Distance & Similarity

Boston University CS 506 - Lance Galletti

Refund	Marital Status	Income	Age

Refund	Marital Status	Income	Age
1	Single	125k	25

Refund	Marital Status	Income	Age
1	Single	125k	25
0	Married	100k	27

Refund	Marital Status	Income	Age
1	Single	125k	25
0	Married	100k	27
0	Single	70k	22

Refund	Marital Status	Income	Age
1	Single	125k	25
0	Married	100k	27
0	Single	70k	22
1	Married	120k	30
0	Divorced	90k	28
0	Married	60k	37
1	Divorced	220k	24
0	Single	85k	23
0	Married	75k	23
0	Single	90k	26

Data

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

Data

n data points

Data point i

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

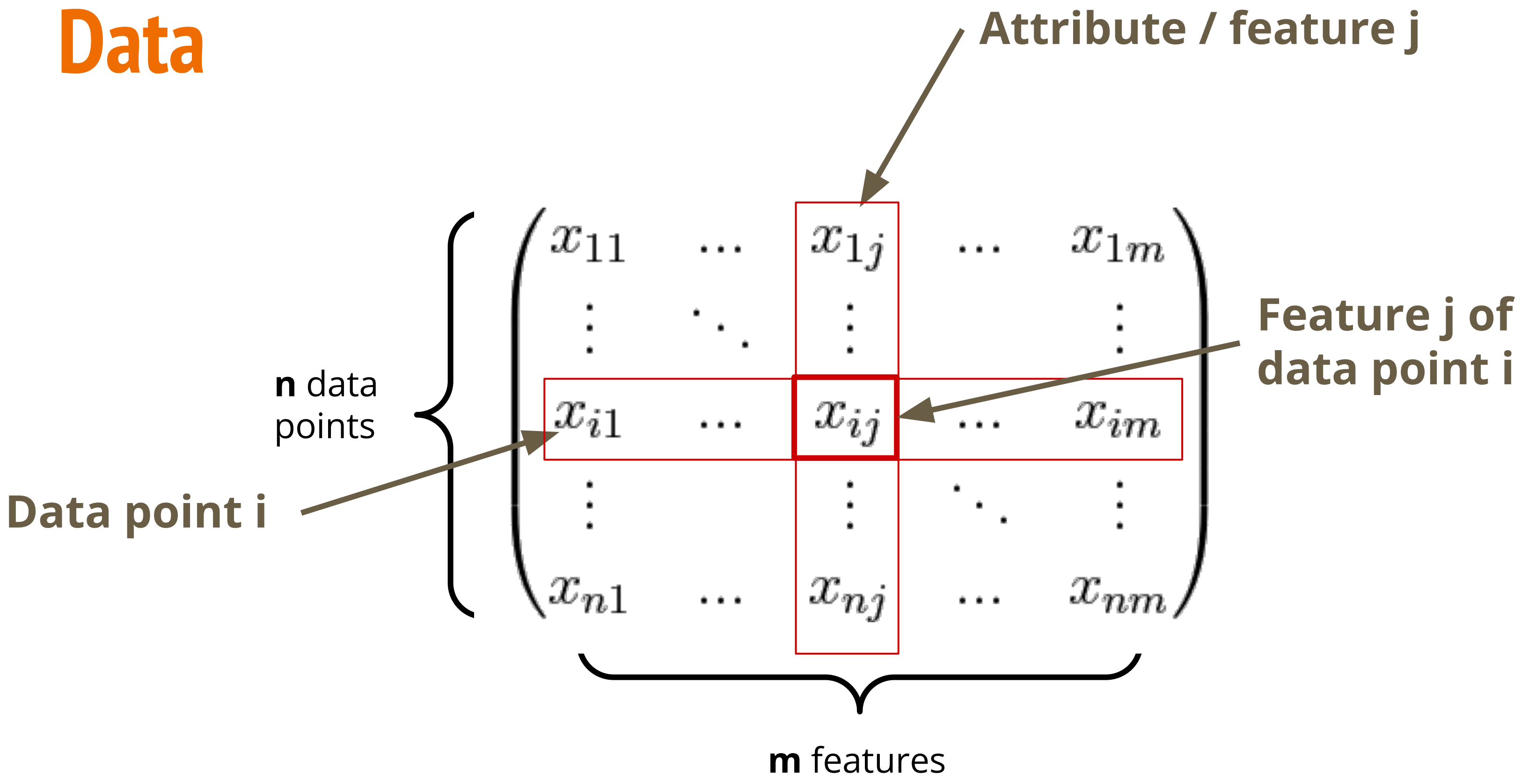
Data

Attribute / feature j

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

Data point i

Data



a data point would be like a person and features/attributes would be age, sex, race, etc.

Feature Space

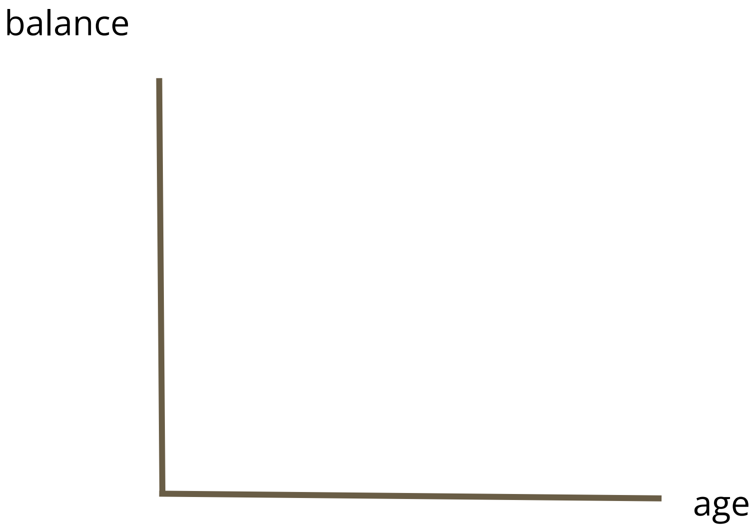
From our data we can generate a **feature space** of all possible values for the set of features in our data.

name	age	balance
Jane	25	150
John	30	100

Feature Space

From our data we can generate a **feature space** of all possible values for the set of features in our data.

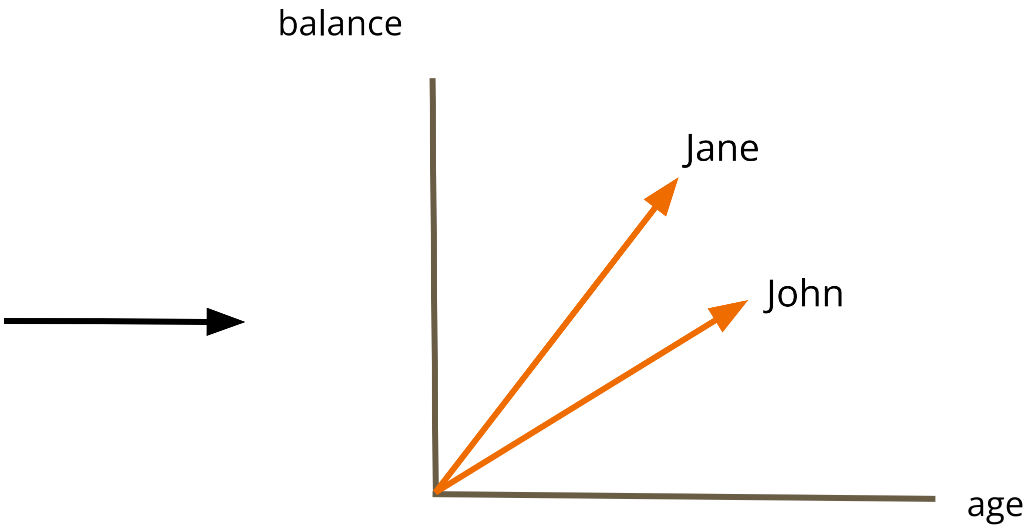
name	age	balance
Jane	25	150
John	30	100



