Distances and similarities

A: ℓ_p norms

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?

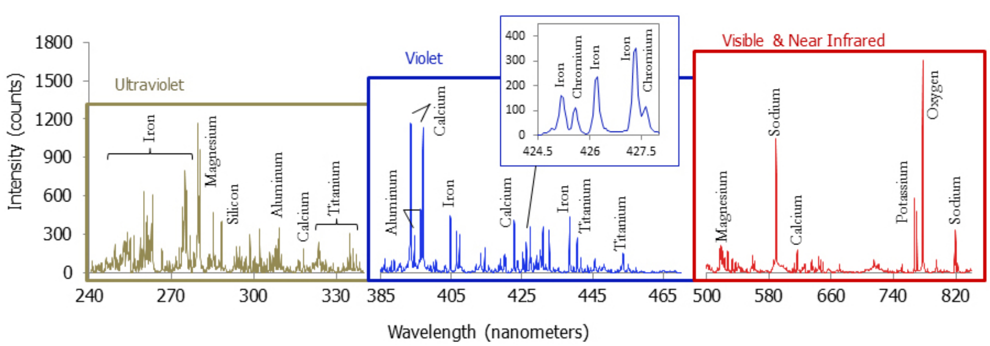
Example 2

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

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

Weighted ℓ_1 norm

For ChemCam, what if we want to emphasize/de-emphasize certain elements?



Weighted ℓ_1 norm between x and x':

$$\sum_{i=1}^{m} w_i \, |x_i - x_i'|.$$

Weighted ℓ_p norm

How would you define a weighted ℓ_p norm?

B: Metrics

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)

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)

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:

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

Other classes of distance functions

- Bregman divergences
- F-divergences
- $\vdots$

C: Similarity functions

Jaccard similarity

A notion of similarity between sets:

$$s(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Widely used in information retrieval (e.g. web search).

- In what range does this lie?
- For what B is $s(A, B)$ maximized?

Cosine similarity

A notion of similarity between vectors:

$$s(x, z) = \frac{x \cdot z}{\|x\| \|z\|}.$$

- In what range does this lie?
- How is it related to the angle between the vectors?
- For what z is $s(x, z)$ maximized?

Dot product

Even simpler than the cosine distance:

$$s(x, z) = x \cdot z.$$

- In what range does this lie?
- Can $s(x, z)$ ever be larger than $s(x, x)$?

Kernel functions

Generalization of dot products:

- Let $\mathcal{X}$ be any instance space
- We say $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel function** if

$$k(x, z) = \phi(x) \cdot \phi(z)$$

for some mapping $\phi : \mathcal{X} \to \mathbb{R}^d$, where $1 \leq d \leq \infty$.

Examples:

$$k(x, z) = (x \cdot z)^2$$
$$k(x, z) = e^{-\|x-z\|^2}$$