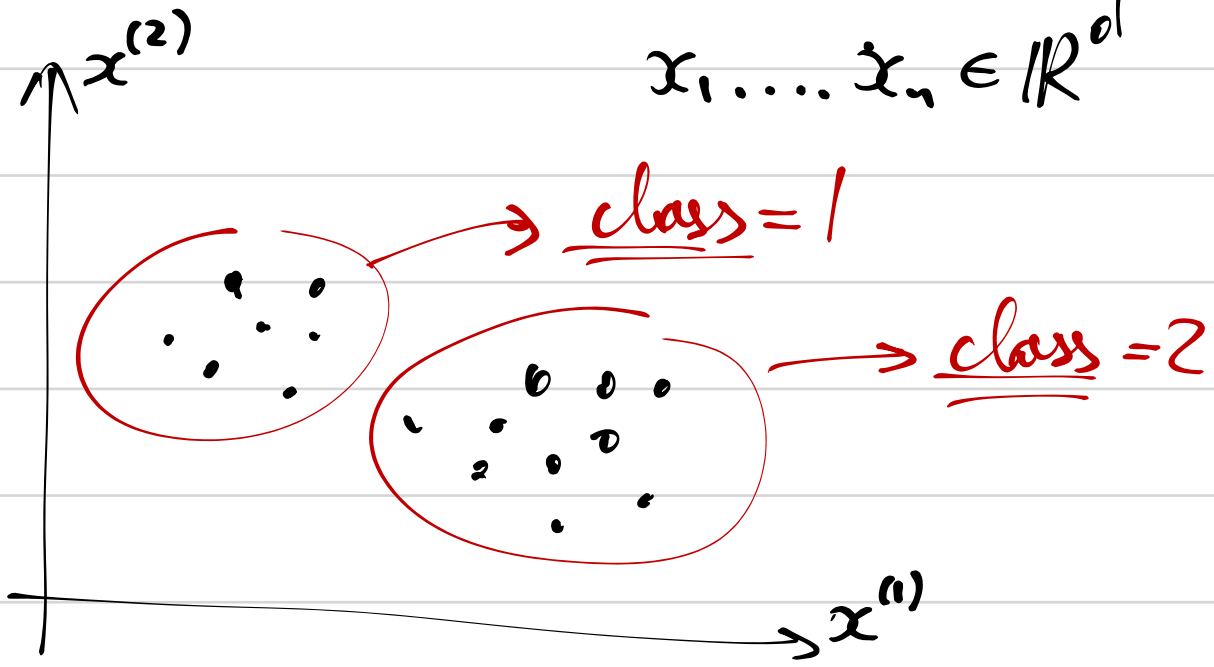
Hi !!

Clusterization methods

① k-means clustering

② hierarchical clustering.

The problem



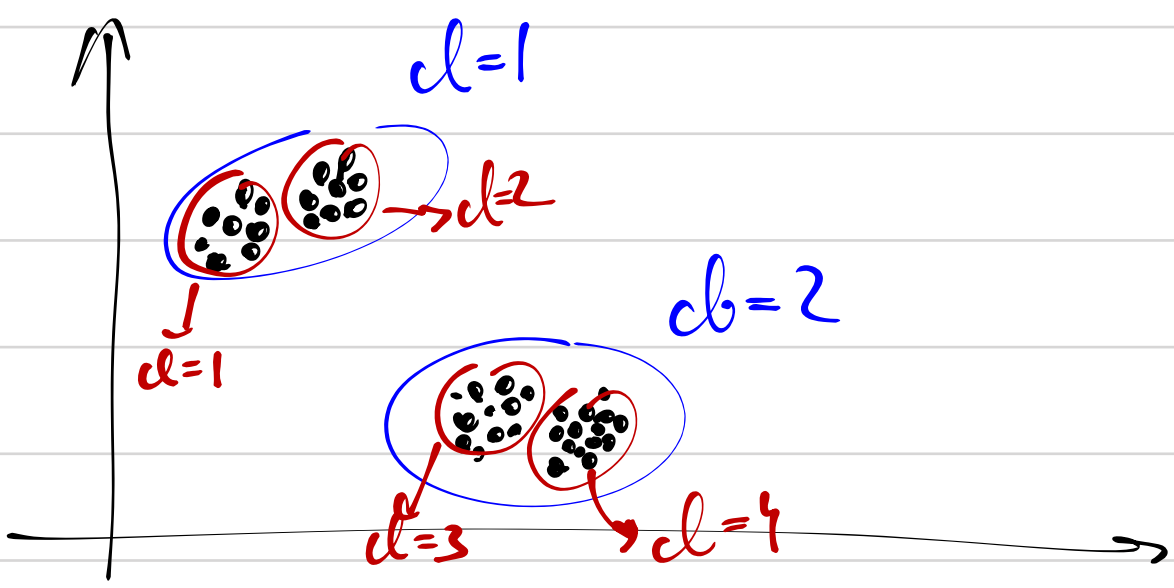
Do not observe the class!