Feature Space

From our data we can generate a **feature space** of all possible values for the set of features in our data.

name	age	balance
Jane	25	150
John	30	100



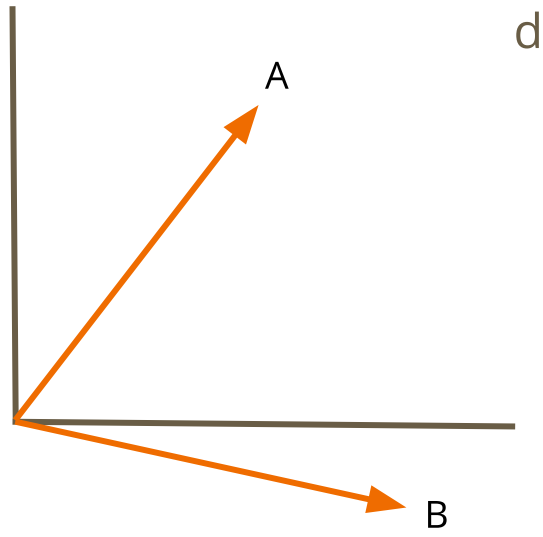
Our feature space is the Euclidean plane

Dissimilarity

In order to uncover interesting structure from our data, we need a way to **compare** data points.

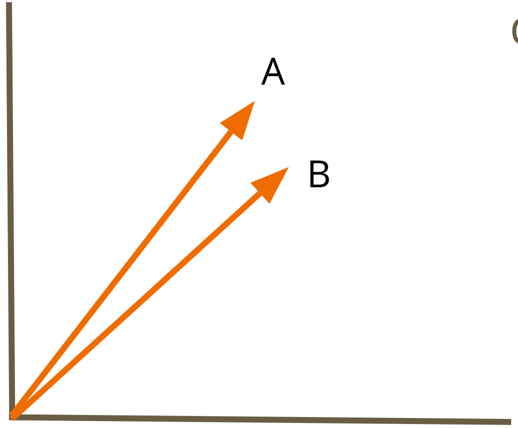
A **dissimilarity function** is a function that takes two objects (data points) and returns a **large value** if these objects are **dissimilar**.

Dissimilarity



dissim(A, B) is large

Dissimilarity



dissim(A, B) is small

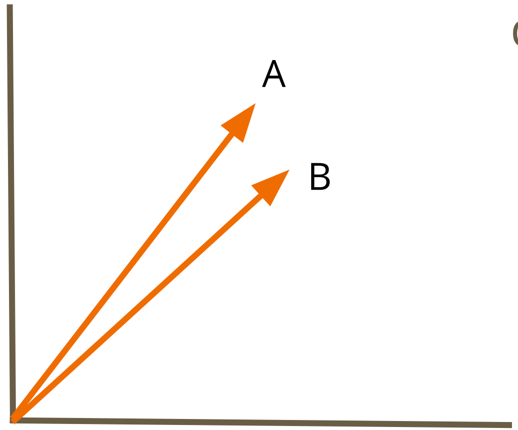
Distance

A special type of dissimilarity function is a **distance** function

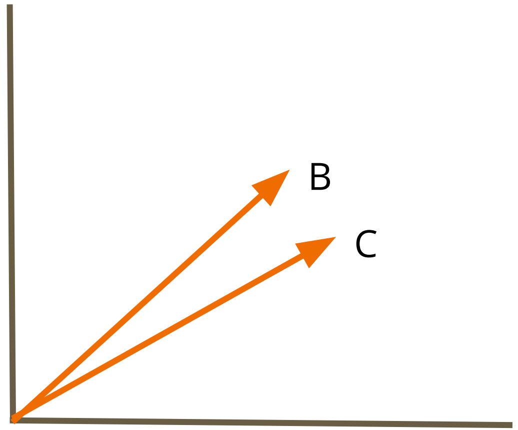
d is a distance function if and only if:

