Lecture 19:

Module 4: Unsupervised learning:

- There is no label (in contrast with supervised ML)

Data: $x_1, x_2, \dots, x_n$

Goal: Discover informative patterns, structures, subgroups, etc within data.

In this course, we'll consider two specific instances of unsupervised learning:

- clustering
- Dimensionality Reduction

Clustering: Partition data into distinct subgroups such that data points within each group are similar to each other and points from different groups are "dissimilar".

$x_i \in \mathbb{R}^p$

"Similarity" will be quantified by a distance function:

$x \bullet \xleftarrow{\quad d(x,y) \quad} \bullet y$

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \in \mathbb{R}^p$

Example:

(1) $d(x,y) = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{1/2} = \|x-y\|_2$

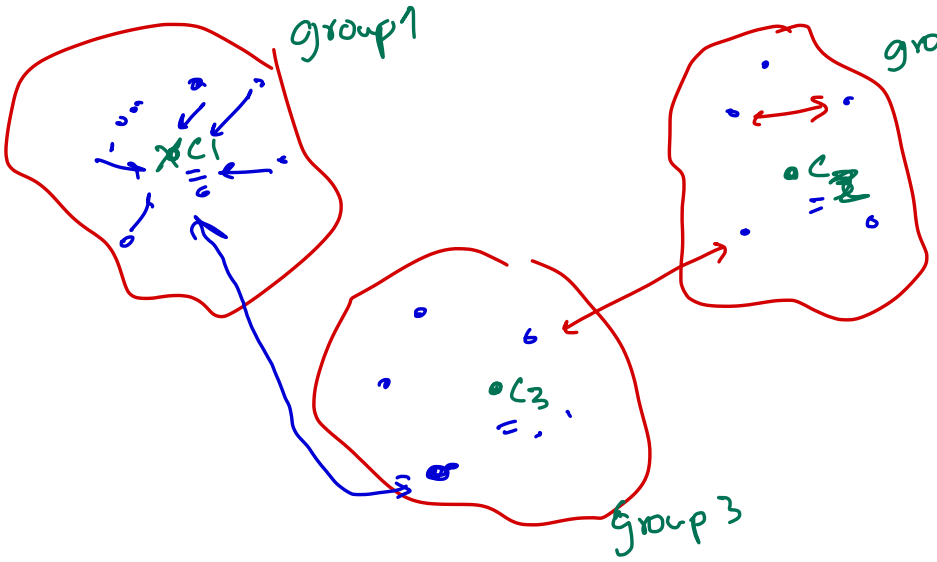
- the Euclidean distance
- the L_2 distance

(2) $d(x,y) = \sum_{i=1}^{p} |x_i - y_i| = \|x-y\|_1$

- L_1 distance
- the Manhattan distance

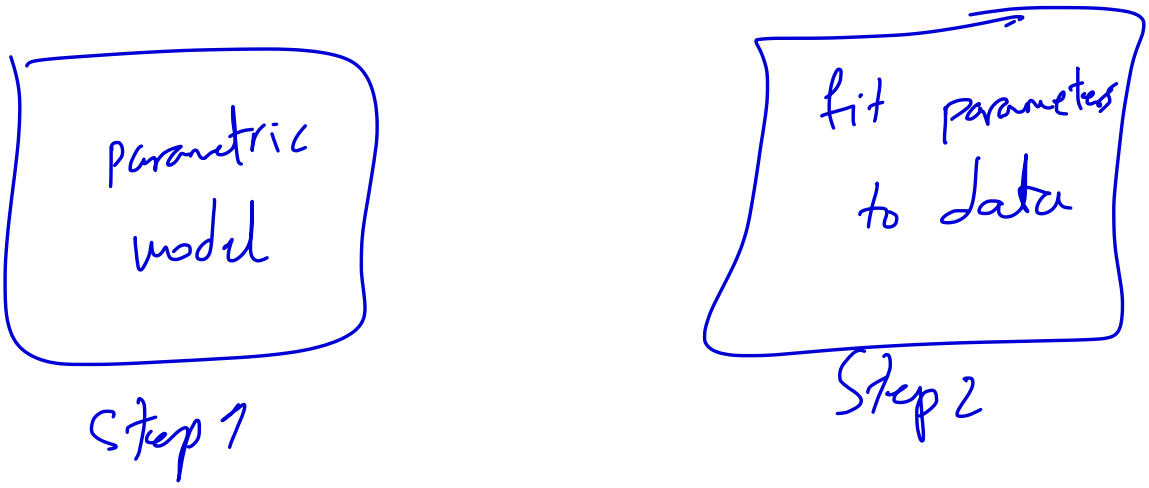
distance ↓ Similarity ↑

Clustering: partition data into K subgroups (K is given) such that points in each group are similar (i.e smaller distance) and points in different groups are dissimilar (relatively larger distance).