training set

$-x_1-$
$-x_2-$
$\vdots$
$-x_n-$

goal: create artificial class label.

problem: we have no true class



$-x_1-$
$-x_2-$
$\vdots$
$-x_n-$

goal

creak

y_1
$\vdots$
y_n

$y_i \in \{1, 2, \dots k\}$

k-means clustering.

Ideal target.

$\underline{k}$ [number of clusters] is fixed

$x_1, x_2 \dots x_n$

goal: create a partition $S_1 \cup S_2 \dots S_k$ of the set $\{1, 2, 3, 4, \dots n\}$

S_j - id. of observations in the cluster n^j.

$n=5$
$k=2$

$S_1 = \{1, 2, 4\}$ $S_2 = \{3, 5\}$

$\{1, 2, 3, 4, 5\}$

$S_i \cap S_j = \emptyset$ if $i \neq j$

$S_1 \cup S_2 \cup S_3 \cup \dots \cup S_k = \{1, 2, 3, 4 \dots n\}$

$\min_{S_1 \dots S_k}$ WCSS

Within Cluster Sum of Squares.

(v) $$WCSS = \sum_{i=1}^{n} \| x_i - c(x_i) \|^2$$

$c(x_i)$ - center of the cluster to which x_i belongs

$$\mu_j = \frac{\sum_{x \in S_j} x}{card(S_j)}$$ ← center of the cluster n^j.

(v) $$WCSS = \sum_{j=1}^{k} \sum_{x \in S_j} \| x - \mu_j \|^2$$

$$\sum_{x \in S_j} \| x - \mu_j \|^2 =$$

μ_j ← center of the cluster S_j

(pythagorean theorem)

$$= \frac{1}{2 \cdot \operatorname{card} S_j} \cdot \sum_{x,y \in S_j} \| x - y \|^2$$

$$\min_{S_1 \dots S_k} \sum_{j=1}^{k} \frac{1}{\operatorname{card} S_j} \cdot \sum_{x,y \in S_j} \| x - y \|^2$$

---

this optimization is not (yet?) possible on modern computers.

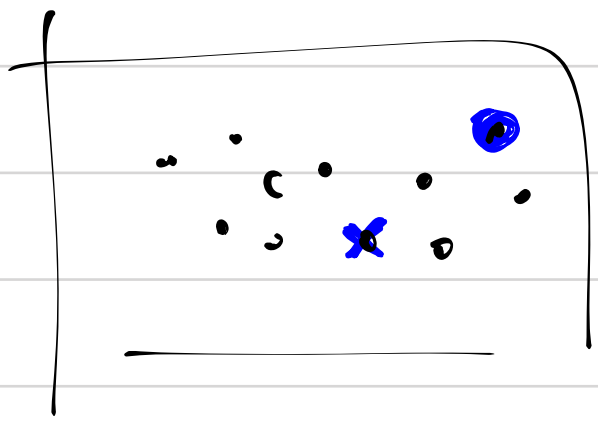
$n = 1000 \qquad k = 2$

$$\frac{2^n}{2} = 2^{999}$$

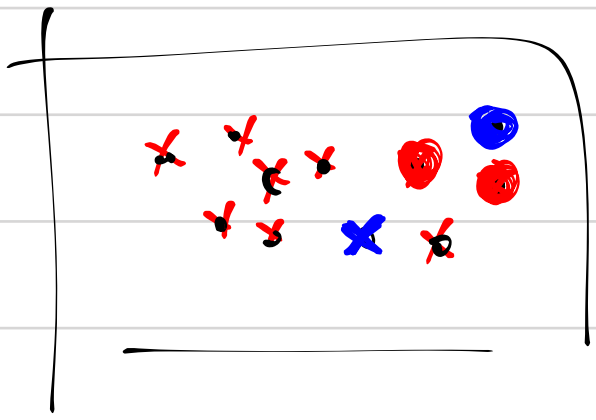
it's not possible to check this huge amount of combinations.

