last time, we started exploring the idea of clustering, a form of dim. reduction.

(write down our for-now defn, b/c might be hard to read on handout)

- looked @ functions for describing how good a clustering is, and explored relationships between them.

- but we didn't talk about how you get a good clustering (i.e., didn't describe any algorithms).

~~There's a good reason for this~~

Note that we're talking about a huge search space

<see "Stirling numbers of the second kind": bounded below by $\frac{1}{2}(k^2+k+2)k^{n-k-1} - 1$, according to Wikipedia>

~~Hatzivassi~~ Hatzivassiloglou; McKeown '93: $n=21$, $k=9$, 1.23×10^{14}. >

... altho' that ~~doesn't im~~ fact by itself doesn't imply that clustering is hard.

<sorting also has a huge search space>.

nb: Srebo et al, "when is clustering hard?"

Today, two algs.

<follow handout 2nd page>

- (1) seems to be $O(n^2)$ just to compute the criterion
- note that (2) requires being able to compute centroids — if your data isn't given as points in $\mathbb{R}^d$, might not be able to do this. Example: only given point-to-point distances.

→ visualization in D3; www.naftaliharris.com/blog/visualizing-k-means-clustering

should say about the alg: r_j^i is not necessarily r_j for C_j!! Maybe $\hat{r}_j^i$ would have been better notation.

→ start w/ packed circles.

- explain ~~colors = closest~~, ● = centroid, colors = ~~closest~~ pts whose closest centroid is that color's.

(needed someone to 'put finger' on ● r_j^{i-1})

reassign

(in class, accidentally picked an r_j^0 that became "empty". oops!)

~~defer~~ smiley ~~to later~~??

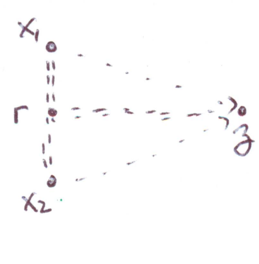
(*) specifically: can say k-means opt fn inspired by k, but it is actually:

$$\min_{\substack{C_1,\dots,C_k \\ \hat{r}_1,\dots,\hat{r}_k}} \sum_j n_j \sum_{x \in C_j} \|x - \hat{r}_j\|_2^2 .$$ so $\hat{r}_j$ may not always be centroid;

- now: pf of (8). rewrite: $\left(\sum_{x \in C_j} \|x - r_j\|_2^2\right) + n_j \|r_j - z\|_2^2 = \sum_{x \in C_j} \|x - z\|_2^2$ for arbitrary z.

last time: note that it seems like this is a violation of $\Delta \neq$:

<clicker results>



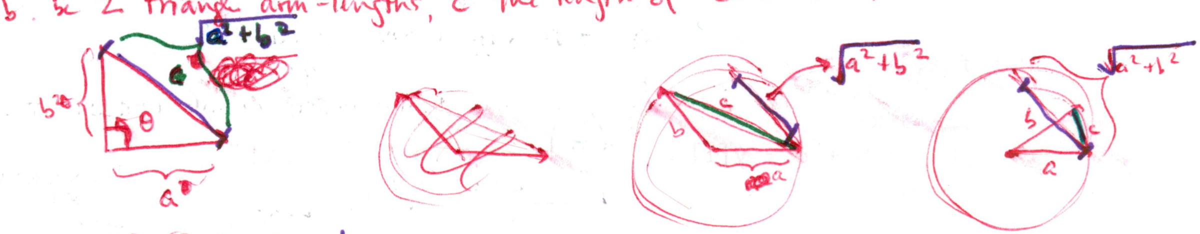
that circle around the exponent was meaningful!

we resolve the paradox by noting that this is the Euclidean distance squared.

