Lecture 11, part I: single-link optimality 3/5/15

last time:

- ~~introduced~~ studied 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" ~~eq~~ ~~alt~~ alternative opt fn:

$$\min_{\substack{c_1,\dots,c_k \\ \hat{r}_1,\dots,\hat{r}_k}} \sum_j n_j \sum_{x \in c_j} \|x - \hat{r}_j\|_2^2$$

two sets of parameters;

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

- introduced single-linkage:

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

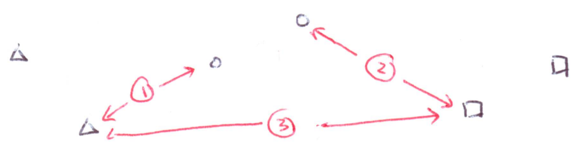
= "~~point of~~ closest approach"

~~spacing(△,○,□) = min{spacing(△,○), spacing(○,□), spacing(△,□)}~~

$$\text{spacing}(\underline{\triangle}, \underline{\circ}, \underline{\square}) = \min\{sp(\underline{\triangle}, \underline{\circ}), sp(\underline{\circ}, \underline{\square}), sp(\underline{\triangle}, \underline{\square})\}$$

①, ②, ③

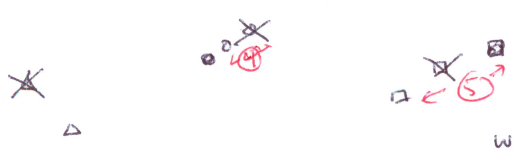
underline = cluster of those shapes



= ①, closest "point of approach".

alternate ~~spacing~~ clustering: <top vs. bottom>

ask: ④ or ⑤ ?



$$\text{spacing}(\underline{X}, \underline{\text{not } X}) = ④$$

~~bad b/c~~ worse ~~...~~ closest pt of approach.

~~claim: every single~~

~~observation:~~

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

=> ~~spacing~~ new clustering's spacing $\geq$ old clustering's.

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

~~that's why we have a greedy alg.~~

• ≪ for sake of argument, should we assume all inter-point distances are diff? ≫

• so you can see we have a greedy algorithm.

But, how do we know if this leads to the best clustering overall?

After all, a common pitfall of greedy algs 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.

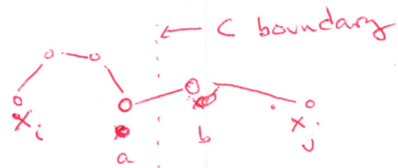
$\text{spacing}(C^*) \geq \text{spacing}(C)$

pf:

observation: every inter-cluster edge in C^* has length less than spacing(C^*).

which is, after all, a potential edge. so if this hadn't been true, we would have picked that edge.

if $C \neq C^*$, there's points x_i, x_j in the same cluster in C^* but not C.



— = C^* edges.

~~a = 1st~~

b = 1st time leaving x_i's cluster.

$\|a-b\|_2^2 \leq \text{spacing}(C^*)$, by our observation.

which?

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

put these 2 ineq together, and what do we get?