- $d(i, j) = 0$ if and only if $i = j$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$

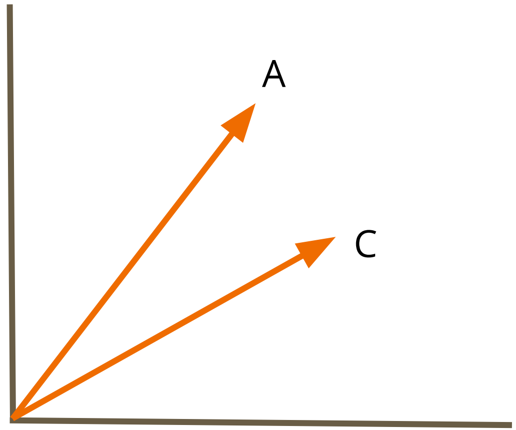
We don't **need** a distance function to compare data points, but why would we prefer using a distance function?



dissim(A, B) is small

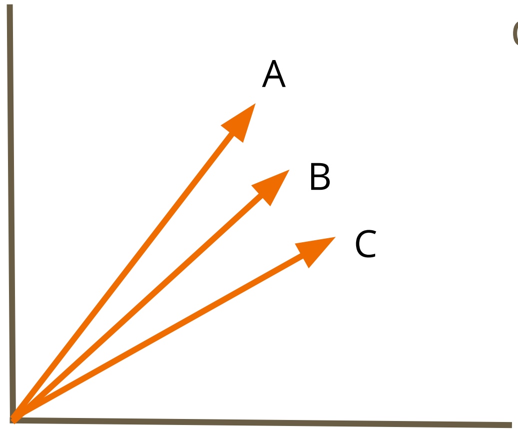


dissim(B, C) is small



dissim(A, C) not necessarily small

Using any random dissim function, we don't get a guarantee that if AB close and BC close, then AC close
But with distance = d(), we can guarantee that if AB close and BC close, then AC close



d(A, B) is small

d(B, C) is small

Triangle inequality guarantees d(A, C) small

dissum is also allowed to not be symmetric, unlike the distance function will be symmetric. AKA if AB close, then BA is close also

Minkowski Distance

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

I.e. $\mathbf{x} = [x_1, \dots, x_d]$ and $\mathbf{y} = [y_1, \dots, y_d]$

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

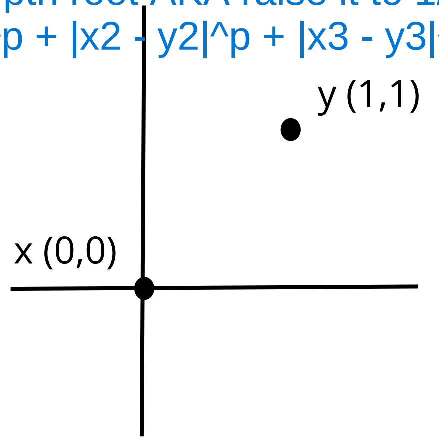
When **p** = 1 -> Manhattan Distance

Example

when d = 2, it is the following
summation from i = 1 to 2 of |xi - yi| to the power of p, and then take the whole thing to the pth root AKA raise it to 1/9
= (|x1 - y1|^p + |x2 - y2|^p) ^ (1/p)

d = 2

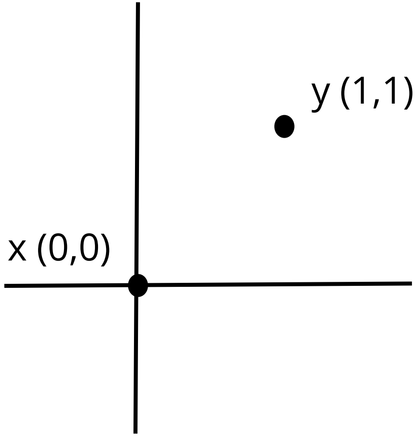
when d = 3, it is the following
summation from i = 1 to 3 of |xi - yi| to the power of p, and then take the whole thing to the pth root AKA raise it to 1/9
= (|x1 - y1|^p + |x2 - y2|^p + |x3 - y3|^p) ^ (1/p)



p is simply a paramter that is up to us to choice.
We can set p to be different things and that will impact how the data is interpreted. If we don't like how a specific p makes some data ask, we can change it

