TDA 1

METRIC SPACES

Eventually: data sets.

Survey questions: $\longrightarrow (x_1, \dots, x_9)$

x_1: age

x_2: height

x_3: # of pets

$\vdots$

$\uparrow$ data point

How are data points compared?

Ex: geographic distance vs. genetic distance

PC : in the US
EC : in France

distance = many miles.

siblings have comparable genetic material

AC : also in the US

distance : short (from DNA point of view).

distance (PC, AC) = small geographically

= large genetically.

Conclusion: different problems require different notions of distance.

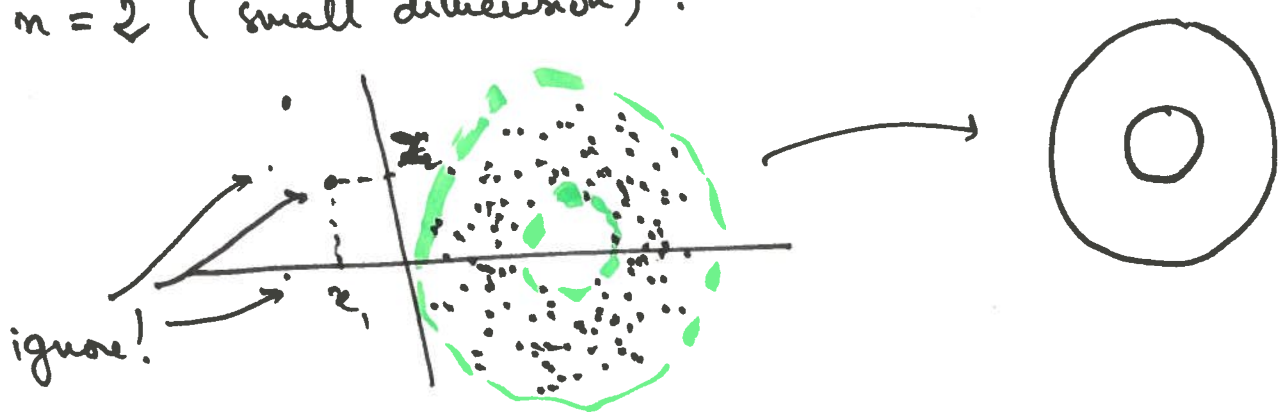
In general: data points belong to $\mathbb{R}^n$

$(x_1, \ldots, x_n)$ n may be large

The number of points may also be large....

Idea of TDA: how to study the shape of a large set in large dimension?

If $n = 2$ (small dimension):



→ what if $n > 3$? No pictures...

→ what if our eyes see no pattern?

Goal: To develop a good theory of "shape" in any dimension.

n = dimension = "number of questions in survey"

We ask n questions or make n measurements.

$\longrightarrow$ for each, we get a point $(x_1, \dots, x_n) \in \mathbb{R}^n$

For each i, $1 \leq i \leq n$, $x_i \in \mathbb{R}$.

x_i = the i^{th} answer for person x

$$x = (x_1, \dots, x_n)$$

If answers are not numerical, make them numerical:

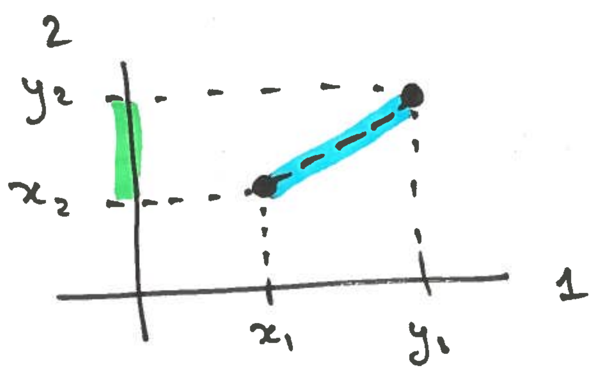
yes — 1 , no — 0

More generally: we want to estimate how close answers are.

Question: how should proximity be measured?

Ex. $n = 2$ $\quad x_1$: age $\quad x = (x_1, x_2)$

x_2 : weight $\quad y = (y_1, y_2)$



$$d(x,y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$$

If we only care about weight: $\delta(x,y) = |y_2 - x_2|$

→ What should a distance satisfy?

