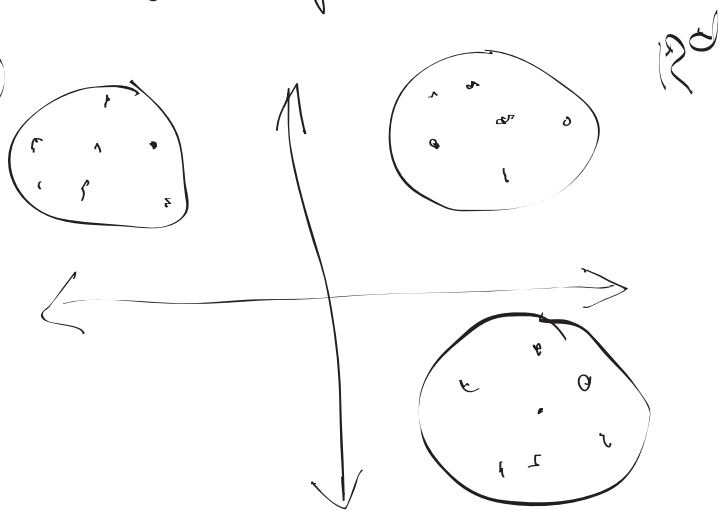
K means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$



Assume K clusters with "centers" (centroids)

$\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$

$$\sum_{j=1}^{k} \sum_{\substack{i \text{ such} \\ \text{that} \\ d_i \text{ is assigned} \\ \text{to cluster } j}} \text{distance}(d_i, \mu_j)$$

$$= \sum_{j=1}^{k} \sum_{\substack{i \text{ such} \\ \text{that} \\ d_i \text{ is assigned} \\ \text{to cluster } j}} \| d_i - \mu_j \|^2 = \mathcal{L}$$

k means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$

Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$

$$\sum_{j=1}^{k} \sum_{\substack{i \text{ such} \\ \text{that} \\ d_i \text{ is assigned} \\ \text{to cluster } j}} \text{distance}(d_i, \mu_j)$$

$$= \sum_{j=1}^{k} \sum_{\substack{i \text{ such} \\ \text{that} \\ d_i \text{ is assigned} \\ \text{to cluster } j}} \| d_i - \mu_j \|^2 = \mathcal{L} =$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \cdot \| d_i - \mu_j \|^2 \quad \text{where}$$

$$a_{ij} = \begin{cases} 1 & \text{if } d_i \text{ is assigned to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

K means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$

Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ (unknown)

$$\mathcal{L} = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \cdot \| d_i - \mu_j \|^2 \quad \text{where}$$

$$a_{ij} = \begin{cases} 1 & \text{if } d_i \text{ is assigned to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

If we were given the μ_j, $1 \le j \le k$,

it's easy to compute $\mathcal{L}$

e.g. $\| d_i - \mu_j \|^2$

$$= \sqrt{(d_{i1} - \mu_{j1})^2 + \cdots + (d_{id} - \mu_{jd})^2}$$

K means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$

Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ (unknown)

$$L = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \cdot \| d_i - \mu_j \|^2$$

Algorithm

0. "Arbitrarily" select centroids $\mu_1 \ldots \mu_k$

Assignment Step

1. Assign each d_i to the closest centroid

Update Step

2. Recompute centroids $\mu_1 \ldots \mu_k$

Repeat until the change in L is sufficiently small.

K means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$

Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ (unknown)

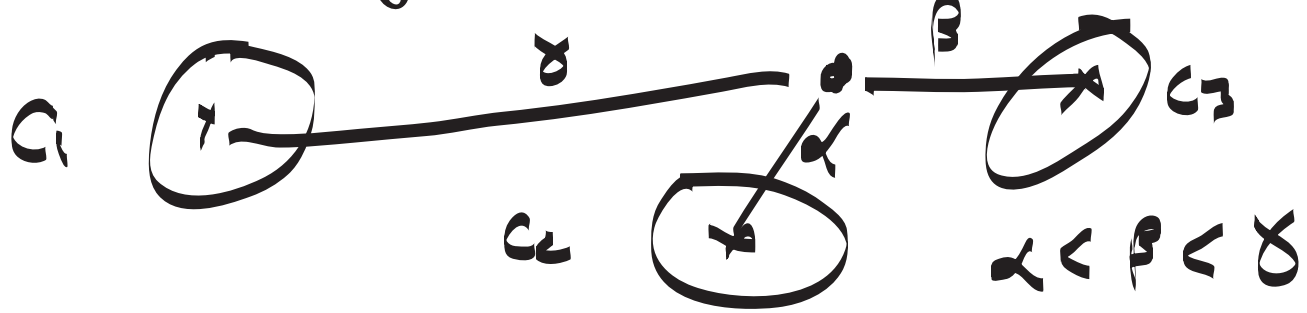
$$\mathcal{L} = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \cdot \| d_i - \mu_j \|^2$$

Assignment Step

① Assign each d_i to the closest centroid

$$C_j = \left\{ d_i : \| d_i, \mu_j \|^2 \leq \| d_i, \mu_p \|^2, \; 1 \leq p \leq k \right\}$$

So C_j is all the data points assigned to the j^{th} cluster.