Example

d = 2

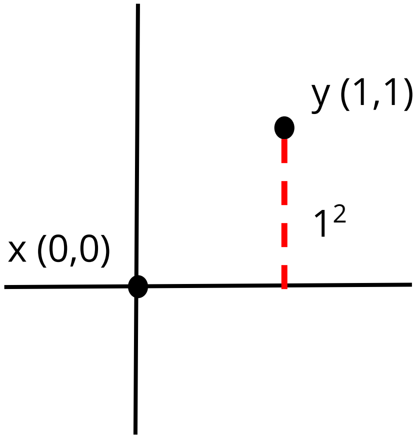


p = 2

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

Example

d = 2

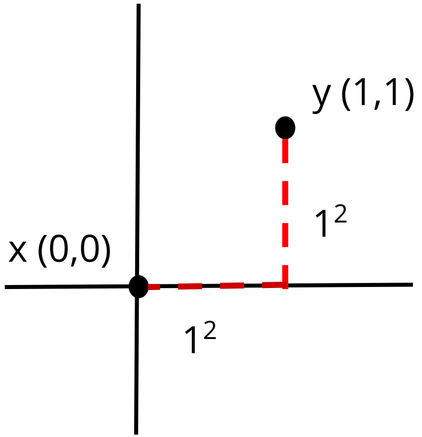


p = 2

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

Example

d = 2



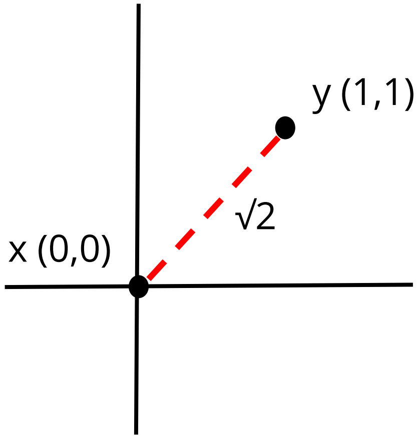
p = 2

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

Example

d = 2

Euclidean distance = sqr root of 2

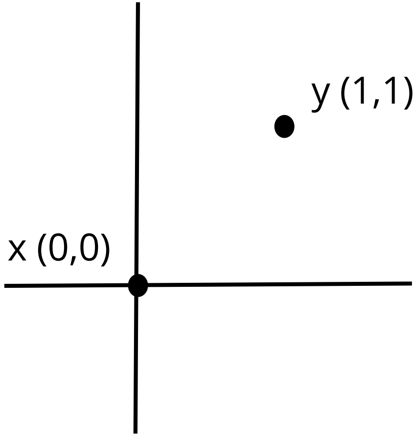


p = 2

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

Example

d = 2

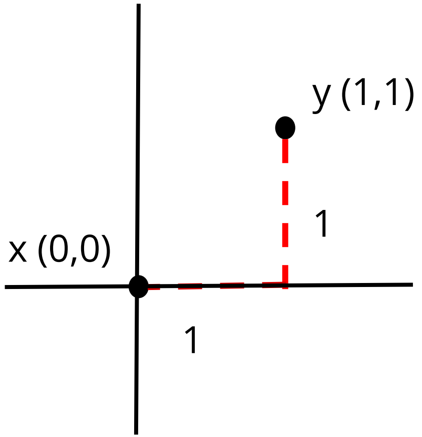


p = 1

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

Example

d = 2



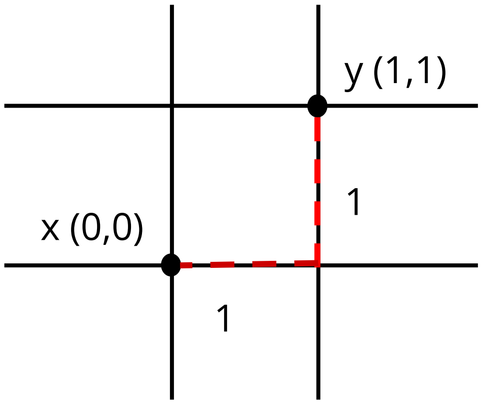
p = 1

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

Example

manhattan distance = 2

d = 2

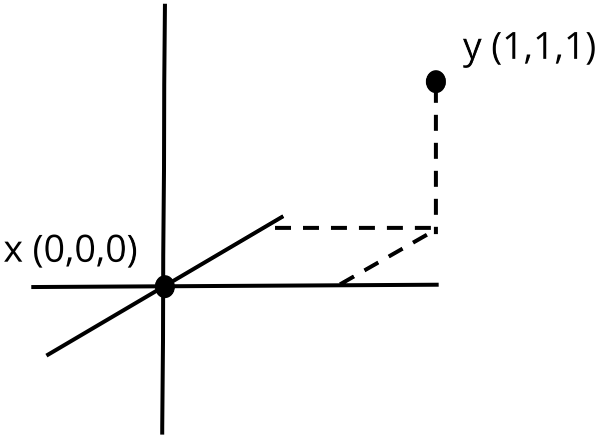


p = 1

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

Example

d = 3

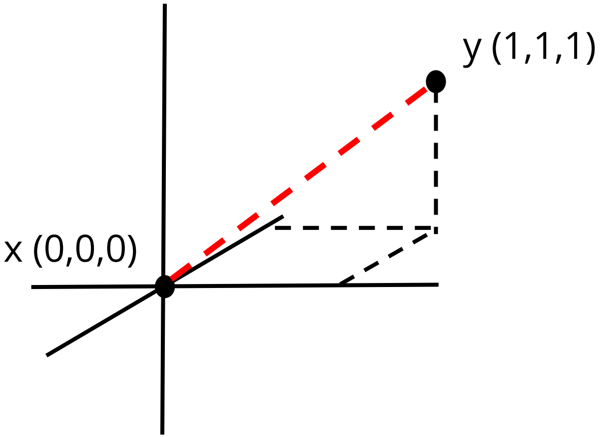


p = 2

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

Example

d = 3

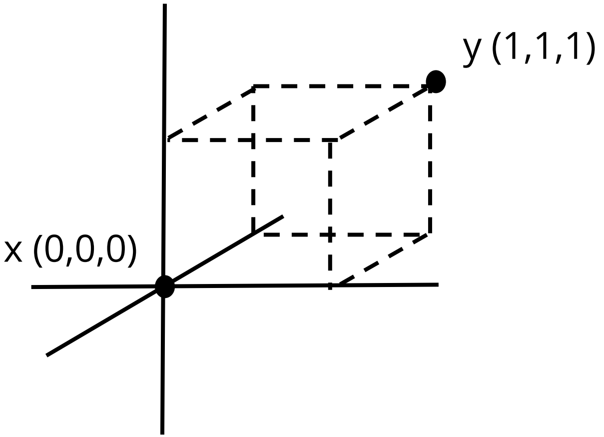


p = 2

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

Example

d = 3

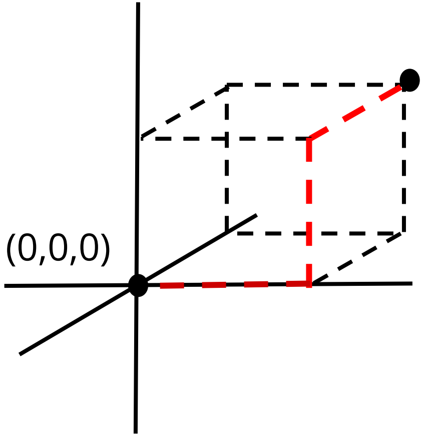


p = 1

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

Example

d = 3



p = 1

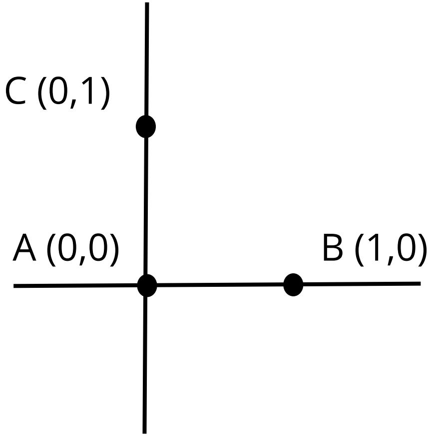
