CS4811
March 15, 2017
wednesday



# Clustering

trying to group things not knowing what the groups are.

n-tuples "points"

↳ the features can be numeric or not.

distance between points or entities to be clustered.

↓

depends on the problem.

## K-means clustering

↳ the number of clusters

n data points

- select k (random) seeds (centers)
- assign all the rest of the points to clusters
- compute new centroids.

[until clusters stop changing] n-k

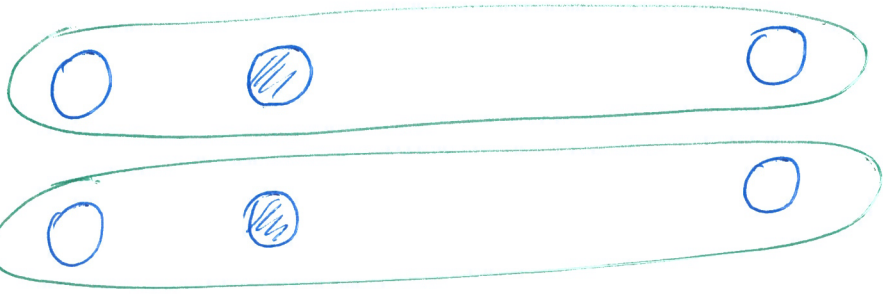
so that each point to be clustered exactly once.

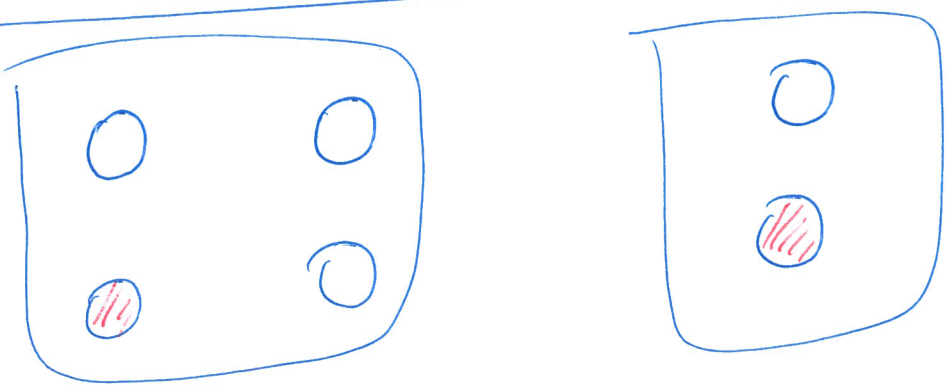
time complexity: n points
$O(n)$

- how to initialize the k-seeds?
- what will k be?
  iteratively generate clusters incrementing k until the change in "quality" of the clustering is not significant.
- Could there be a situation where a point is clustered

idk

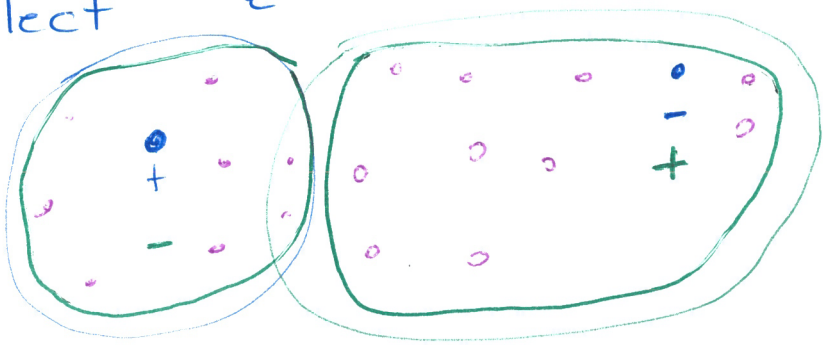
- The order we process the points effects clustering.





## Cluster/2 Algorithm

- select k seeds



provide descriptions so that all the other seeds or negative

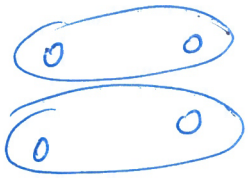
- classify all object with respect to the description
- minimize or tighten the description. (to remove overlap)
- adjust more to remove overlaps
- iterate: to get more "satisfactory" clusters
-

k - clusters

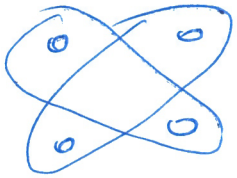
n points

how many possible clusters can I get?

