L7: Building Embeddings : Metric Embeddings

For a data $X \subset \mathbb{R}^d$ we know $d_E(x,p) = \|x-p\|_2 = \|x-p\|$ is a metric.

Recall a **metric** $D : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$ satisfies for any $a, b, c \in \Omega$: - $D(a,b) \geq 0$ (*non-negativity*, by definition of $\mathbb{R}_{\geq 0}$) - $D(a,b) = 0$ if and only if $a = b$ (*identity*) - $D(a,b) = D(b,a)$ (*symmetry*) - $D(a,b) \leq D(a,c) + D(c,b)$ (*triangle inequality*)

ℓ_p distance $\ell_p(a,b) = \left(\sum_{j=1}^d |a_j - b_j|^p\right)^{1/p}$
is a metric for $p \in [1, \infty)$.
- not technically defined for $p = \infty$, but in limit
$\ell_\infty(a,b) = \max_{j=1}^d |a_j - b_j|$ is a metric - ℓ_1 is the "smallest" normed metric - perhaps surprisingly, ℓ_2 may be only the second most common distance in high-dimensions!
… because of …

Cosine Distance

$$D_{\cos}(a,b) = 1 - \frac{\langle a,b\rangle}{\|a\|\|b\|} = 1 - \frac{\sum_{j=1}^d a_j b_j}{\|a\|\|b\|}$$

If $\theta_{a,b}$ is the angle between a and b (wrt origin) then $\cos(\theta_{a,b}) = \frac{\langle a,b\rangle}{\|a\|\|b\|}$.
So $D_{\cos}(a,b) = 1 \cos(\theta_{a,b})$.

$D_{\cos}(a,b) \in [0,2]$ and does not depend on magnitude of a or b.
Useful for when origin matters, but norm may not (*scale of word embeddings may depend on frequency more so that meaning*)
But not invariant to origin (like D_E is)

Is $D_{\cos}$ a metric?
No. - Does not satisfy *identity*.
But is it a pseudo-metric? No
Consider $\Omega = \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$.

- Does not satisfy *triangle inequality*
  Consider: $a = (0,1)$, $b = (1,0)$, $c = (1/\sqrt{2}, 1/\sqrt{2})$.
  Now $D_{\cos}(a,b) = 1$ and $D_{\cos}(a,c) = D_{\cos}(c,a) = (1 - 1/\sqrt{2}) \approx 0.29$
  So $D_{\cos}(a,c) + D_{\cos}(c,b) \approx 0.58 < 1 = D_{\cos}(a,b)$

Angular Distance:

$$D_{\mathsf{ang}}(a,b) = \arccos(\langle a,b\rangle) = \mathsf{radians}(\theta_{a,b})$$

Angular distance **is a metric** on $\mathbb{S}^{d-1}$.
Arclength along $\mathbb{S}^{d-1}$

Minimizing Cosine Distance

A common ML/AI task is to maximize a sum of dot-products == minimize sum of cosine distances
For pairs $(x_1, x_1'), (x_2, x_2'), \ldots$ - minimize $\sum_i D_{\cos}(x_i, x_i')$ or maximize $\sum_i \langle x_i, x_i'\rangle$
where some parameter space α controls the location of $x_1, x_1', x_2, x_2' \ldots$

If we assume $x, x' \in \mathbb{S}^{d-1}$ then

$$D_{\cos}(x,x') = 1 - \langle x, x'\rangle = \frac{1}{2}(\|x\|^2 + \|x'\|^2 - 2\langle x,x'\rangle) = \frac{1}{2}\|x - x'\|^2$$

So if we assume data is normalized (or scale irrelevant wrt 0), then
minimizing $\sum_i D_{\cos}(x_i, x_i')$ is same as minimizing sum of square errors!

Distortion-Bounded Metric Embeddings

Start with **metric space** (X, D) where - X is a domain, or sometimes a finite data set, and - D is a metric distance defined on X.

An **embedding** $(X, D) \hookrightarrow^{\rho} (Y, D')$ is both a - mapping $\phi : X \to Y$ from one domain / point set to another if $|X| = n$ finite, then each each $x_i \in X$ then $y_i = \phi(X) \in Y$.
- new metric D' so for all $x_1, x_2 \in X$ has **distortion** ρ

$$\frac{1}{\rho} \le \frac{D(x_1, x_2)}{D'(\phi(x_1), \phi(x_2))} \le \rho$$

If $\rho > 0$, for $x_1 \ne x_2 \in X$, can we have $\phi(x_1) = \phi(x_i)$?
No, then divide by 0, and the $\le \rho$ is not bounded.

Embeddings in ℓ_∞

Recall $\ell_\infty(a, b) = \max_{j=1}^{d} \|a_j - b_j\|$

Theorem: Every metric space $(X, D) \hookrightarrow^1 \ell_\infty^n$ (no distortion!)

$\phi : X \to \mathbb{R}^d$
$\phi(x_i) = (D(x_1, x_i), D(x_2, x_i), \ldots, D(x_d, x_i))$

Since D is a metric, satisfies triangle inequality.
$D(x_k, x_i) - D(x_k, x_j) \le D(x_i, x_j)$. So

$$\max_k |D(x_k, x_i) - D(x_k, x_j)| \le D(x_i, x_j)$$

$$\|\phi(x_i) - \phi(x_j)\|_\infty \le D(x_i, x_j)$$

But also, jth coordinate of $\phi(x_i) - \phi(x_j) = D(x_j, x_i) - D(x_j, x_j) = D(x_i, x_j)$
So

$$\|\phi(x_i) - \phi(x_j)\|_\infty \ge D(x_i, x_j)$$

Hence $\|\phi(x_i) - \phi(x_j)\|_\infty = D(x_i, x_j)$

Embeddings in ℓ_p

We consider (X, D), summarized as ℓ_p^d, where - $X \subset \mathbb{R}^d$ of size n, and - $D = \ell_p$

Bourgain [1985] has the following famous result:
For any metric space (X, D), for $p \in [0, \infty)$ assumed a constant, we have

$$(X, D) \hookrightarrow^{O(\log n)} \ell_p^{O(\log^2 n)}$$

- tight for $p = 2$, original target dimension was exponential in n - Matousek [1996]: $(X, D) \hookrightarrow^{O(\frac{\log n}{p})} \ell_p^{O(\log^2 n)}$

Some special cases can do better!

- $\ell_2 \hookrightarrow^1 \ell_1^{\binom{n}{2}}$ on n points
- $\ell_1 \hookrightarrow^{O(\sqrt{\log n} \log \log n)} \ell_2$ on n points
  (distortion lower bound is $\Omega(\sqrt{\log n})$)

Edit Distance

Edit distance measures between two strings how many substitutions are needed to get from one to the other.
Is a metric
Closely associated with dynamic programming.

It is a "counting" measure, so most naturally associated with ℓ_1.
Requires $\Omega(\log n)$ distortion to embed into ℓ_1.