Def. Let X be a set. A <u>metric</u> or <u>distance on X</u> is a function

$$d : X \times X \longrightarrow \mathbb{R}$$
$$(x,y) \longmapsto d(x,y)$$

with the following properties:

(M1) $d(x,y) \geq 0$ and $d(x,y) = 0 \iff x = y$

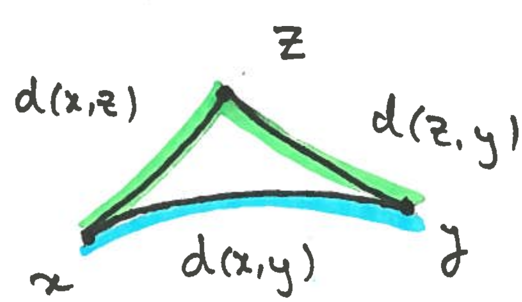
(M2) $d(x,y) = d(y,x)$ - symmetry

(M3) $d(x,y) \leq d(x,z) + d(z,y)$ - triangle ineq.

(M1) says that the only points at distance 0 are identical.

$$\underbrace{x = y}_{\text{in } X} \iff \underbrace{d(x,y) = 0}_{\text{in } \mathbb{R}}$$

(M3) - Triangle...



$$d(x,y) \leq d(x,z) + d(z,y)$$

Ex 1. $X = \mathbb{R}$. $d(x,y) = |x-y|$

Show that d is a distance.

Ex 2. $X = \mathbb{R}^2$. $\delta(x,y) = |y_2 - x_2|$ $x = (x_1, x_2)$, $y = (y_1, y_2)$

Is δ a distance?

Ex 1. $d(x,y) = \underbrace{|x-y|}_{\geqslant 0} = |y-x| = d(y,x)$

(M1) ✓ (M2) ✓

$$d(x,y) = |x-y| = |x - \underbrace{z + z}_{=0} - y|$$ $|a+b| \leq |a| + |b|$

$$\leq |x-z| + |z-y| = d(x,z) + d(z,y)$$

(M3) ✓

Ex 2. $\delta(x,y) = |y_2 - x_2|$ is <u>not</u> a distance.

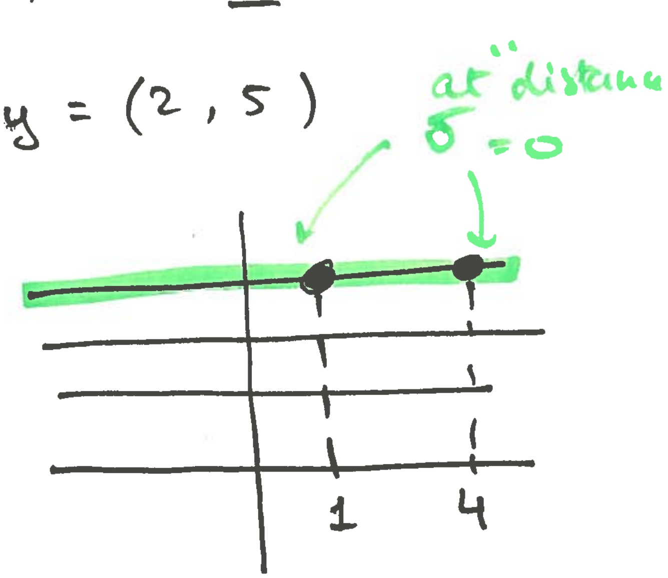
$x = (1,3)$ $y = (2,5)$

$z = (4,3)$

$\delta(x,y) = |5-3| = 2$

$\delta(x,z) = |3-3| = 0$

<u>but</u> $x \neq z$!



HW. The Euclidean distance on $\mathbb{R}^n$.

