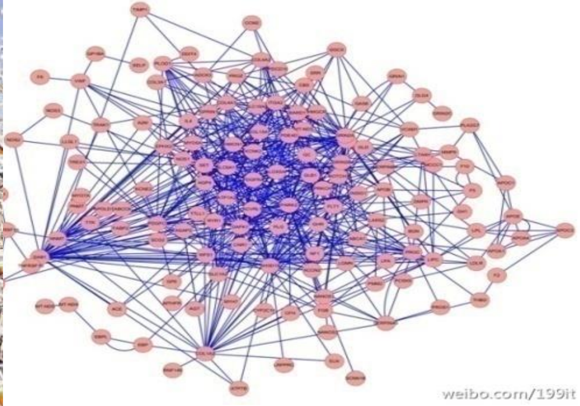


Chapter 9 Clustering: K-means

2025 Autumn

Lei Sun

西安电子科技大学
1931
XIDIAN UNIVERSITY

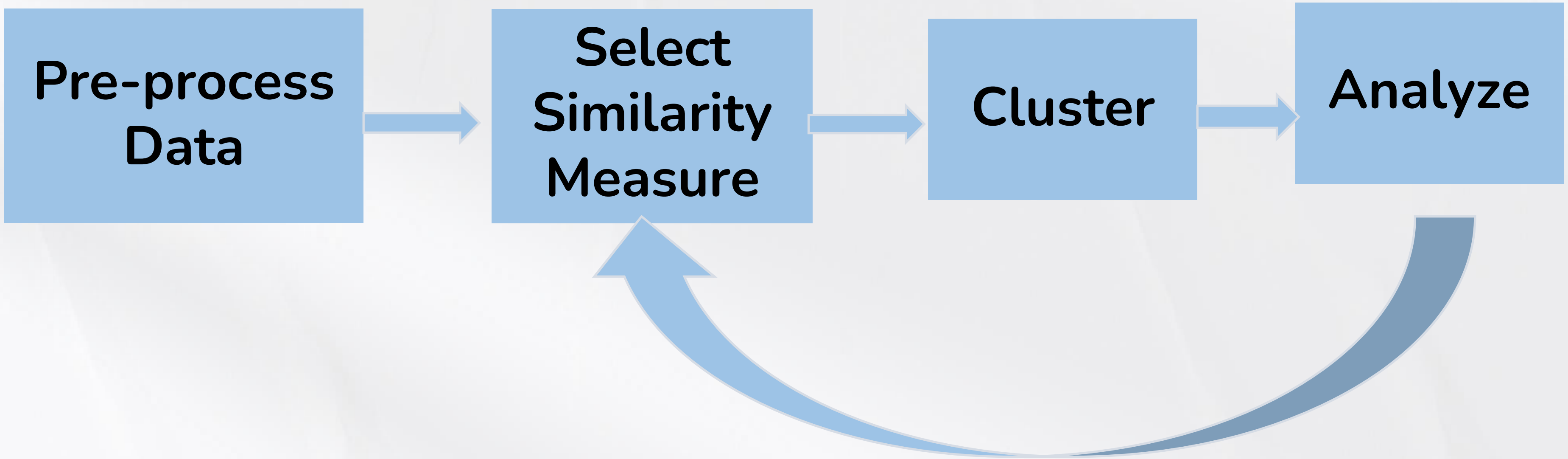
01 K--means

02 Choosing the value of K

03 K-means characteristics

Clustering

Process



Clustering

What is K-means Clustering?

- ✓ Unsupervised Learning – worked with unlabeled data
- ✓ Groups data according to its similarity and distinct patterns
- ✓ The goal of clustering is to divide the set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

Clustering

Application

Clustering has been widely used across industry for years:

- Biology – for genetic and species grouping
- Medical imaging – for distinguishing between different kinds of tissues
- Market research – for differentiating groups of customers based on some attributes
- Recommender system – giving better Amazon purchase suggestions or movie matches

02

Basic Idea

- ✓ K means clustering, assigns data points to one of the K clusters depending on their **distance** from the center of the clusters.
- ✓ Group unlabeled data into clusters
  - – Similar to one another within the same cluster (high intra-class similarity)
  - – Dissimilar to the objects in other clusters(low inter- class similarity)

$$E = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \mu_i\|^2$$

The only information used in clustering is the similarity between examples.

Goal: Group the examples into k partitions

02

Basic Idea

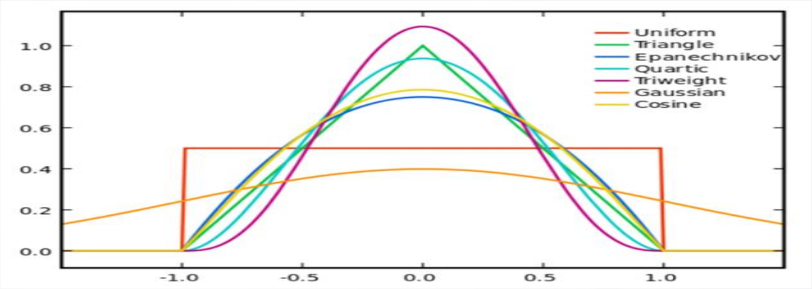
Different ways exist to measure distances. Some examples:

–Euclidean distance:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

–Kernelized(non-linear) distance:

$$K(d) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{d^2}{2}) \cdot I(|d| < 1)$$



03

K-means algorithm

Input: n examples $\{x_1,\ldots,x_n\}$, and the number of partitions $\boldsymbol{k}$

① Define the “centroid” of a group: point whose fields are the average of the fields’ values of points in the group.

