Worksheet 2 # 1, 2, 3, 4, 5

Some distance functions for machine learning

DSE 220

Useful families of distance functions

1. ℓ_p norms
2. Metric spaces

Measuring distance in $\mathbb{R}^m$

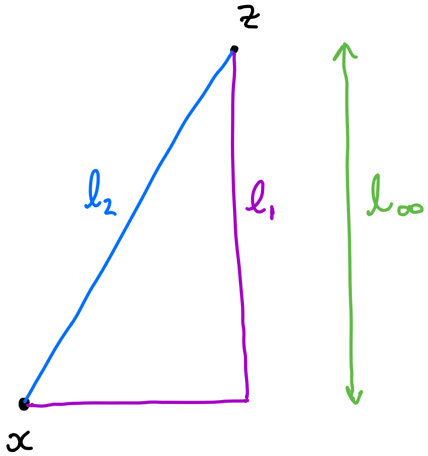
Usual choice: **Euclidean distance**:

$$\|x - z\|_2 = \sqrt{\sum_{i=1}^{m} (x_i - z_i)^2}.$$

Measuring distance in $\mathbb{R}^m$

Usual choice: **Euclidean distance**:

$$\|x - z\|_2 = \sqrt{\sum_{i=1}^{m} (x_i - z_i)^2}.$$



For $p \geq 1$, here is ℓ_p **distance**:

$$\|x - z\|_p = \left(\sum_{i=1}^{m} |x_i - z_i|^p \right)^{1/p}$$

- $p = 2$: Euclidean distance
- ℓ_1 distance: $\|x - z\|_1 = \sum_{i=1}^{m} |x_i - z_i|$
- ℓ_∞ distance: $\|x - z\|_\infty = \max_i |x_i - z_i|$

Example 1

Consider the all-ones vector $(1, 1, \ldots, 1)$ in $\mathbb{R}^d$.
What are its ℓ_2, ℓ_1, and ℓ_∞ length?

distance to origin

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \; d$$

$$\|x\|_2 = \sqrt{1^2 + 1^2 + \cdots + 1^2} = \sqrt{d}$$

$$\|x\|_1 = 1 + 1 + \cdots + 1 = d$$

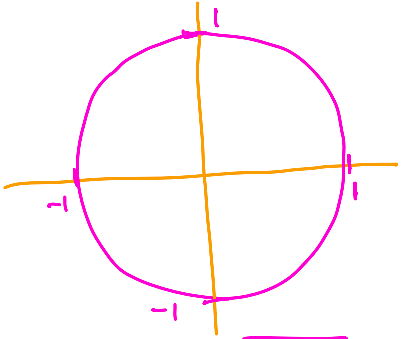
$$\|x\|_\infty = 1$$

Example 2

In $\mathbb{R}^2$, draw all points with:

1. ℓ_2 length 1 ← circle of radius 1
2. ℓ_1 length 1 ← diamond
3. ℓ_∞ length 1 ← square

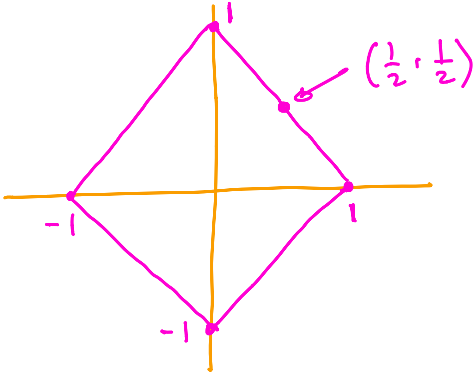
1



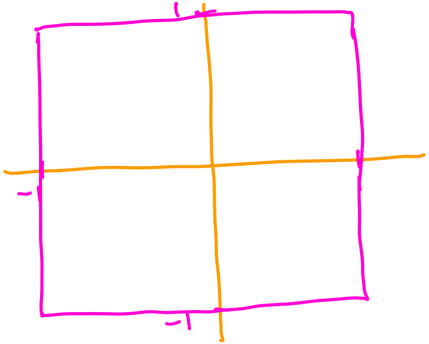
all (x,y) with $\sqrt{x^2+y^2} = 1$

ie. $x^2+y^2=1$

2



3



Metric spaces

Very useful family of distance functions

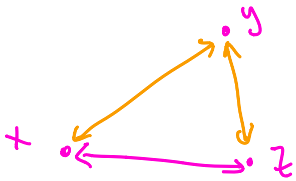
could be anything: vectors, strings, graphs, documents

