Useful distance functions for machine learning

Topics we'll cover

1. L_p norms
2. Metric spaces

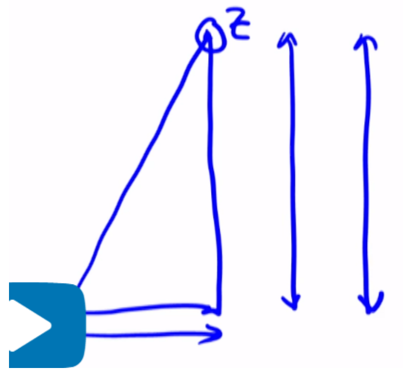
Measuring distance in $\mathbb{R}^m$

Usual choice: **Euclidean distance**:

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

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?

$\|x\|_2$

$= \sqrt{1^2 + 1^2 + \cdots + 1^2}$

$= \sqrt{d}$

$\|x\|_1 =$

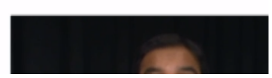
$|x_1| + \cdots + |x_d|$

$= d$

$x = (1, 1, \ldots, 1)$

0

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

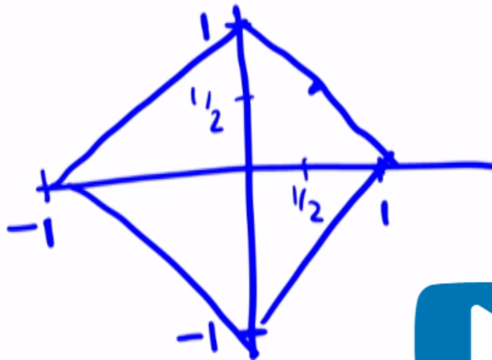


Example 2

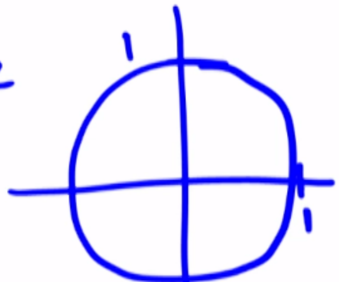
In $\mathbb{R}^2$, draw all points with:

1. ℓ_2 length 1
2. ℓ_1 length 1
3. ℓ_∞ length 1

$\ell_1 : \{(x_1, x_2) : |x_1| + |x_2| = 1\}$

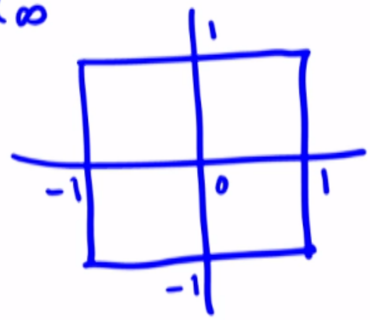


ℓ_2



$\{(x_1, x_2) : \sqrt{x_1^2 + x_2^2} = 1\}$

ℓ_∞



Metric spaces

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

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)

Example 1

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

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)

$\mathcal{X} = \{A, C, G, T\}^*$

$x = \text{A C C G T}$

$y = \text{C C G T}$

Example 2

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

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)

$d(x,y) = \#$ of insertions, deletions, substitutions to get from x to y.

$d(x,y) \geq 0$

$d(x,y) = 0 \iff x = y$

$d(x,y) = d(y,x)$

triangle inequality

Is it symmetric?

Is the number of steps to go from x to y,

the number of insertions, deletions, and substitutions

to go from x to y the same as to go from y to x?

Yes, it is, because a deletion is

the reverse operation of an insertion.

And finally, does it satisfy the triangle inequality?

It does and that's something you can also convince yourself.

A non-metric distance function

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

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

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

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

$q = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$

$d(p,q) = \frac{1}{2}\log\frac{1/2}{1/6} + \frac{1}{4}\log\frac{1/4}{1/3} + \frac{1}{8}\log\frac{1/8}{1/3} + \frac{1}{8}\log\frac{1/8}{1/6}$

Now, thankfully, it turns out,

that this can never be negative.

But that's about where the good news ends.

This is not a symmetric distance.

So the distance from p to q is in general

not the same as the difference from q to p.

And it doesn't come close to satisfying

the triangle inequality.

But, it's a distance function we use all the time.