$\| \ \|_2^2$ → square! not 1. $\Delta \not\leq$ doesn't apply:

<again, sometimes the little things matter.

let a, b be 2 triangle arm-lengths, c the length of "connector".



here a^2+b^2 = |—|

for right angles, $c = \sqrt{a^2+b^2}$

$a^2+b^2 < c^2$ $\qquad$ $a^2+b^2 > c^2$.

so now that we @ least don't think ~~the lemma~~ that (8) is false a priori, why don't we go ahead & prove it?

$$\sum_{x \in C_j} \|x - z\|_2^2$$ need to introduce r_j. let's brute-force it.

$$= \sum_{x \in C_j} \|\underline{x - r_j} + \underline{r_j - z}\|_2^2$$

$\left(\|a - b\|_2^2 = \sum_i a[i]^2 - 2\sum_i a[i]b[i] + \sum_i b[i]^2 \right.$ (two vectors)

$= \|a\|_2^2 - 2\sum_i a[i]b[i] + \|b\|_2^2$

$= \|a\|_2^2 - 2\,a \cdot b + \|b\|_2^2$

$$= \left(\sum_{x \in C_j} \|x - r_j\|_2^2 \right) - 2\sum_{x \in C_j} (x - r_j) \cdot (r_j - z) + \sum_{x \in C_j} \|r_j - z\|^2$$

$$\sum_{x \in C_j} \|r_j - z\|^2 \to n_j \|r_j - z\|_2^2$$

second term in lemma

$$-2\sum_{x \in C_j} (x - r_j) \cdot (r_j - z) \to -2(r_j - z) \cdot \underbrace{\sum_{x \in C_j} (x - r_j)}_{=\left(\sum_{x \in C_j} x\right) - n_j r_j}$$

$$n_j r_j = n_j \frac{\sum_{x' \in C_j} x'}{n_j}$$

- corollary: the z that minimizes ~~the LHS~~ either LHS / RHS is the centroid

making it a good choice for r_j

• and, again, we assert w/out proof that this lemma ~~can~~ yields that (1) ≡ (2). See Hopcroft & Kannan (link on webpg) for pf.

→ each $\|x-\hat{r}_j\|$ is being reduced, where $\hat{r}_j$'s are fixed.

So: considering RHS:

⟨k-means⟩ step 1 reduces distance to the "representative".

step 2 ~~minimizes w/in cluster~~ minimizes per/cluster terms.

~~[are we playing fast and loose w/ defn of r?]~~ ?.

b/c we're choosing the best $\hat{r}_j$'s for the ~~given~~ fixed C_j's.

• convergence given by ~~# of possible~~ improvements in LHS having to be in discrete increments (if we assume no cycling btwn clusterings).

(this alternating minimization)

- does it optimize? no (NP-hard problem).

↓

partitioning problems are "pretty much always" NP-complete.

[illegible] k-means.

-- except when they aren't: lead-in to single-link clustering: an optimal alg. (for a certain clustering criterion)

• note: "best friend" criterion ~~turns out to not~~ is not the single-link criterion.

↑ from last time

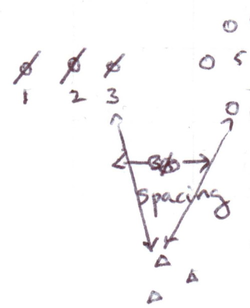
(~~probable hw~~ - probable hw exercise! Interesting to understand these subtleties.

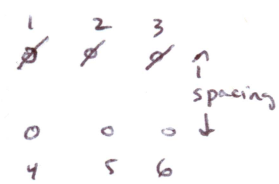
~~cons~~

instead: take inter-cluster spacing (front of handout).

~~instead~~

$\max_{j,j'}$ the minimum of $d(C_j, C_{j'}) = \min_{x\in C_j,\, x'\in C_{j'}} \|x-x'\|_2^2$





note that ~~this w/~~ not all pairwise intra-cluster distances are small.

- idea of single-linkage: "knock out" the minimum 2-cluster spacing

⇒ the max increases.

but what about optimality?