Heuristic algorithm. "naive k-means"

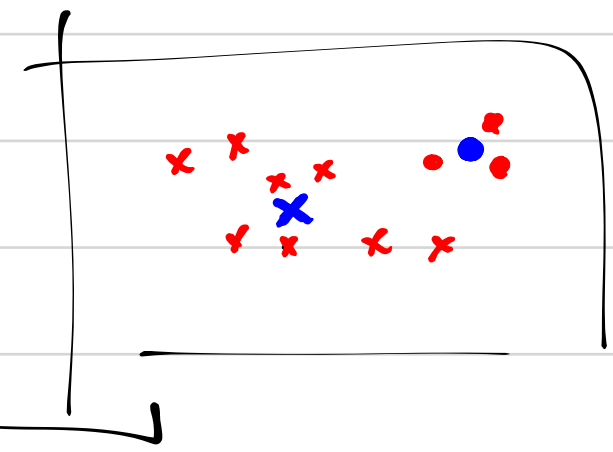
Step 1. Choose k points among our n points randomly as preliminary cluster centers.



→ Step 2. Preliminary assign each point to the closest center.



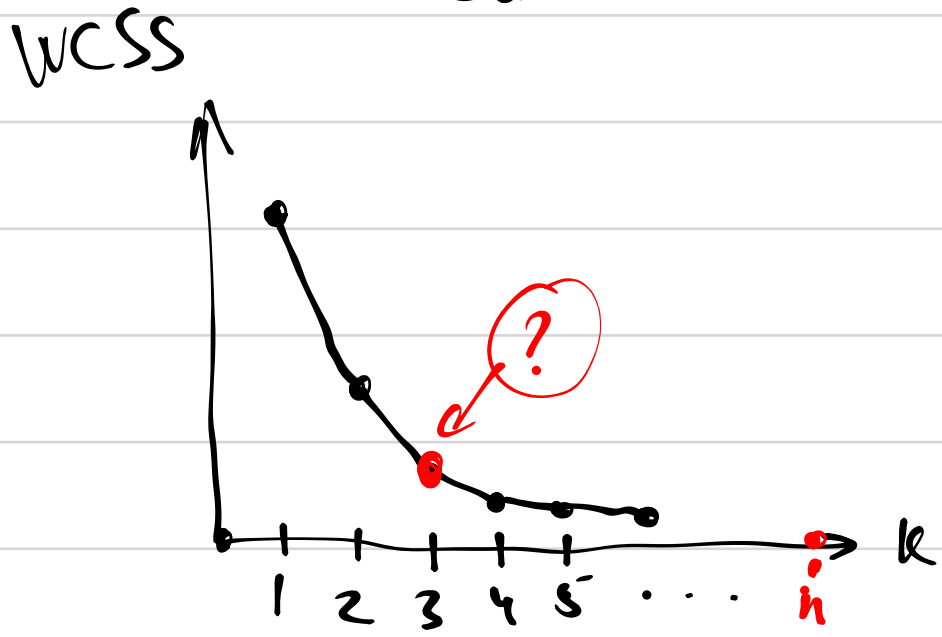
Step 3. Preliminary update cluster centers.



no guarantee that it will converge at all !!

it works well in general

"Elbow method" (graphical approach) to select k.



