clustering algorithms

how to find a shape where there is no shape

what is clustering?

- grouping objects that are similar each other into classes
- method for data exploration

example – news articles

- consider some vocabulary $V = \{v_1, \cdots, v_d\}$
- map each news article to a vector $(x_1, \cdots, x_d)$ where $x_i = 1$ if v_i appears in the article and $x_i = 0$ if v_i does not appear in the article
- the articles are points of a d dimensional space
- articles with similar sets of points (near in the space) correspond to similar topics

a possible definition of distance

- Euclidean distance between two points a and b with coordinates $a_1 \cdots a_d$ and $b_1 \cdots b_d$
  - d is the number of dimensions

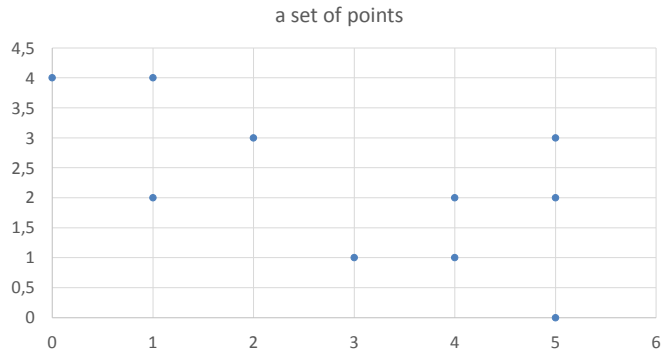
$$dist(a,b) = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2}$$

k-means algorithm

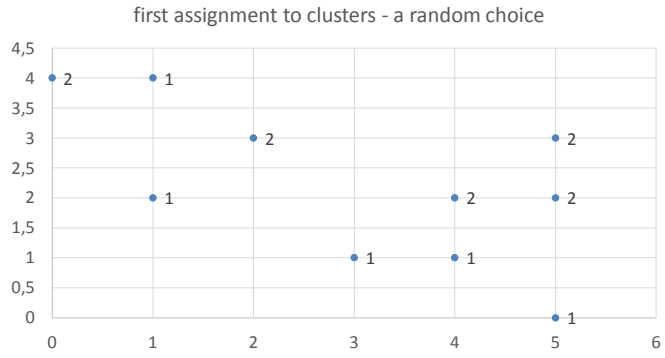
- input: n points in some d–dimensional space and a number k

1. randomly place k points into the space; each point represents the centriod of a cluster
2. assign each point to the cluster that has the closest centroid
3. recompute the positions of the k centroids
4. repeat steps 2 and 3 until a stopping criterion is met

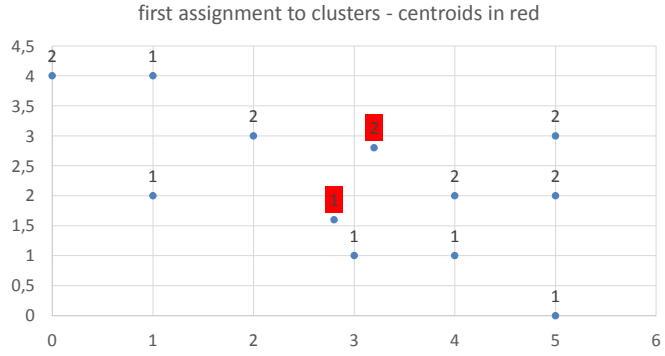
clustering 10 points in 2 clusters ($k = 2$)



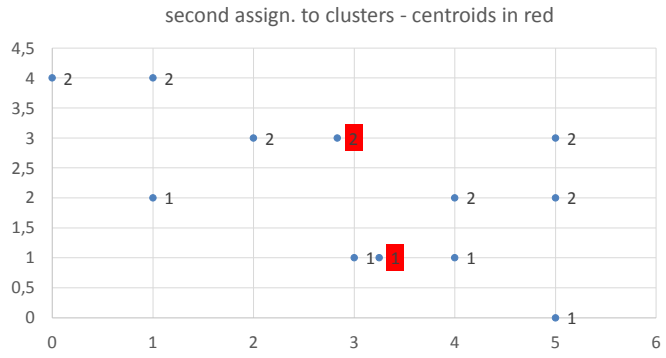
clustering 10 points in 2 clusters ($k = 2$)



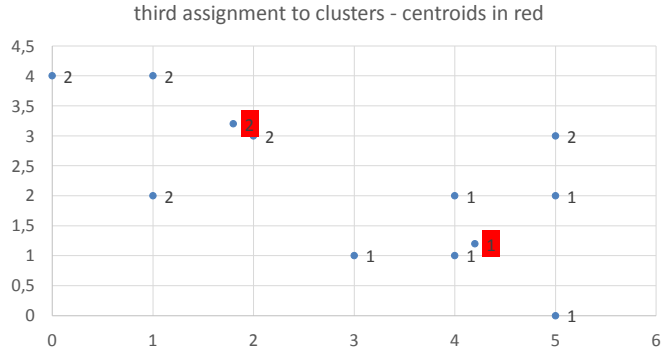
clustering 10 points in 2 clusters ($k = 2$)



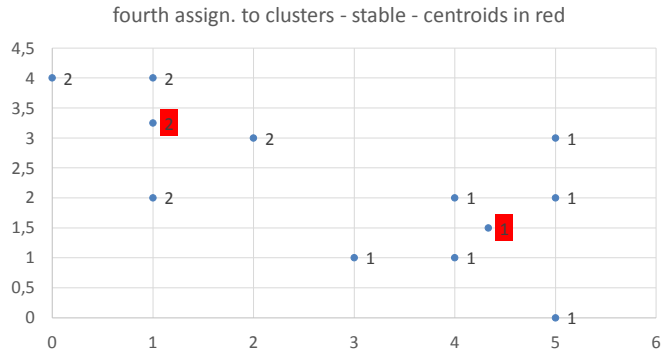
clustering 10 points in 2 clusters ($k = 2$)



clustering 10 points in 2 clusters ($k = 2$)



clustering 10 points in 2 clusters ($k = 2$)



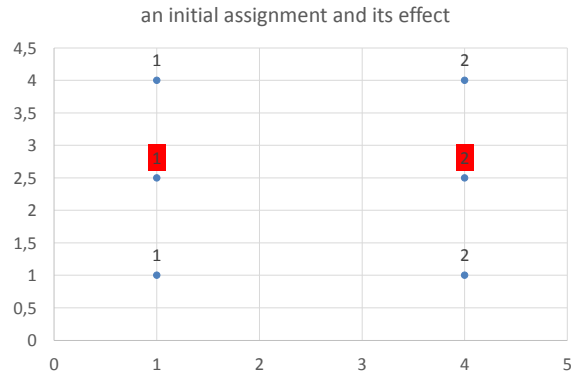
does it take too much time?

- if the number of iterations is t the algorithm spends $O(nkt)$ time

cons

- necessity of specifying k
- sensitive to noise and outlier data points
  - outliers: a small number of such points can substantially influence the mean value
- clusters are sensitive to initial assignment of centroids
  - clusters can be inconsistent from one run to another

sensitivity to initial assignment



hierarchical clustering

- start giving to each point its own specific cluster
- repeat
  - merge the two closest clusters
- until (e.g.) the number of clusters is «small»

hierarchical clustering