$$L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$?

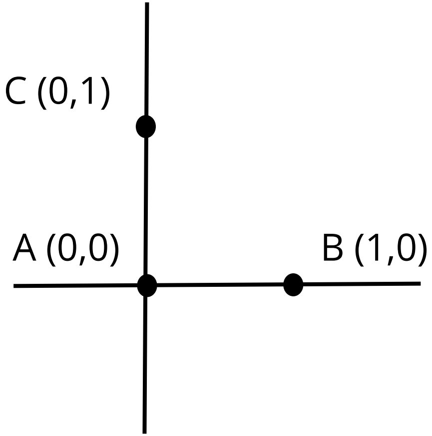
Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$?



Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$?

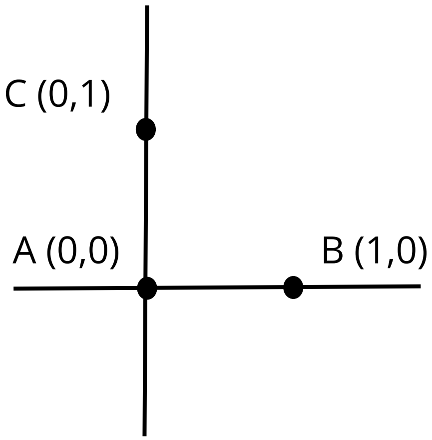


D(B,A) = D(A, C) = 1

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

Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$?



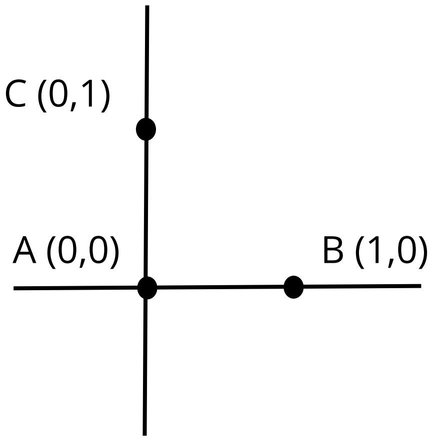
D(B,A) + D(A, C) = 2

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

But… if $\mathbf{p < 1}$ then $\mathbf{1/p > 1}$

Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$?



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

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

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

Jaccard Similarity

How similar are the following documents?

w1 in doc x and in doc y
w2 not in doc x, in doc y
wd in doc x, not in doc y

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

Jaccard Similarity

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

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

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

Jaccard Similarity

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

doesn't account for how big the documents are

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

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

Will only be 1 when $x_i \neq y_i$

Jaccard Similarity

But how can we distinguish between these two 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

Jaccard Similarity

But how can we distinguish between these two 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

Jaccard Similarity

We need to account for the size of the intersection!

Given two documents x and y:

Here x and y are sets of words in the docs, not the binary of the presence of the word in the doc

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

the words in common between doc x and y
over
all the words in doc x and doc y

Jaccard Similarity

We need to account for the size of the intersection!

Given two documents x and y:

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

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

Here, x is the set of words (not the binary vector representation)

Jaccard Similarity

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

assume d = 100

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

	w_1	w_2
x	0	1
y	1	0

Only differ on the last two words

Completely different

Maybe?
1- (98/100) = 1- 0.02 = 0.98

1 - (0/2) = 1- 0 = 1

What is the jaccard distance in each?