$x = (x_1, \dots, x_n)$

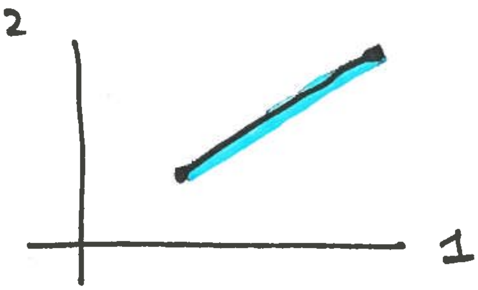
$y = (y_1, \dots, y_n)$

$$d_{\text{Eucl.}}(x,y) = \sqrt{(y_1-x_1)^2 + (y_2-x_2)^2 + \dots + (y_n-x_n)^2}$$

Fact: this is a distance (satisfies (M1), (M2), (M3))

$n=1$ $\quad d_{\text{Eucl.}}(x,y) = \sqrt{(y_1-x_1)^2} = |y_1-x_1|$

$n=2$



HW. $X = \mathbb{R}^2$ $\qquad x = (x_1, x_2)$

$y = (y_1, y_2)$

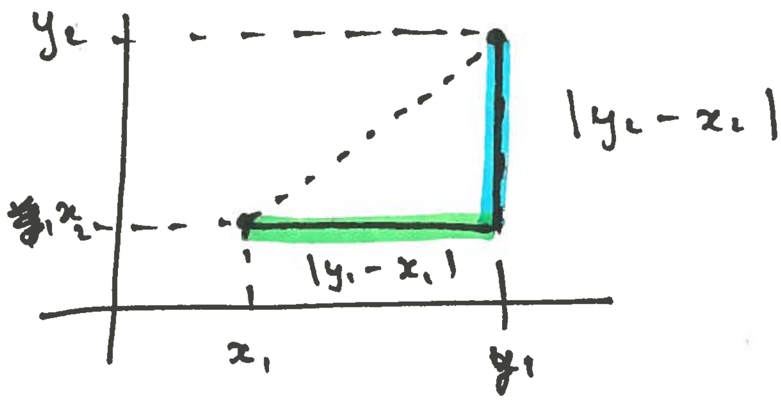
Define $d_1(x,y) = |y_1-x_1| + |y_2-x_2|$

and $d_\infty(x,y) = \max\{|y_1-x_1|, |y_2-x_2|\}$

Q: prove that d_1 and d_∞ are distances on $\mathbb{R}^2$.

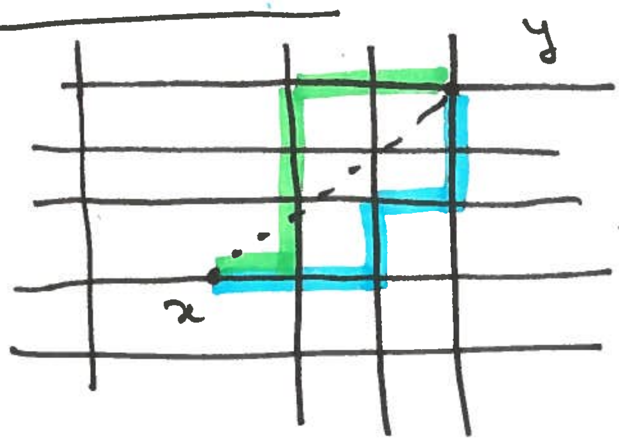
What does "shape" mean if we measure distances with d_1 or d_∞ instead of $d_{Eucl.}$?

$$d_1(x,y) = |y_1 - x_1| + |y_2 - x_2|$$



This is called the Manhattan distance

or taxicab distance:



length: $\underbrace{3 \text{ right}}_{=|y_1 - x_1|} + \underbrace{3 \text{ up}}_{=|y_2 - x_2|} = 6$

<u>Concerning d_∞:</u> can define it on $\mathbb{R}^n$

$$x = (x_1, \ldots, x_n)$$
$$y = (y_1, \ldots, y_n)$$

$$d_\infty(x,y) = \max\{|y_1 - x_1|, |y_2 - x_2|, \ldots, |y_n - x_n|\}$$

<u>$n = 3$</u> $\quad x = (1; 2; 3)$

$y = (1.1; 2,2; 5)$

$$\left.\begin{array}{l} |y_1 - x_1| = 0.1 \\ |y_2 - x_2| = 0.2 \end{array}\right\} \text{small difference}$$

$$|y_3 - x_3| = 2 \quad \} \text{ bigger difference}$$

$$d_\infty(x,y) = 2$$

d_∞ only measures the greatest difference between x and y.

<u>HW</u>. There is also a Manhattan distance on $\mathbb{R}^n$:

$$d_1(x,y) = |y_1 - x_1| + |y_2 - x_2| + \ldots + |y_n - x_n|$$

$\longrightarrow$ Check this is a distance!