- n



$c\binom{4}{2}$



$p_1$ $p_2$ $p_3$ - - - $p_n$

k k - - - k

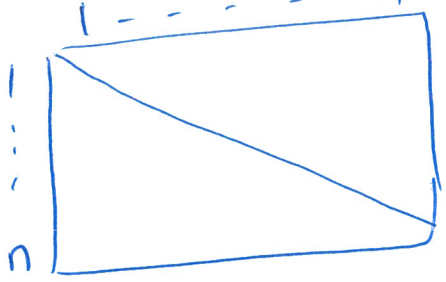
$k^n$

k

$c_1$ $c_2$ ... $c_k$

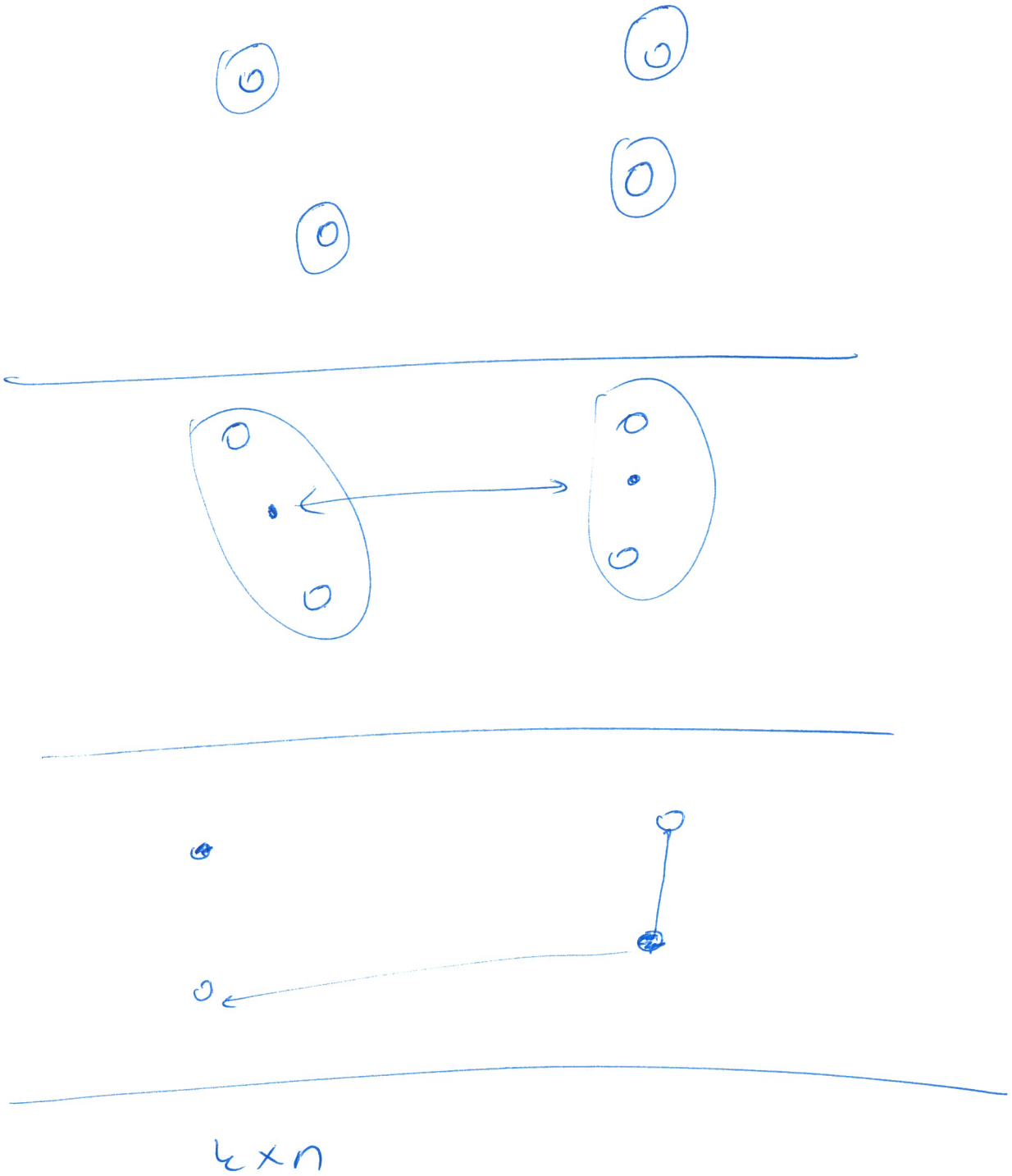
## Hierarchical Agglomerative clustering (HAC)

n points




similarity

OO



4 x n