K-means

Tuan Nguyen

Ngày 8 tháng 2 năm 2023

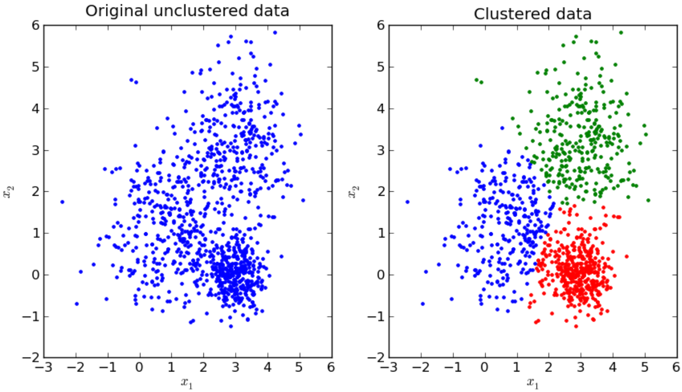
Overview

Unsupervised Learning

K-means

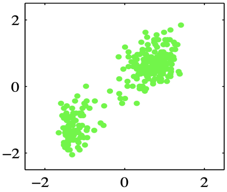
Choose k

Hình 1: Supervised vs Unsupervised



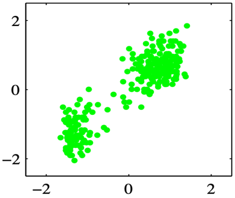
Hình 2: Clustering

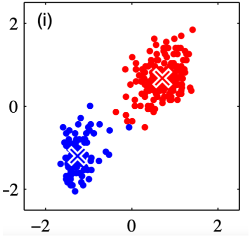
Clustering (cont.)



- Given the dataset x_1, x_2,...,x_N, each $x_i \in \mathrm{R}^D$, partition the dataset into K clusters.
- Intuitively, a cluster is a group of points, which is close together and far from other.

Distortion Measure





- Formally, introduce cluster center $\mu_k \in \mathbb{R}^D$.
- Use binary r_{nk}, 1 if point n is in cluster k, 0 otherwise (1 of K coding scheme again).
- Find $\{\mu_k\}$, $\{r_{nk}\}$ to minimize distortion measure:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \qquad (1)$$

- e.g. two clusters

$$J = \sum_{x_n \in C_1} \|x_n - \mu_1\|^2 + \sum_{x_n \in C_2} \|x_n - \mu_2\|^2 \qquad (2)$$

Minimizing Distortion Measure

Minimizing J directly is hard. Why?

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| x_n - \mu_k \right\|^2 \tag{3}$$

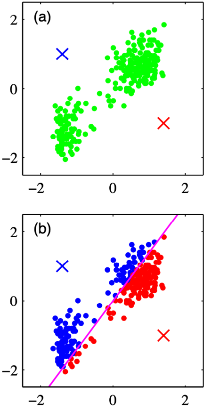
However, two things are easy:

- if we know μ_k, minimizing J wrt r_{nk}
- if we know r_{nk}, minimizing J wrt mu_k

$\Rightarrow$ Iterative procedure

- Start with initial guess for μ_k
- Iteration of two steps:
  - Minimizing J wrt r_{nk}
  - Minimizing J wrt mu_k

Minimizing Distortion Measure (cont.)



- Minimizing J wrt r_{nk}

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \qquad (4)$$

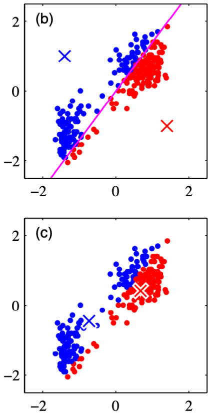
- Loss for each item

$$J_n = \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \qquad (5)$$

$\Rightarrow$ find r_{nk} to minimize J

- Simply set $r_{nk} = 1$ for the cluster center μ_k with smallest distance

Minimizing Distortion Measure (cont.)



