last time:

- introduced K-means clustering:

- greedy alternating optimization algorithm: finds local minimum

- there's more detail on the lecture notes for last time:

but think of the "approximate" alternative opt fn:

\[ \min \sum_{j} w_j \sum_{x \in C_j} \|x - \hat{r}_j\|^2 \]

- two sets of parameters:

- each step keys one set fixed, and chooses best other set of params.

- introduced single-linkage:

- greedy optimization alg for maximizing: the spacing between clusters of a clustering

- "closest approach"

- spacing \((\triangle, \circ, \square)\):

- \[ \min \left\{ \text{spacing} (\triangle, \circ), \text{spacing} (\circ, \square), \text{spacing} (\triangle, \square) \right\} \]

- spacing \((\triangle, \circ, \square)\):

- \[ \frac{\text{spacing (\triangle, \circ), spacing (\circ, \square), spacing (\triangle, \square)}}{3} \]

- underline = cluster of these shapes

- = closest "point of approach"

- alternate spacing clustering: (top vs. bottom)

- ask: (4) or (5)?

- spacing \((X, \text{not } X)\) = closest pt of approach.
Claim every single
observation.

Single-link: add an edge (join the clusters of) the two (different cluster)
points that are closest.

\[ \Rightarrow \text{spacing new clustering's spacing > old clustering's.} \]

b/c you just eliminated one of the places where two clusters
were closest together.

that's why we have a greedy alg.

\[ \text{< for sake of argument, should we assume all inter-point distances are diff? >} \]

\[ \overset{\text{so you can see we have a greedy algorithm.}}{\text{But, how do we know if this leads to the best clustering overall?}} \]

After all, a common pitfall of greedy algo's is they make some
choice early on that "blocks off" a better solution later on.

Claim: let \( C^* \) be the \( k \)-clustering by single-link, let \( C \) be any alternate.

\[ \text{spacing}(C^*) \geq \text{spacing}(C) \]

Pf:

**Observation:** every inter-cluster edge in \( C^* \) has length less than \( \text{spacing}(C^*) \).

which is, after all, a potential edge.

So if this hadn't been true, we would have picked that edge.

\[ \text{if } C \neq C^* \text{, then's points } X_i, X_j \text{ in the same cluster in } C^* \text{ but not } C. \]

\[ \overset{\text{C boundary}}{\text{- C edges,}} \]

\[ \overset{\text{C* edges,}}{\text{a+b}} \]

\[ b = 1^{st} \text{ time leaving } X_i \text{'s cluster.} \]

\[ \|a-b\|_2^2 \leq \text{spacing}(C^*) \]

by our observation.

\[ \|a-b\|_2^2 \geq \text{spacing}(C) \]

put these 2 in= together, and what do we get?