$K = 3$

remember our ~~opr~~ general approach:



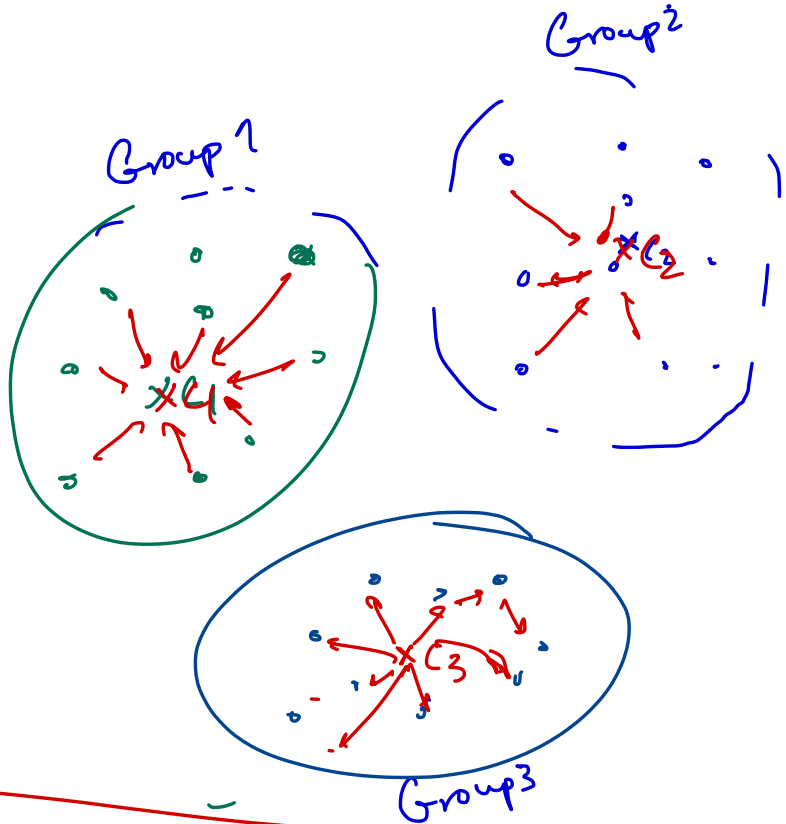
Step 1 (parametric modeling):

How do we represent a group?

- Assign to group i a "center", $c_i \in \mathbb{R}^p$, such that points in group i are close to the center c_i and points outside group i and far from c_i.

$\Rightarrow$ parameters of the clustering problem are $= c_1, c_2, \dots, c_k \in \mathbb{R}^p$

step 2:



$k = 3$

c_1, c_2, c_3

k-means objective $\leftarrow$

$$\min_{c_1, \dots, c_k} \sum_{j=1}^{k} \sum_{i \in \text{Group } j} \| x_i - c_j \|^2 \quad \text{(k-means)}$$

$\| x_i - c_j \|^2$ = internal distance of Group j

k-means clustering: solve (k-means) and find $c_1, \dots, c_k$ along with the groups.

- Difficulty: Finding the exact solution of (k-means) is very difficult (it's an NP-hard problem)

- we'll describe an algorithm that finds an approximate (but sub-optimal) solution for the k-means problem which works well in practice!

The k-means Algorithm:

the k-means algorithm is based on two simple principles:

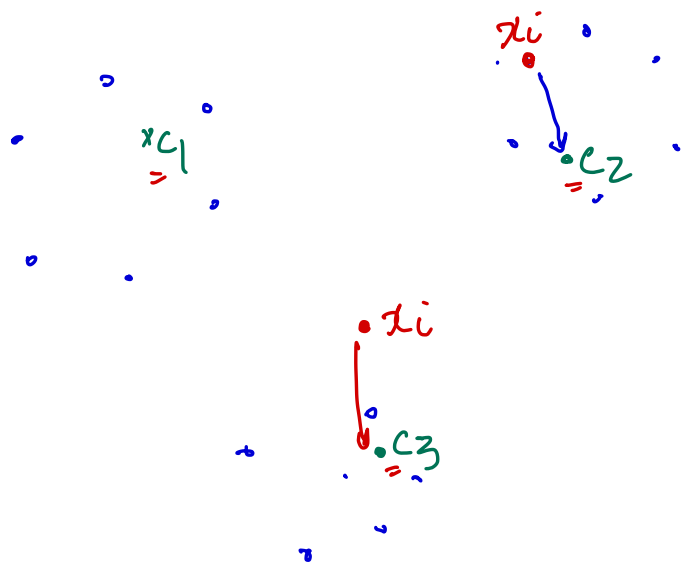
$$\sum_{j=1}^{K} \sum_{x_i \in \text{Group}_j} \|x_i - c_j\|^2$$