→ Given a set X, is there always a distance on X?

On $\mathbb{R}^n$, we saw: $d_{Eucl.}$, d_1, d_∞ ....

If X is a set, define $d : X \times X \longrightarrow \mathbb{R}$ by

$$\left[\begin{array}{lll} d(x,y) = 0 & \text{if} & x = y \\ = 1 & \text{if} & x \neq y \end{array} \right]$$

Ex (HW): verify that this is a distance!

Remark: if $n = 1$ $(X = \mathbb{R})$

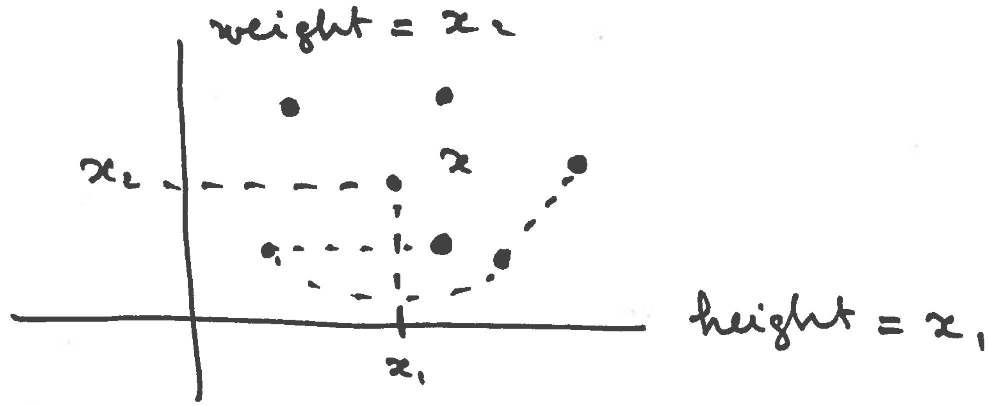
$$\begin{aligned} d_{Eucl.}(x,y) &= \sqrt{(y-x)^2} = |y-x| \\ &= d_1(x,y) \\ &= d_\infty(x,y) \end{aligned}$$

In that case, they are all the same!

Not true for $\mathbb{R}^n$ with $n \geq 2$.

Example: survey asks about height and weight:

put results in $\mathbb{R}^2$:



For this survey, we may use $d_{\text{Eucl.}}$, d_1, d_∞, ...

Special distance on any set: $d_0(x,y) = 0$ if $x = y$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad = 1$ if $x \neq y$.

$\to$ We want to check (M1), (M2), (M3).

(M1) $\quad d_0(x,y) = 0 \iff x = y$ ✓

(M2) $\quad d_0(y,x) = 0 \iff x = y \iff y = x$
$\qquad\qquad\qquad\quad = 1 \iff x \neq y \iff y \neq x$ ✓

(M3) Compare $d_0(x,y)$ and $d_0(x,z) + d_0(z,x)$

1st case: $x = y$

Then $d_0(x,y) = 0$ and $\underbrace{d_0(x,z)}_{\geq 0} + \underbrace{d_0(z,y)}_{\geq 0} \geq 0$

so $d_0(x,y) \leq d_0(x,z) + d_0(z,y)$ ✓

2nd case : $x \neq y$

Then $d_0(x,y) = 1$.

and $d_0(x,z) + d_0(z,y) = 0 + 1$ if $x = z,\ y \neq z$

$= 1 + 0$ if $x \neq z,\ y = z$

$= 1 + 1$ if $x \neq z,\ y \neq z$

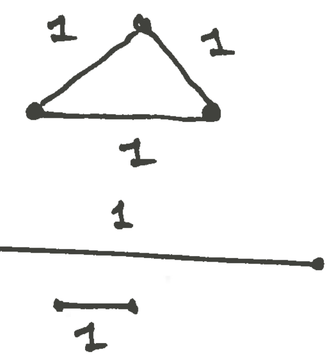
- $d_0(x,y) = 1$ and $d_0(x,z) + d_0(z,y)$ is 1 or 2

  so $d_0(x,y) \leq d_0(x,z) + d_0(z,y)$

($d_0(x,z) + d_0(z,y)$ cannot be 0, that would mean $x = z$ and $z = y$ and we know $x \neq y$)

- $\underbrace{d_0(x,z) + d_0(z,y)}_{}$ $= 2$ can happen

  not the distance between two pts.