Jaccard Similarity

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

Here, x is the set of words (not the binary vector representation)

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(}\theta\mathbf{)}$$

where θ is the angle between **x** and **y**

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\boldsymbol{\theta})}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of:

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\boldsymbol{\theta})}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of:

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

two opposite vectors have a similarity of:

Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

two opposite vectors have a similarity of: - 1

Cosine Similarity

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**

Here, we can use

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

Cosine Similarity

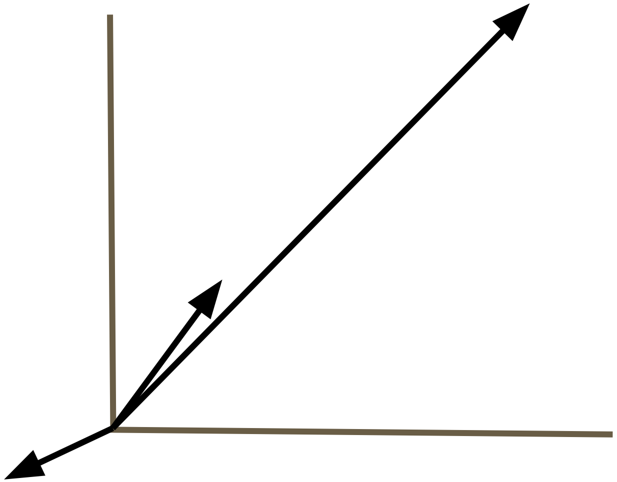
When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

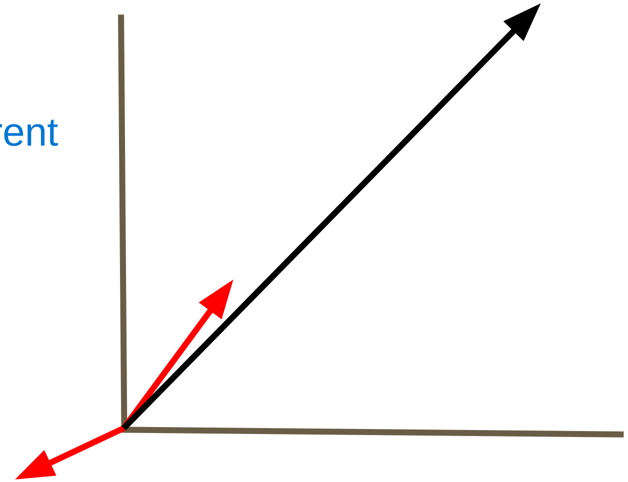


Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

example:
2 abstracts of completely different papers



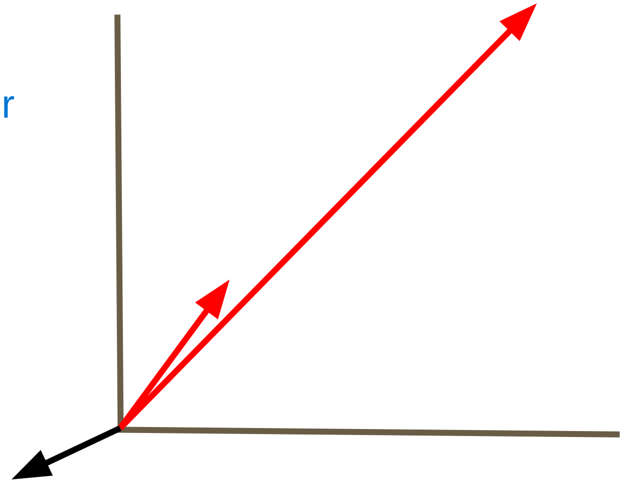
Close under
Euclidean distance

Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

example:
an abstract and the full paper
of that abstract



Close under Cosine
Similarity

A quick Note on Norms

$$d(A, B) = \|A - B\|$$

Size = Distance from the origin $\quad d(0, X) = \|X\|$

- Minkowski Distance <=> Lp Norm
- Not all distances can create a Norm