k-means

Reviews

Unsurprised learning

$x_1, x_2, \cdots, x_n \qquad x_i \in R^d$

Goal: Assign labels to each data point $y_1, y_2, \cdots, y_n \quad y_i \in \{1, \cdots, k\}$

k = number of clusters

- **Distance-based clustering**

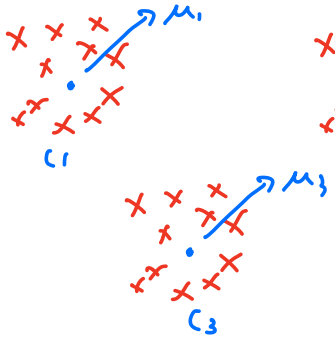
Given centers → Assignment is easy

Define: $u = \frac{1}{n} \sum x_i$ (centroid)

$$F(C_1, C_2, \cdots, C_k) = \sum_{j=1}^{k} \left(\sum_{x_i \in C_j} \| x_i - u_j \|_2^2 \right)$$

k-cluster

Variance in cluster j



If we have centroid, it is quite easy to cluster. But given the dataset, how to choose k suitable centroid?

→ Lloyd's algorithm.

Lloyd's algorithm. (K-means)

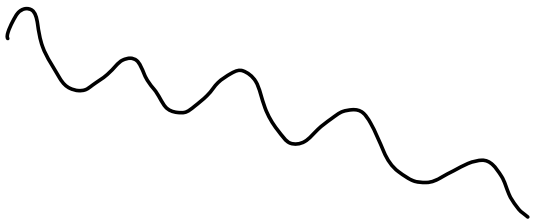
- Initialize: pick random centers $u_1, u_2, \cdots, u_k$
  - ① Assignment $\quad z_i = \underset{j=1,\cdots,k}{\arg\min} \| x_i - u_j \|$
  - ② Update centers $\quad u_i = \frac{1}{n} \sum_{x_i \in C_j} x_i$

Step 1

Calculate the distance between each point and each centroid. and the point belongs to the group of certain centroid once it has the closest distance!

Step 2. Update the centers using existing points belongs to current center.

Not a convex optimization problem.



Some questions:

① How to initialize?

② Do we always find optimal clustering?

③ Convergence?

④ For what kind of data is this suited?

Question 1

$F(C_1, C_2, \dots, C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|_2^2$,

$\mu_1, \dots, \mu_k \equiv$ centroids of $C_1, \dots, C_k$ respectively

- For a fixed k, prove/argue that F achieve a minimum value.

Solution: - Finite ways to assign points into cluster $C_1, \dots, C_k$

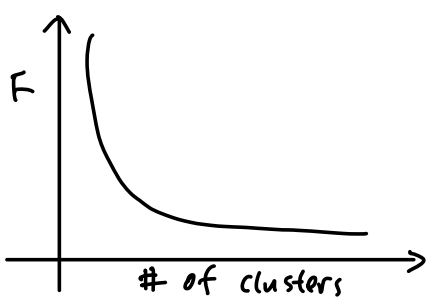
- Among these clusterings, pick the one with smallest value for F

Question 2

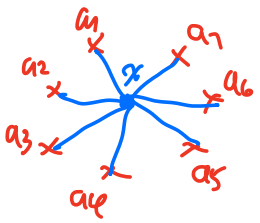
What is the minimum value if $k = n$? ($n \equiv$ number of point)

Solution: Zero. Each point is its own cluster.

Elbow method



Lemma : Let $\{a_1, \cdots, a_n\}$ be a set of n points. Let x be an arbitrary point.



Find the blue point which has the minimum distances among all red points. → the blue point must be the centroid!

Prove: $\sum_{i=1}^{n} \|a_i - x\|^2 = \sum_{i=1}^{n} \|a_i - c\|^2 + n\|c - x\|^2$

where c is the centroid of $\{a_1, \cdots, a_n\}$

Proof: $\sum_{i=1}^{n} \|a_i - x\|^2$

$$\Rightarrow \sum_{i=1}^{n} \|a_i - x + (-c)\|^2$$

$$\Rightarrow \sum_{i=1}^{n} \|\underbrace{a_i - c}_{A} + \underbrace{c - x}_{B}\|^2$$

$$\Rightarrow \sum_{i=1}^{n} \|a_i - c\|^2 + \|c - x\|^2 + 2(a_i - c)^T(c - x)$$

$$= n\|c - x\|^2 + \sum_{i=1}^{n} \|a_i - c\|^2 + \sum_{i=1}^{n} 2(a_i - c)^T(c - x)$$

$$\sum_{i=1}^{n} (a_i - c) = \sum_{i=1}^{n} \left(a_i - \frac{1}{n}\sum_{j=1}^{n} a_j\right)$$

$$= \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \left(\frac{1}{n}\sum_{j=1}^{n} a_j\right)$$

$$= nc - nc = 0$$

$$\|x + y\|^2 = (x+y)^T(x+y)$$
$$= x^Tx + y^Ty + 2x^Ty$$
$$= \|x\|^2 + \|y\|^2 + 2x^Ty$$

$$c = \frac{1}{n}\sum a_i$$

Conclusion: $\sum \|a_i - x\|^2 \geq \sum_{i=1}^{n} \|a_i - c\|^2$

Example

N points in five tight clusters.