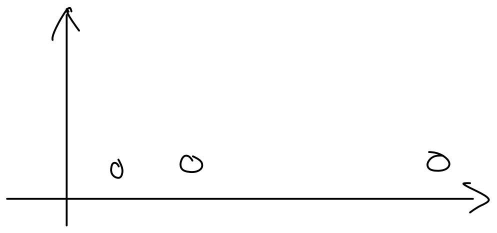
K-MEANS

Main Idea:

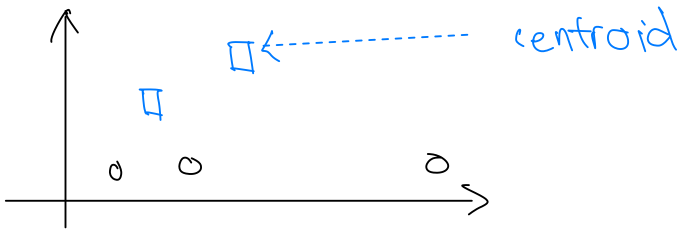
Separating data points into K clusters by alternatively set the assignments and centroids.

For example:

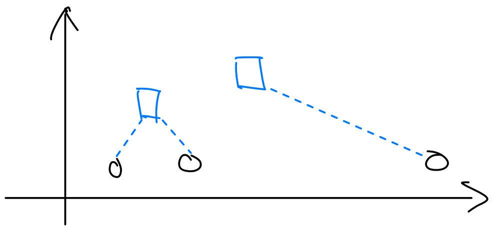
- Given some data points:



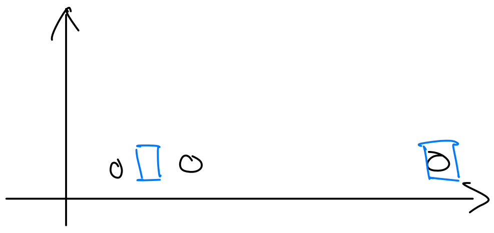
- Initialize random centroids:



- Iterate through these steps:
  - Assignment: Assign each point to the nearest centroid



  - Centroid: Update position of centroids (mean of assignments)



K-means algorithm

- Input: dataset $\{x_i\}_{i=1}^{n}$ and a number K of clusters
- Algorithm:
  - Initialization: Randomly placed K centroids
  - Iterate till converge:

    i) Assignment: For each x_i, assign it to closet centroid

$$z_i = \underset{k=1,..,K}{\text{argmin}} \| x_i - \mu_k \|_2$$

centroid index assigned to x_i

i^{th} data point

k^{th} centroid

    ii) Centroids: For each centroid, update based on new assignments

$$\mu_k = \frac{1}{|S_k|} \sum_{i \in S_k} x_i \quad , \text{where } S_k = \{ i : z_i = k \}$$

centroid of k^{th} cluster

k^{th} cluster

K-means as Optimization

K-means can be viewed as a "coordinate descent" algorithm for optimizing a objective function of centroids and assignments.

Problem definition:

- Given : $\{x_i\}_{i=1}^n$
- Define objective function:

$$L(\mu, z) = \sum_{i=1}^n \| x_i - \mu_{z_i} \|^2$$

centroid assigned to x_i

- We try to minimize that function:

$$\boxed{\min_{\mu, z} L(\mu, z)}$$

Tricky : ◦ Mixed optimization (not convex optimization)

Because : ◦ μ is continuous
◦ z is discrete

$\Rightarrow$ Many local optima

Solution to mixed optimization problem:

Use coordinate descent to find the local optima

- Coordinate descent:
  - Initialize μ_0
  - Repeat: (at iteration t^{th})
    i) Update z, with fixed μ

    vector $z^t = \arg\min_z L(\mu^t, z)$

    ii) Update μ, with fixed z

    vector $\mu^{t+1} = \arg\min_\mu L(\mu, z^t)$

actually K-means algorithm (Proof next page)

- To find the global optima, run many coordinate descent with different initialization values and return the best optima value

Proof coordinate descent on $\mathcal{L}(\mu, z)$ is actually K-means algorithm

- We know that: $\quad L(\mu, z) = \sum^{n} \| x_i - \mu_{z_i} \|^2$
- Coordinate descent says that:
  - Update z, fixed μ

$$\underbrace{z^t}_{\text{vector}} = \underset{z}{\arg\min}\ L(\mu^t, z)$$

$$= \underset{z}{\arg\min} \sum^{n} \| x_i - \mu_{z_i} \|^2$$

scalar

$$\Rightarrow z_i = \underset{z_i}{\arg\min} \| x_i - \mu_{z_i} \|^2$$

This is <u>assignment step</u> in K-means

  - Update μ, fixed z

$$\underbrace{\mu^{t+1}}_{\text{vector}} = \underset{\mu}{\arg\min}\ L(\mu, z^t)$$

$$= \underset{\mu}{\arg\min} \sum^{n} \| x_i - \mu_{z_i} \|^2$$

Scalar

$$\Rightarrow \mu_k = \underset{\mu_k}{\arg\min} \sum_{i \in S_k} \| x_i - \mu_k \|^2 \quad , \text{ where } S_k = \{ i : z_i = k \}$$

$$= \arg\min \Big[\underbrace{\sum \|x_i\|^2}_{\text{const}} - \underbrace{2\left(\sum x_i\right)\mu_k + |S_k| \cdot \|\mu_k\|^2}_{} \Big]$$

$$\nabla_\mu f(\mu_k) = 0$$

$$\boxed{\Rightarrow \mu_k = \frac{1}{|S_k|} \sum_{i \in S_k} x_i}$$

This is <u>centroids step</u> in K-means

- <u>Conclusion</u>:

Coordinate descent on $\mathcal{L}(\mu, z) \iff$ K-means algorithm