Useful distance functions for machine learning

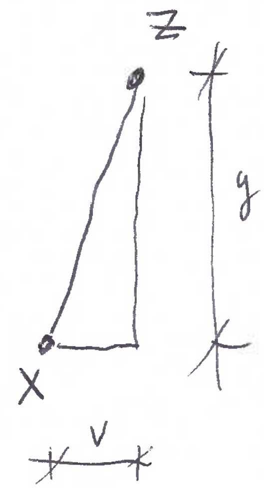
Topics we'll cover

1. L_p norms
2. Metric spaces

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

Usual choice: **Euclidean distance**:

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



difference along each coordinate

ℓ_p distance between 2 vectors x & z,

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: $\|x - z\|_1 = \sum_{i=1}^{m} |x_i - z_i|$ $v + y$
- ℓ_∞ distance: $\|x - z\|_\infty = \max_i |x_i - z_i|$ y

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

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?

ℓ_2 norm

$\|x\|_2$

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

$= \sqrt{d}$

ℓ_1 norm

$\|x\|_1 =$

$|x_1| + \ldots + |x_d|$

$= d$

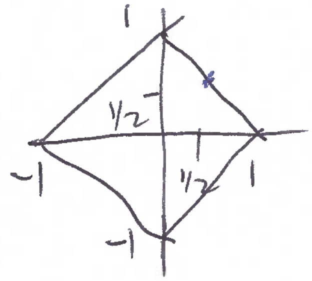
$\|x\|_\infty = 1$

Example 2

ℓ_1: $\{(x_1, x_2) : |x_1| + |x_2| = 1\}$

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

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

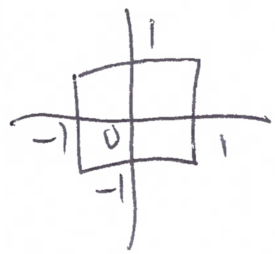


unit ball for ℓ_1

ℓ_2

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

ℓ_∞



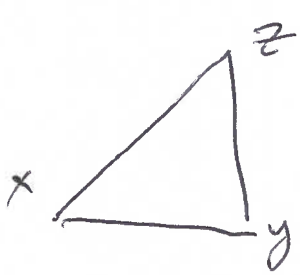
Metric spaces

$d(x, x') = 3.6$

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

A distance function (data) $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 $d(x, y) = \|x - y\|_p$

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

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) ✓

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

Sum over all i

ℓ_1 satisfies metric distance 4 properties

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)

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

$x =$ ACCGT

$y =$ CCGT

edit distance

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

$d(x,y) \geq 0$

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

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

triangle inequality

$d \to$ distance

A non-metric distance function

$d(p,q)$ distance between p & q

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

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

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

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

$q = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$ we can compute l_1, l_2 but

the choice is designed KL div

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

w/c in $d(p,q)$