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

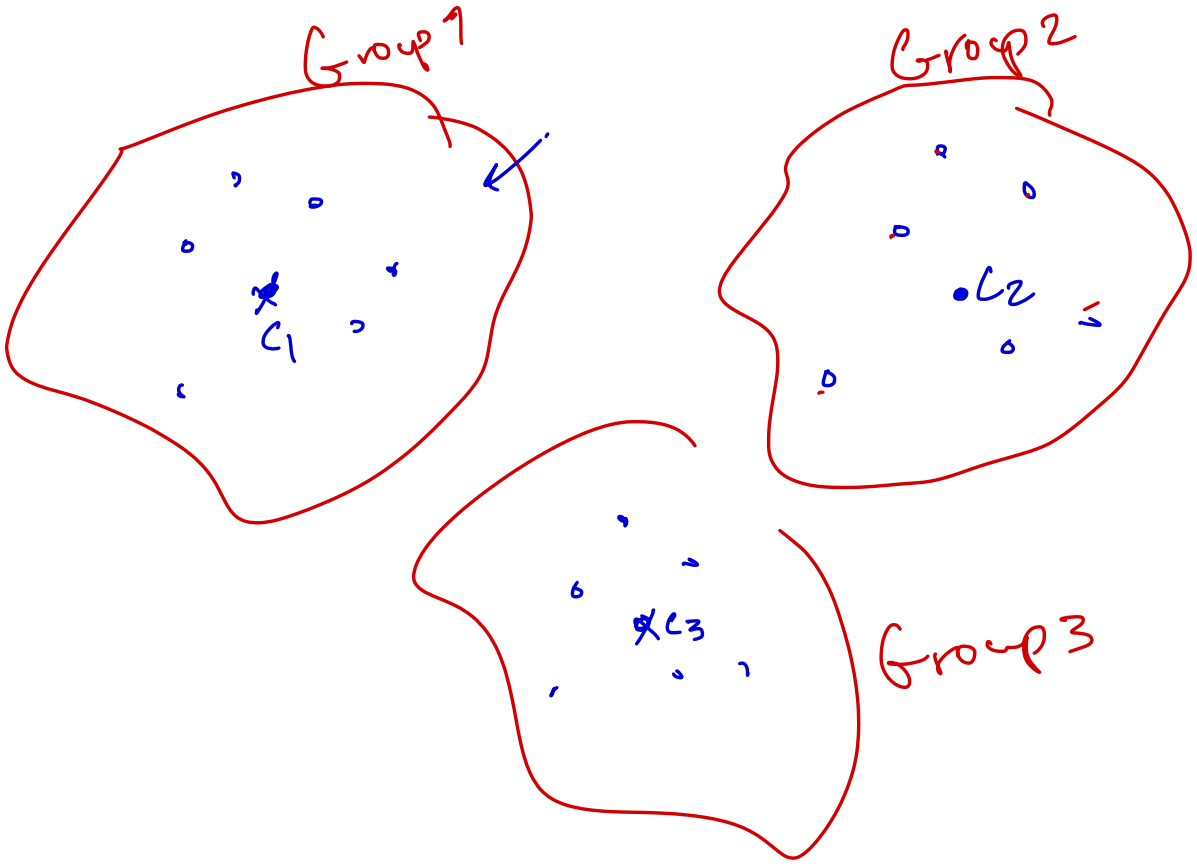
$c(x_i)$ is the center of the group that x_i belongs to.