K means Problem

Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$

Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ (unknown)

$$\mathcal{L} = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \cdot \| d_i - \mu_j \|^2$$

Update Step

(2) Recompute centroids $\mu_1 \ldots \mu_k$

Updated $\mu_j = \frac{1}{|C_j|} \cdot \sum_{d_j \in C_j} d_j$

assigned to μ_j

$d_1 = (4, 8, 16)$

$d_2 = (3, 9, 2)$

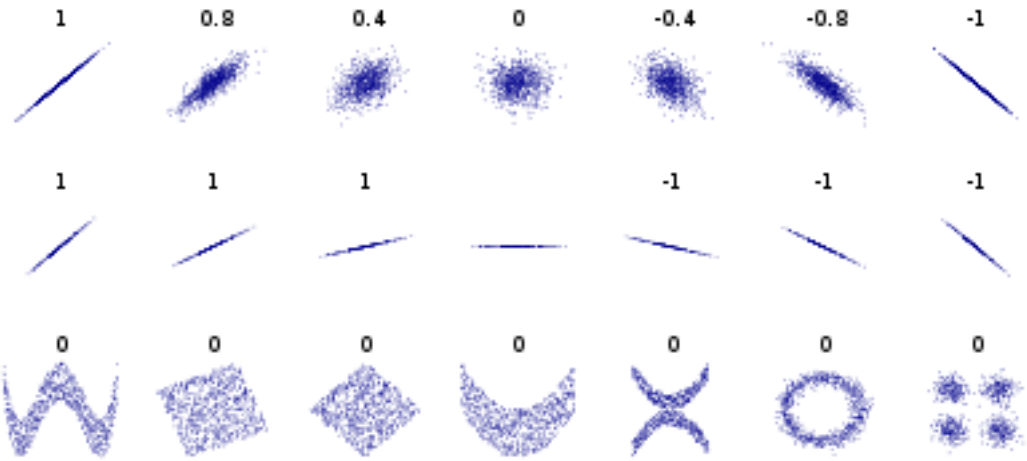
$d_3 = (4, 7, 5)$

$\mu_j = \frac{1}{3} \cdot (11, 24, 23) = (11/3, 8, 23/3)$

Pearson Correlation

Let $x = (x_1, x_2, \cdots, x_d)$

& $y = (y_1, y_2, \cdots, y_d)$



x = (1, 2, 4, 6)
y = (2, 4, 8, 16)

x = (1, 2, 4, 6)
y= (8, 16, 32, 48)

x= (1, 2, 4, 6)
y= (-1, -2, -4, -7)

Pearson Correlation

Let $x = (x_1, x_2, \cdots, x_d)$

& $y = (y_1, y_2, \cdots, y_d)$

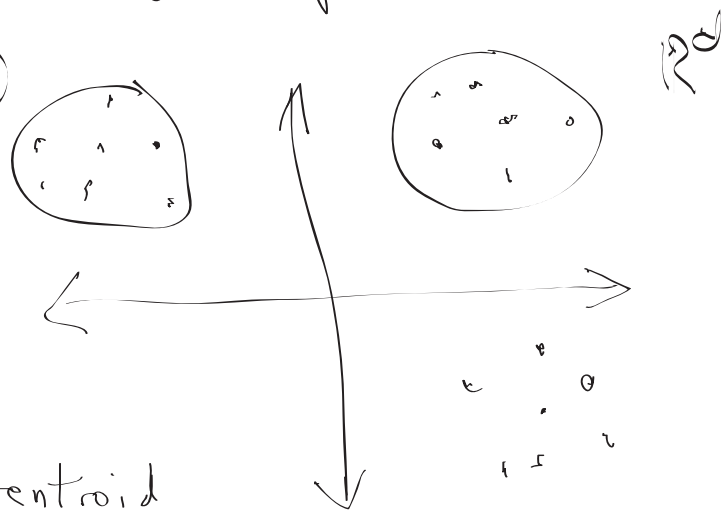
$$r_{xy} = \frac{\sum_{i=1}^{d} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{d} \cdot \sum_{i=1}^{d} x_i$

Kmeans Problem

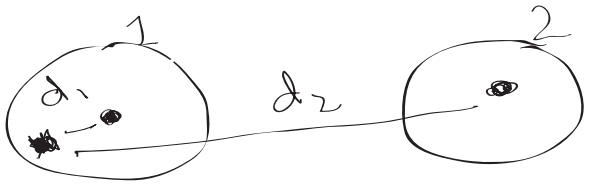
Given: $D = (d_1, \ldots, d_n)$ data points

$d_i = (d_{i1}, \ldots, d_{id})$



Intuition:

Points are assigned to a cluster so that the distance to the centroid of the cluster is shorter than the distance to any other cluster



Assume K clusters with "centers" (centroids)

$$\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^d$$

$$\sum_{j=1}^{K} \sum_{\substack{i \text{ such} \\ \text{that} \\ d_i \text{ is assigned} \\ \text{to cluster } j}} \text{distance}(d_i, \mu_j)$$