In this setting, two points can only be at distance 0 (same) or 1 (different).

(HW) Comparison of $d_{\text{Eucl.}}$, d_1 and d_∞ on $\mathbb{R}^2$:
For $x = (x_1, x_2)$, $y = (y_1, y_2)$ in $\mathbb{R}^2$, prove:

$$\frac{1}{2} d_1(x,y) \leq \frac{1}{\sqrt{2}} d_{\text{Eucl.}}(x,y) \leq d_\infty(x,y)$$

Special sets in a metric space.

Assume X is a set with a distance d.

For $a \in X$ (a element of X)

and $r > 0$ (r is a real number).

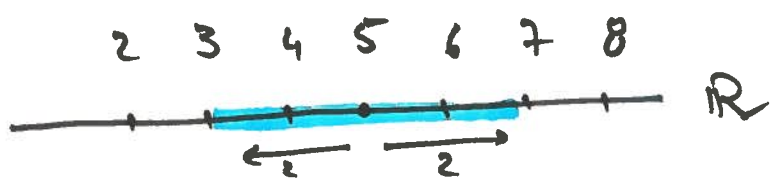
Consider: $B(a,r) = \{ x \in X \text{ s.t. } d(a,x) < r \}$

such that

$B(a,r)$ is called the ball with radius r and center a.

Ex: $X = \mathbb{R}$, $r = 2$, $a = 5$

$$B(a,r) = \{ x \in \mathbb{R} \text{ s.t. } |5-x| < 2 \}$$



$$|5-x| < 2 \iff |x-5| < 2$$

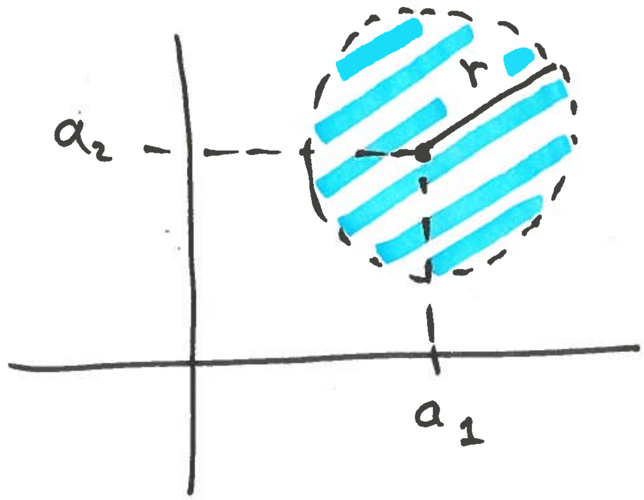
$$\iff -2 < x-5 < 2$$

$$\iff 3 < x < 7$$

$$B(5,2) = (3,7)$$

In $X = \mathbb{R}^2$ with d_{Eucl} :

$$B(a,r) = \{(x_1, x_2) \text{ s.t. } (x_1-a_1)^2 + (x_2-a_2)^2 < r^2\}$$



open disk
↳ without the boundary circle.

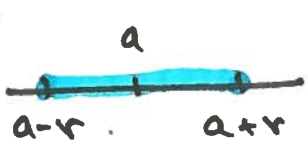
In $\mathbb{R}^3$: a full open ball in the ordinary sense:

without the boundary sphere

If we want to include the boundary, define:

$$B_c(a,r) = \{x \in X \text{ s.t. } d(a,x) \leq r\}$$

$\underline{X = \mathbb{R}}$:



$\underline{X = \mathbb{R}^2}$:

$\underline{X = \mathbb{R}^3}$:

Balls are the basic objects to define <u>metric</u> <u>topology</u>.

We need to understand them visually in small dimension, for several metrics.

$\longrightarrow$ What do balls look like for d_1, d_∞, d_0 ?

Consider d_∞ (Do d_1 as HW) in the case of

$X = \mathbb{R}^2$. Pick $a = \underbrace{(0,0)}_{0}$ and $r = 1$.