- Minimizing J wrt μ_k

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \qquad (6)$$

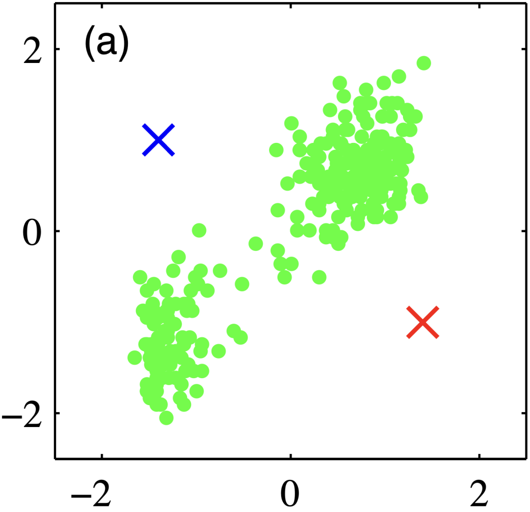
- we can minimize wrt each μ_k separately

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

$$\Leftrightarrow \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} (7)$$

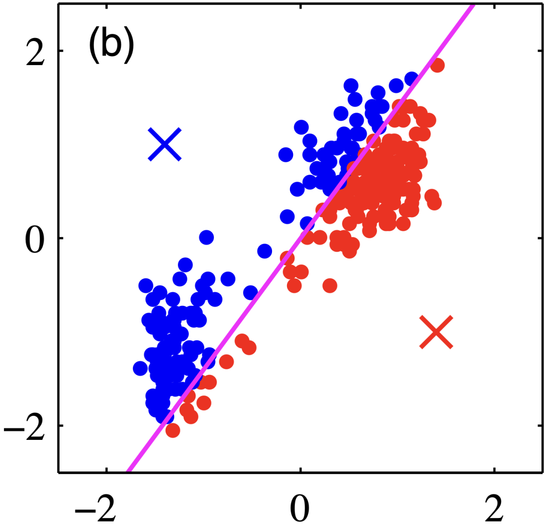
- mean of datapoints x_n assigned to cluster k

Minimizing Distortion Measure (cont.)



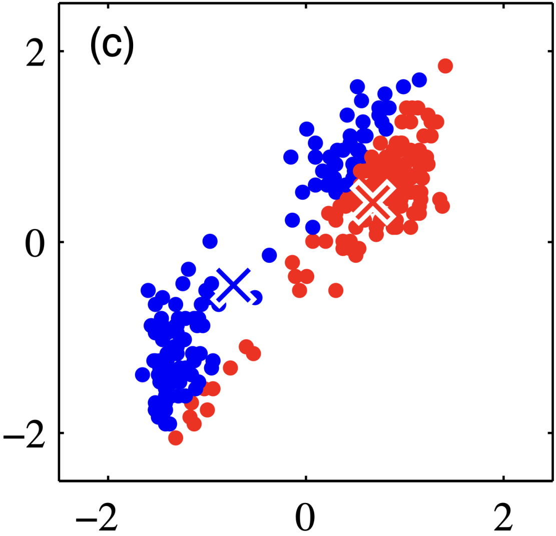
Hình 3: Initialize the cluster center

Minimizing Distortion Measure (cont.)



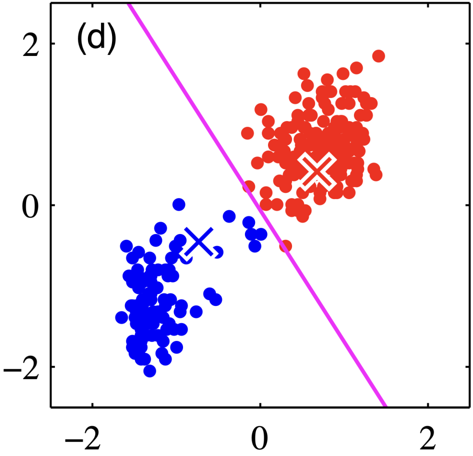
Hình 4: Initialize the cluster center

Minimizing Distortion Measure (cont.)