Principle 1: Given centers $c_1, \dots, c_k$, the center (group) that we assign to data point x_i is the one that's closest to x_i

$$c(x_i) = \underset{c_j \in \{c_1, \dots, c_k\}}{\operatorname{argmin}} \|x_i - c_j\|$$



Principle 2: Let's assume that we fix the groups:



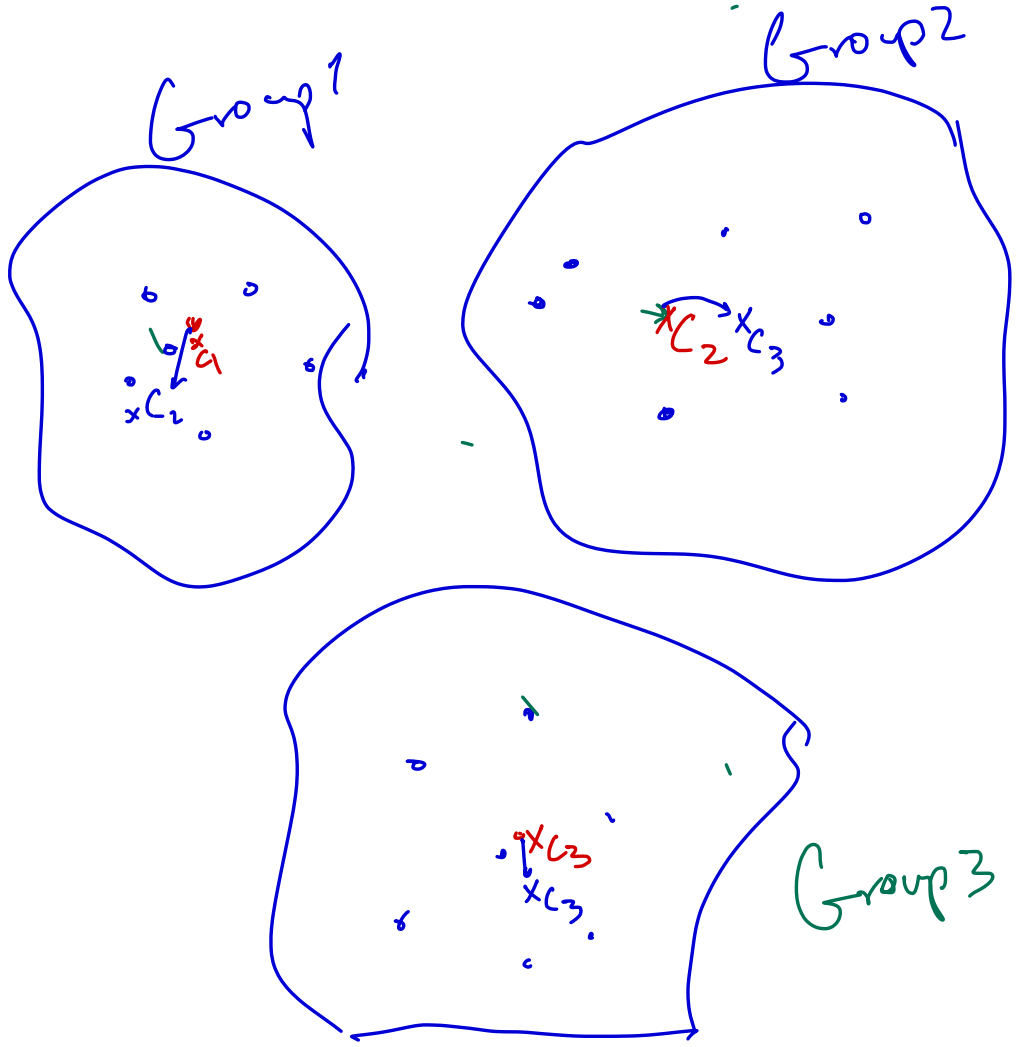
$$c_j = \frac{\sum_{x_i \in \text{Group } j} x_i}{N_j}$$

N_j = # points in group j

Let's try to solve the 1-means problem

$$\min_{c_1 \in \mathbb{R}^p} \sum_{i=1}^{n} \|x_i - c_1\|_2^2$$

$$\rightarrow c_1 = \frac{\sum x_i}{n}$$



k-means algorithm:

- choose $\{c_1^0, c_2^0, \ldots, c_k^0\}$ according to some (arbitrary) choice.

- For $t = 1, 2, 3, \ldots$

  (1) Define Group j to be the set of all the data points that are closest to c_j^{t-1}. → (principle 1)

(2) update $c_j^t = \frac{\sum_{x_i \in \text{Group } j} x_i}{N_j}$ → (principle 2)

(3) if $c_j^t = c_j^{t-1}$ for all $j \in \{1 \to k\}$

we will stop!