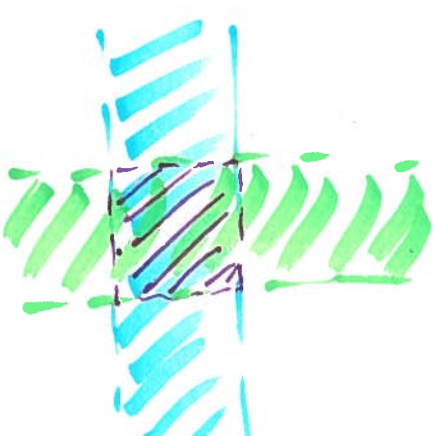
$$B(a,r) = B(0,1) = \, ?$$

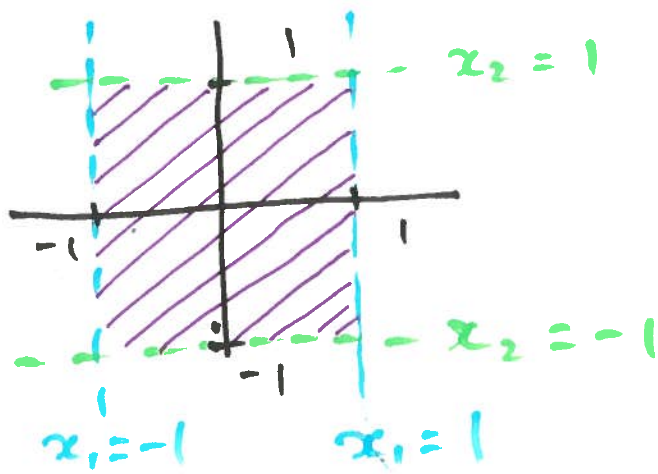
$$= \{ (x_1,x_2) \text{ s.t. } \max\{|x_1 - 0|, |x_2 - 0|\} < 1 \}$$

$$= \{ (x_1,x_2) \text{ s.t. } \max\{|x_1|, |x_2|\} < 1 \}$$

$$= \{ (x_1,x_2) \text{ s.t. } |x_1| < 1 \text{ and } |x_2| < 1 \}$$

$$= \{ (x_1,x_2) \text{ s.t. } \underline{-1 < x_1 < 1} \text{ and } \underline{-1 < x_2 < 1} \}$$





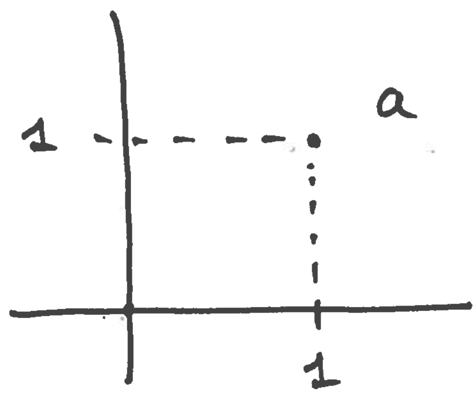
$B(0,1)$

Ex. Draw $B(0,1)$ for d_1 (Manhattan) (HW)

Another example: $X = \mathbb{R}^2$ with $d_0(x,y) = 0$ if $x = y$
$= 1$ if $x \neq y$.

Consider $a = (1,1)$.

Draw $B(a, \frac{1}{2})$, $B(a, \frac{4}{5})$, $B(a, 2)$



Let $x = (x_1, x_2)$.

- $d_0(a,x) < \frac{1}{2} \iff d_0(a,x) = 0$
$\iff a = x$

so $B(a, \frac{1}{2}) = \{a\}$

- $d_0(a,x) < \frac{4}{5} \iff d_0(a,x) = 0$
$\iff a = x$

so $B(a, \frac{4}{5}) = \{a\}$!!!

- Now, for <u>any</u> $x \in \mathbb{R}^2$,

$$d_0(a,x) = 0 \text{ or } 1 \quad < 2$$

so $d_0(a,x) < 2$ for all x in $\mathbb{R}^2$, that is,

$$\boxed{B(a,2) = \mathbb{R}^2} \quad !!$$

<u>Conclusion</u>: $B(a,r) = \{a\}$ if $r \leq 1$

$B(a,r) = \mathbb{R}^2$ if $r > 1$