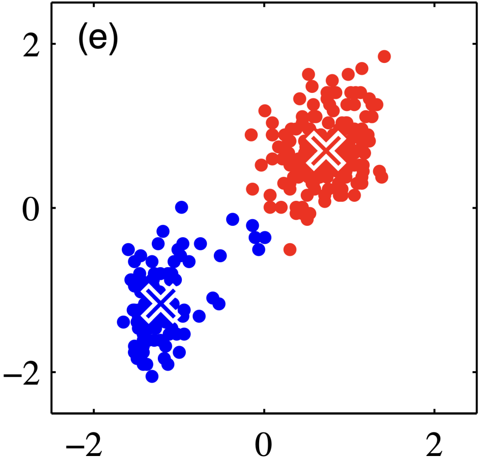
Hình 5: Assign points to the cluster

Minimizing Distortion Measure (cont.)



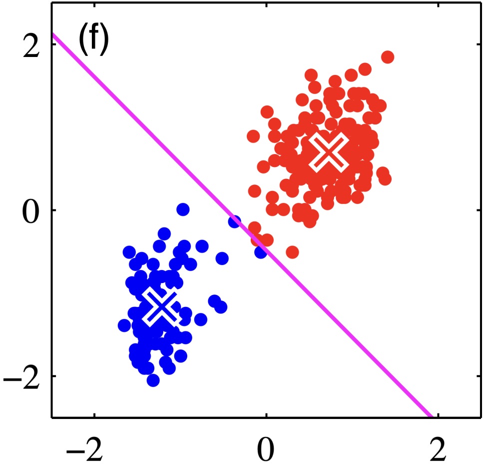
Hình 6: Update cluster center

Minimizing Distortion Measure (cont.)



Hình 7: Assign points to the cluster again

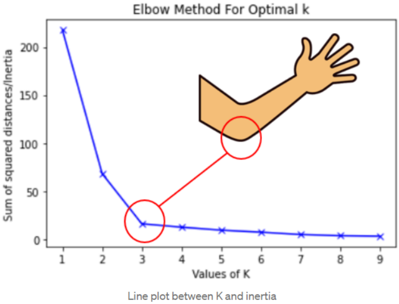
Minimizing Distortion Measure (cont.)



Hình 8: Update cluster center again

Elbow method

- Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k
- Choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.



Hình 9: Elbow method

Silhouette analysis

The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation)

$$S(i) = \frac{b(i) - a(i)}{max(\{a(i), b(i)\})} \tag{8}$$

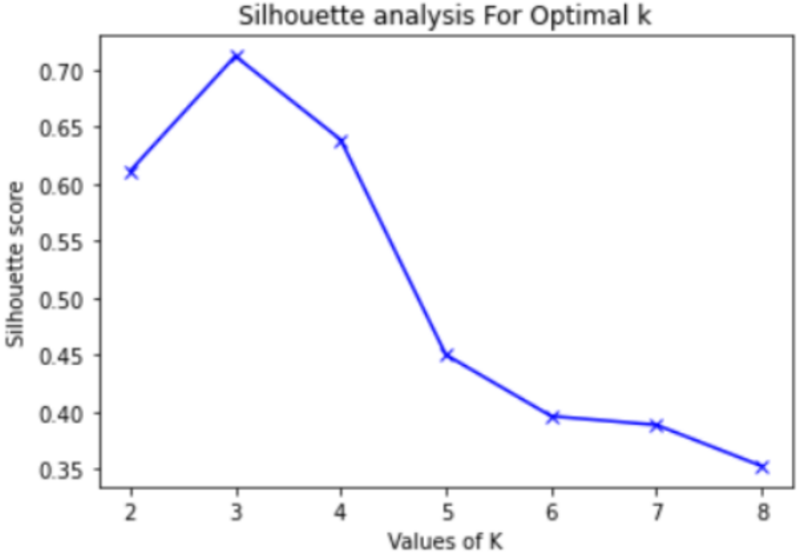
- S(i) is the silhouette coefficient of the data point i.
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.

Silhouette analysis (cont.)

a(i)

b(i)

Silhouette analysis (cont.)



Line plot between K and Silhouette score