Some important (practical) points about the k-means algorithm:

- Let

$L_t \triangleq L(\{c_1^t, \ldots, c_k^t\}) = \sum_{i=1}^{n} \|x_i - c^t(x_i)\|^2$

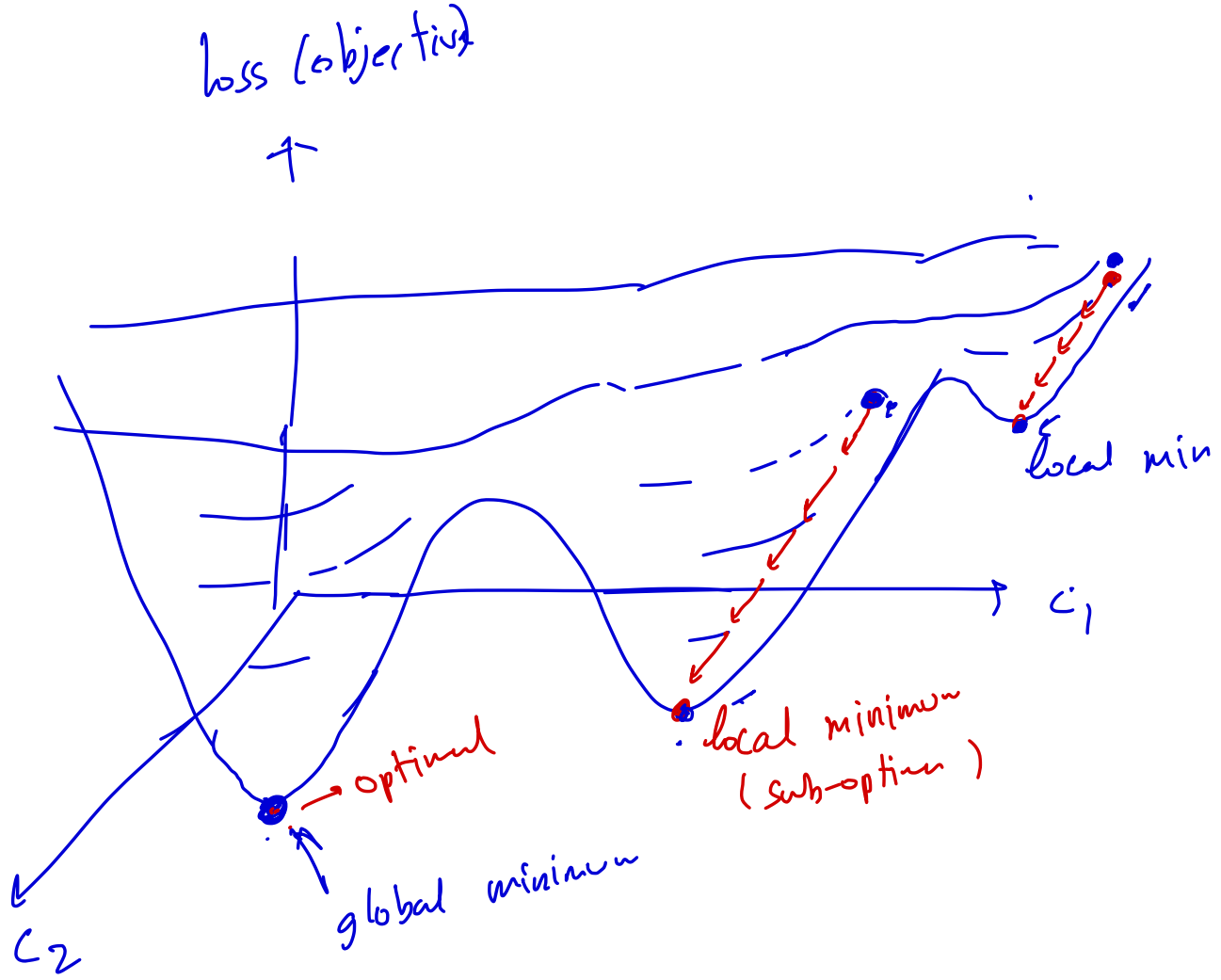
objective computed for $\{c_1^t, \ldots, c_k^t\}$

- then it can be proven that

$$L_t \leq L_{t-1}$$

- Also, it can be proven that

after a "finite" number of iterations, the k-means algorith will stop (converge.)



- The quality of the final set of centers that we obtain from the k-means algorithm depends on the initial centers.
- In practive, we often choose 10-20-50 (random) different initializations and run the k-means algorithm for each initialization, and choose the set of

centers that has the best loss.