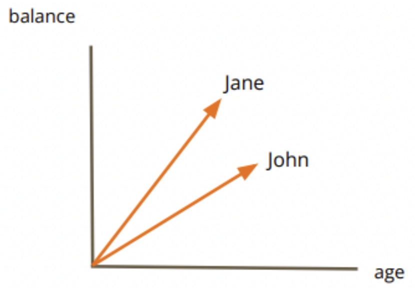
CAS CS 506
Lec 04

Distance & Similarity

1. Data
    a. $$\textbf{n}\text{ data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\text{ features}}\right.$$
2. Feature Space
    a. From our data we can generate a feature space of all possible values for the set of features in our data
        i.

name	age	balance
Jane	25	150
John	30	100

$\longrightarrow$



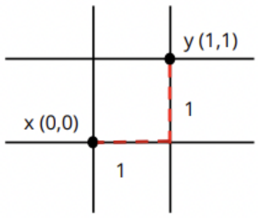
        ii. This is 2-dimensional data
3. Dissimilarity
    a. In order to uncover interesting structures from our data, we need a way to compare data points
    b. A dissimilarity function is a function that takes two objects (data points) and returns a large value if these objects are dissimilar
    c. A special type of dissimilarity function is a distance function
4. Distance
    a. d is a distance function if and only if:
        i. d(i, j) = 0 if and only if i = j
        ii. d(i, j) = d(j, i)
        iii. d(i, j) <= d(i, k) + d(k, j)
    b. We don't need a distance function to compare data points, but why would we prefer using a distance function?
        i. It is intuitive

5. Minkowski Distance
    a. For x, y points in d-dimensional real space
    b. I. e. $x = [x_1, \ldots, x_d]$ and $y = [y_1, \ldots, y_d]$
    c. $p >= 1$

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

    d. d is the dimensional space.
    e. When p = 2, Euclidean Distance
    f. When p = 1, Manhattan Distance
6. Example

d = 2

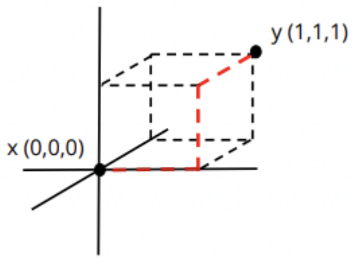


p = 1

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

    a.

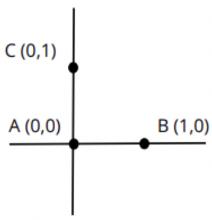
d = 3



p = 1

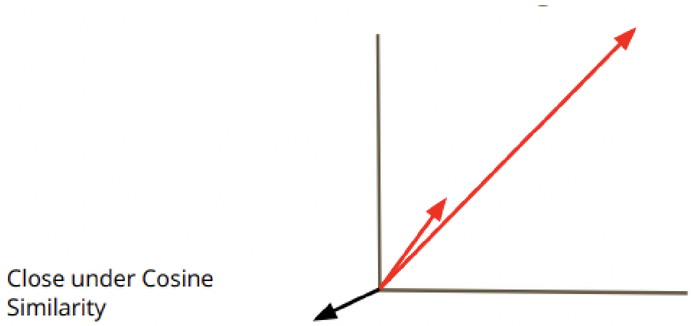
$$L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

    b.
7. Minkowski Distance
    a. Is L_p a distance function when $0 < p < 1$?
        i. no



    b.

   c. D(B, A) + D(A, C) = 2
      D(B, C) = 2 ^ (1/p)
   d. So D(B, C) > D(B, A) + D(A, C) which violates the triangle inequality
8. Cosine Similarity
   a. A similarity function is a function that takes two objects (data points) and returns a large value if these objects are similar
      s(x,y) = cos(theta)
      where theta is the angle between x and y
   b. Two proportional vectors have a cosine similarity of 1
   c. Two orthogonal vectors have a similarity of 0
   d. Two opposite vectors have a similarity of -1
   e. To get a corresponding dissimilarity function, we can usually try
      d(x,y) = 1/s(x,y)
      or
      d(x,y) = k - s(x,y) for some k
   f. Here, we can use
      d(x,y) = 1 - s(x,y)
   g. We use cosine dissimilarity over Euclidean distance when direction matters more than magnitude



9. Jaccard Similarity
   a. How similar are the following documents?

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

   b. One way is to use the Manhattan distance which will return the size of the set difference

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

Will only be 1 when $x_i \neq y_i$

c. How to distinguish between the two cases where both the Manhattan distance is the same (2 in this case)

	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

d. We need to account for the size of the intersection
e. Given two documents x and y:

$$JSim(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

   i. x is the set of words, not the binary vector representation

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

f.

10. A Quick Note on 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 following properties
      i. p(x + y) <= p(x) + p(y)
      ii. p(ax) = |a| p(x)
      iii. p(x) = 0 iff x = 0
      iv. p(x) >= 0 for all x