high decrease of WCSS | low decrease of WCSS

opt k

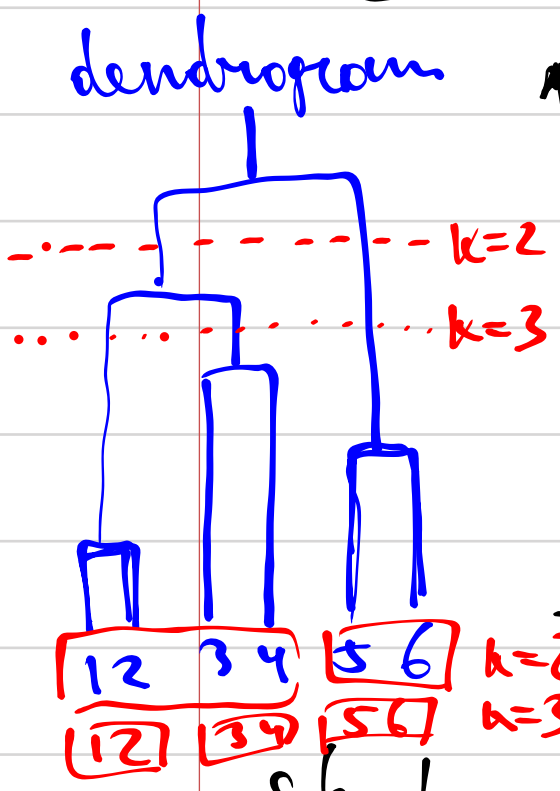
If n = k then WCSS = 0

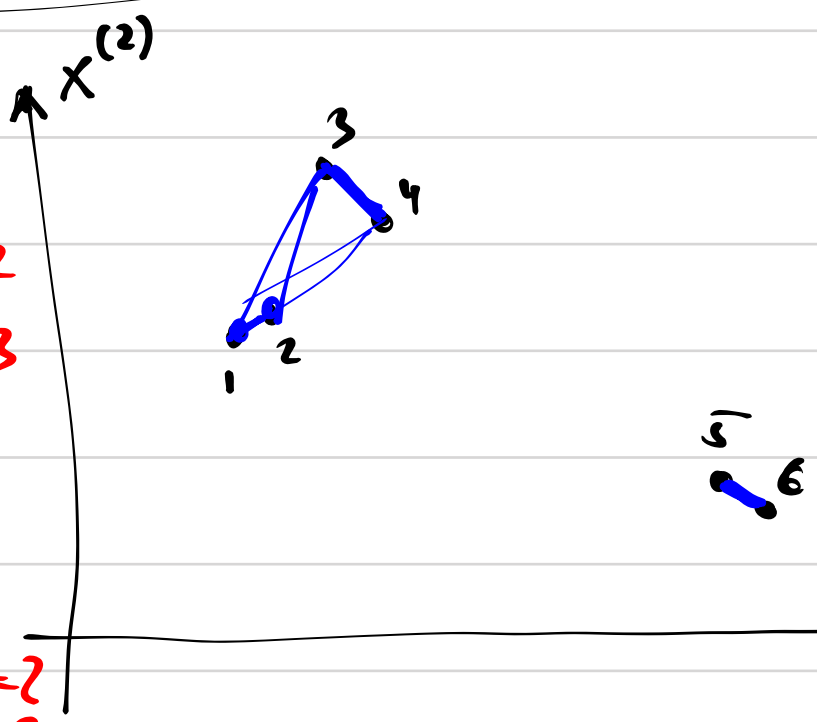
Important point: Do not forget to (scale) standardize the variables !

② hierarchical clustering.

① Scale all the variables.

dendrogram





$X_1 \dots X_n \in \mathbb{R}^d$

1. (1 — 2)
2. (5 — 6)
3. (3 — 4)
4. (1-2) ? (3-4)
5. (1-2-3-4) ? (5-6)

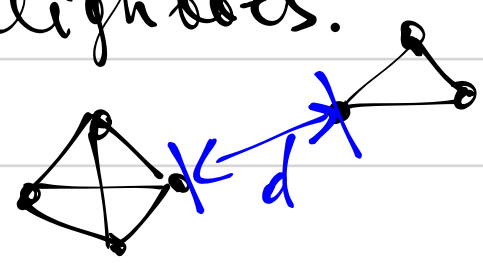
Step 1 Assign each point to its own cluster.

→ Step 2. Find two closest clusteres and join them.

repeat step 2 until all points are in one big cluster

→ How to calculate distance between clusteres?

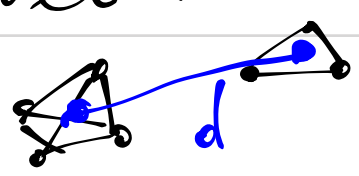
* take the distance between closest neighbors. (most popular)



* take the average distance between neighbors.

$$d(S_i, S_j) = \frac{\sum_{x \in S_i, y \in S_j} d(x, y)}{(\text{card } S_i) \cdot (\text{card } S_j)}$$

* take the distance between centeres of clusteres.



...