Lecture 3: Distance and Similarity

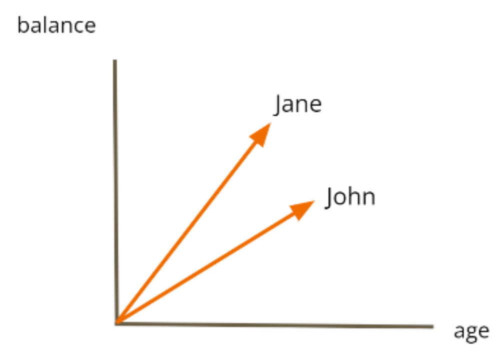
1) Data

$$\textbf{n}\text{ data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\text{ features}}\right.$$

2) Feature Space

a) From data generate feature space of all possible values for set of features in data

name	age	balance
Jane	25	150
John	30	100



Our feature space is the Euclidean plane

3) Distance

a) **d** is distance function iff:
- d(i, j) = 0 if and only if i = j
- d(i, j) = d(j, i)
- d(i, j) ≤ d(i, k) + d(k, j)

b) Don't need distance function to compare points, but prefer it

4) Minkowski Distance

a) For **x**, **y** points in **d**-dimensional real space

i) x=[x_1, …, x_d] and y=[y_1, …, y_d]

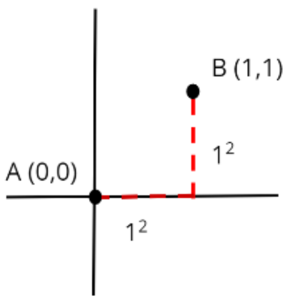
p ≥ 1

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

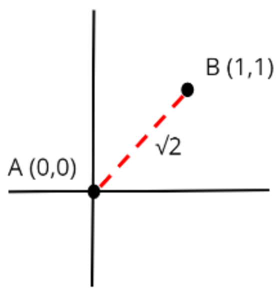
When **p** = 2 -> Euclidean Distance

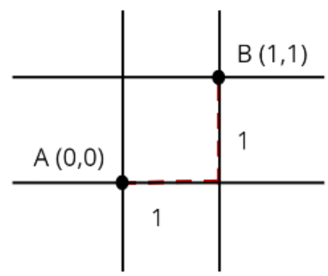
When **p** = 1 -> Manhattan Distance

b) Ex: d = 2 and p = 2

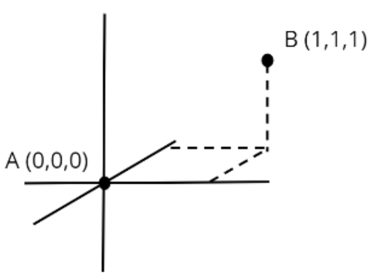


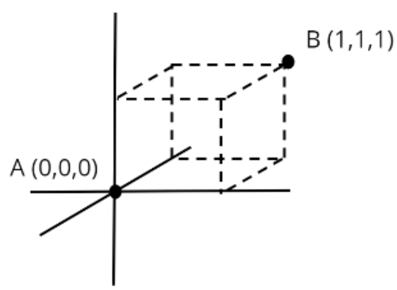
$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

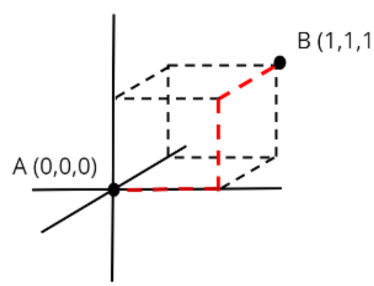




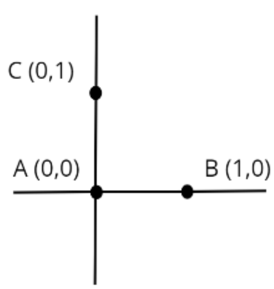
c) Ex: d = 3 and p = 2







d) Is L_p distance function when 0 < p < 1



D(B,A) = D(A, C) = 1

D(B, C) = $2^{1/p}$

D(B,A) + D(A, C) = 2

D(B, C) = $2^{1/p}$

But… if **p < 1** then **1/p > 1**

So **D(B, C) > D(B, A) + D(A, C)** which violates the triangle inequality

5) Cosine Similarity

a) Similarity function: function that takes 2 objects(data points) and return a large value if these objects are similar

  i) $s(x, y) = \cos(\theta)$ where θ is angle between x and y

b) To get dissimilarity function, try:

$$d(x, y) = 1 / s(x, y)$$

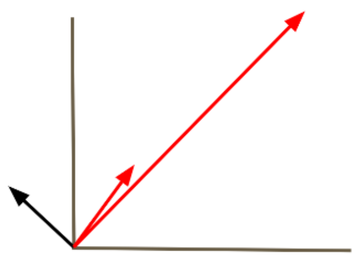
or

$$d(x, y) = k - s(x, y) \text{ for some } k$$

Here, we can use

$$d(x, y) = 1 - s(x, y)$$

c) Use cosine (dis)similarity over euclidean distance when direction matters more than magnitude



Close under Cosine Similarity

6) Jaccard Similarity

a) How similar are the following documents?

	w_1	w_2	...	w_d
x	1	0	...	1
y	1	1	...	0

  i) One way is to use Manhattan distance -> return size of set difference

$$L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

Will only be 1 when $x_i \neq y_i$

b) How to distinguish between 2 cases?

	w_1	w_2	...	w_{d-1}	w_d
x	1	1	1	0	1
y	1	1	1	1	0

Only differ on the last two words

	w_1	w_2
x	0	1
y	1	0

Completely different

Both have Manhattan distance of 2

i) Need to account for size of intersection

$$JSim(x,y) = \frac{|x \cap y|}{|x \cup y|} \qquad JDist(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

7) Norms

a) Distance from the origin
   i) Minkowski Distance <-> Lp Norm
   ii) Not all distances can create a norm
b) Notion of size
c) Has the properties:
   - $p(x + y) \leq p(x) + p(y)$
   - $p(ax) = |a| p(x)$
   - $p(x) = 0$ iff $x = 0$
   - $p(x) \geq 0$ for all x