Let $\mathcal{X}$ be the space in which data lie.

function $d(x, y)$

A distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **metric** if it satisfies these properties:

- $d(x, y) \geq 0$ (nonnegativity)
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)



Many of our algorithmic and statistical results hold not just for ℓ_2 distance but for any metric.

e.g. Methods for fast NN search.

Example 1

Let's look at ℓ_1 distance.

$\mathcal{X} = \mathbb{R}^m$ and $d(x, y) = \|x - y\|_p$

$$\|x - y\|_1 = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_m - y_m|$$

Check:

- $d(x, y) \geq 0$ (nonnegativity) ✓ because of absolute values
- $d(x, y) = 0$ if and only if $x = y$ ✓
- $d(x, y) = d(y, x)$ (symmetry) ✓ yes because $|a-b| = |b-a|$
- $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

We have $|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|$

sum over all i.

$\therefore$ ℓ_1 distance is a metric. [So is any ℓ_p distance, $p \geq 1$]

Example 2

$\mathcal{X} = \{$strings over some alphabet$\}$ and $d =$ edit distance

e.g. DNA sequences

Check:

- $d(x, y) \geq 0$ (nonnegativity)
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

$x =$ A C G G T

$z =$ C G G T

Edit distance between x and z is 1.

Difference between strings and vectors:

① strings can have different lengths

② vectors have numeric entries

E.g. Bag-of-words vector:

$$\begin{array}{c|c|} 1 & 6 \\ 2 & 0 \\ 3 & 1 \\ \vdots & \\ 1000 & 2 \end{array}$$

The edit distance between x and z is the minimum number of insertions, deletions, and substitutions needed to transform x into z.

$$\left.\begin{aligned} x &= \text{C G A T} \\ z &= \text{A G T} \end{aligned}\right\}$$

$d(x, z) = 2$

CGAT $\rightarrow$ AGAT $\rightarrow$ AGT

① $d(x, z) \geq 0$ ✓

② $d(x, z) = 0$ if and only if $x = z$ ✓

③ $d(x, z) = d(z, x)$ ✓ because operations are reversible

④ $d(x, z) \leq d(x, y) + d(y, z)$ ✓

$\therefore$ Edit distance is a metric.

A non-metric distance function

$d(p,q) \geq 0$ ✓

$d(p,q) = 0$ if and only if $p=q$ ✓

$d(p,q) = d(q,p)$ ✗

triangle inequality ✗

Let p, q be probability distributions on some set $\mathcal{X}$.

The **Kullback-Leibler divergence** or **relative entropy** between p, q is:

$$d(p,q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Example: $p = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right)$

$q = \left(\frac{1}{8}, \frac{1}{4}, \frac{1}{8}, \frac{1}{2}\right)$

$$d(p,q) = \frac{1}{4} \log \frac{1/4}{1/8} + \frac{1}{2} \log \frac{1/2}{1/4} + \frac{1}{8} \log \frac{1/8}{1/8} + \frac{1}{8} \log \frac{1/8}{1/2}$$

$$= \frac{1}{4} \log 2 + \frac{1}{2} \log 2 + \frac{1}{8} \log 1 + \frac{1}{8} \log 1/4$$ base 2

$$= \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 1 + \frac{1}{8} \cdot 0 + \frac{1}{8} \cdot (-2) = \boxed{\frac{1}{2}}$$