their similarity or distance carries more weight. In these cases, we can assign different weights to individual dimensions or attributes.

E.g.: $d(i,j)=2\cdot\,|x_{i1}\quad x_{j1}|+|x_{i2}\quad x_{j2}|$

$$similarity(p,q)=\frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q)=\left(\sum_{k=1}^{n} w_k |p_k - q_k|^r\right)^{1/r}.$$

SIMILARITY AND DISTANCE MEASURE IN R

```
# define a function that calculates the Euclidean distance between two vectors a and b
ED = function(a,b) sqrt(sum((a-b)^2))

# define a function that calculates the cosine similarity between two vectors a and b
CS = function(a,b) a%*% b/sqrt(a%*%a*b%*%b)

# given two vectors a and b
A = c(1,2,3)
B = c(4,5,6)

# call functions to calculate distance
ED(a,b) = 5.196
CS(a,b) = 0.975
```

SUMMARY OF DISTANCE AND SIMILARITY

Manhattan and Euclidean distance for numeric variables

Properties of distance measure

Distance for nominal variables: Count matches or convert to binary

Symmetric vs. asymmetric binary variables

Convert ordinal to either numeric or nominal

Calculate distance or similarity on each attribute or attribute group and then average over all

Cosine similarity

Importance of normalization