② Initialize: k cluster centers $\mu_1,\ldots,\mu_k$. Several initialization options:

–Randomly initialized anywhere

–Choose any k examples as the cluster centers

③ Iterate:

–Assign each of example to its **closest centroid**

–Move the centroids to the average of all records assigned to it

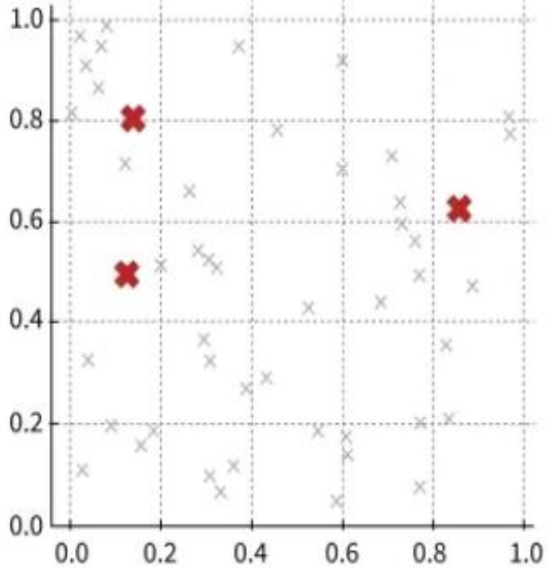
④ A possible convergence criteria: cluster centers do not change anymore

Maximum loop number

Choose Initial Centroids

Centroids are randomly chosen from the data points. These represent the initial cluster centers.

✖ Centroids × Data Points

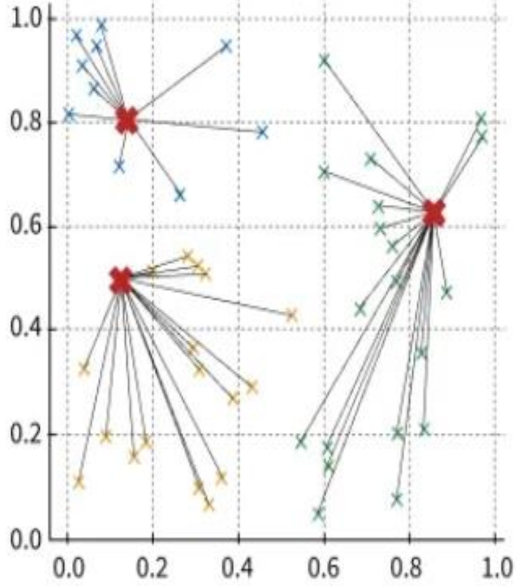


Assign Points to Nearest Centroid

Each point is assigned to the nearest centroid, forming clusters

✖ Centroids × Cluster 1

× Cluster 2 × Cluster 3



Update Centroids

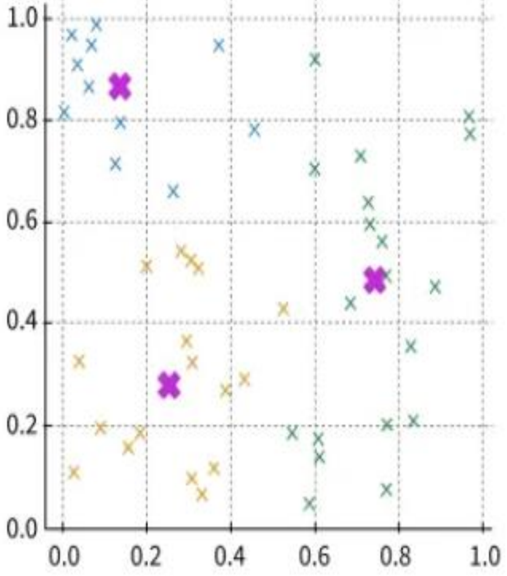
Centroids are recalculated as the mean of the points in each cluster

✖ New Centroids

× Cluster 1

× Cluster 2

× Cluster 3



Repeat Until Convergence

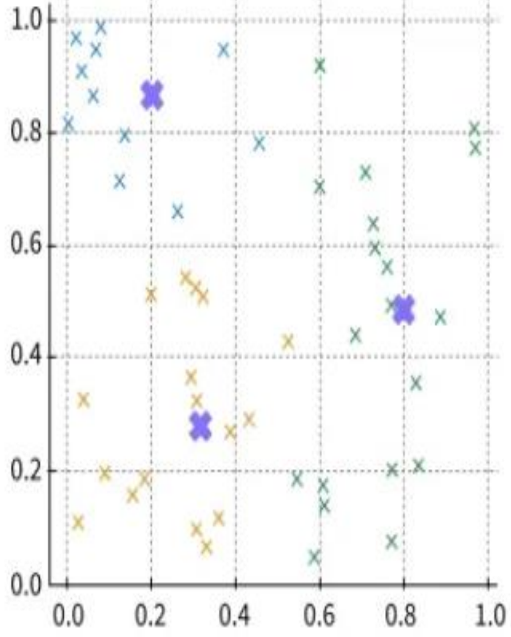
This process repeats until the centroids stabilize and do not move further.

✖ Final Centroids

× Cluster 1

× Cluster 2

× Cluster 3



K-means algorithm

Example： insurance data

No.	age	sex	amount	Ticket number	precedent	agent	result
1	52	M	2000	0	1	Jones	No fraud
2	38	M	1800	0	0	none	No
3	21	F	5600	1	2	Smith	Fraud
4	36	F	3800	0	1	none	No
5	19	M	600	2	2	Adams	No
6	41	M	4200	1	2	Smith	Fraud
7	38	M	2700	0	0	none	No
8	33	F	2500	0	1	none	Fraud
9	18	F	1300	0	0	none	No
10	26	M	2600	2	0	none	No

03

K-means algorithm

- agent:

name： score=0

no： score=1

- age:

Age<20	score=0.0
Age 20-40	score=(age-20)/20
Age 40-60	score=1.0
Age 60-70	score=1.0-(age-60)/10
Age>70	score=0.0

- For claim amount, using the formula: Min-Max
- ticket number

0 ticket	score=1.0
1 ticket	score=0.6
2 or more	score=0

- claim precedent

0 prior claims	score=1.0
1 prior claims	score=0.5
2 or more	score=0

03

K-means algorithm

Select No1. as the seed of cluster 1 and
No.3 as the seed of cluster 2

No.	age	sex	amount	Ticket number	precedent	agent	result
1	1	1	0.28	1	0.5	0	No fraud
2	0.9	1	0.24	1	1	1	No
3	0.05	0	1	0.6	0	0	Fraud
4	0.8	0	0.64	1	0.5	1	No
5	0	1	0	0	0	0	No
6	1	1	0.72	0.6	0	0	Fraud
7	0.9	1	0.42	1	1	1	No
8	0.65	0	0.38	1	0.5	1	Fraud
9	0	0	0.14	1	1	1	No
10	0.3	1	0.4	0	1	1	No

03

K-means algorithm

Distance between No.2 and cluster1:

$$d_{21} = \sqrt{(1-0.9)^2 + (1-1)^2 + (0.28-0.24)^2 + (1-1)^2 + (0.5-1)^2 + (0-1)^2}$$
$$= \sqrt{1.2526}$$

Distance between No.2 and cluster2:

$$d_{23} = \sqrt{(0.9-0.05)^2 + (1-0)^2 + (0.24-1)^2 + (1-0.6)^2 + (1-0)^2 + (1-0)^2}$$
$$= \sqrt{4.4601}$$

No.	age	sex	amount	Ticket number	precedent	agent	result
1	1	1	0.28	1	0.5	0	No fraud
2	0.9	1	0.24	1	1	1	No
3	0.05	0	1	0.6	0	0	Fraud
4	0.8	0	0.64	1	0.5	1	No
5	0	1	0	0	0	0	No
6	1	1	0.72	0.6	0	0	Fraud
7	0.9	1	0.42	1	1	1	No
8	0.65	0	0.38	1	0.5	1	Fraud
9	0	0	0.14	1	1	1	No
10	0.3	1	0.4	0	1	1	No

03

K-means algorithm

- compute the mean points of cluster1 and cluster2:

No.	age	sex	amount	Ticket number	precedent	agent	result
cluster1	0.304	0.714	0.329	0.8	0.714	0.714	No fraud
Cluster 2	0.283	0.333	0.545	0.1667	0.1667	0.333	No

- respectively compute the distance between each point and the mean points of cluster1 and cluster2.

Summary of the example:

- sex: almost male is in cluster1, and female is in the cluster2.
- claim amount、ticket number and agent： no much difference on these three attributes in the two clusters.
- The customer in cluster1 has less fraud than the customer in cluster2.
- conclusion： In cluster1 the customer’s age is older, and more male, has claim precedent, and usually has agent.

No.	age	sex	amount	Ticket number	precedent	agent	result
1	52	M	2000	0	1	Jones	No fraud
2	38	M	1800	0	0	none	No
3	21	F	5600	1	2	Smith	Fraud
4	36	F	3800	0	1	none	No
5	19	M	600	2	2	Adams	No
6	41	M	4200	1	2	Smith	Fraud
7	38	M	2700	0	0	none	No
8	33	F	2500	0	1	none	Fraud
9	18	F	1300	0	0	none	No
10	26	M	2600	2	0	none	No

04

Choosing the value of K

Two indexes for model performance

Elbow method

Silhouette Coefficient
轮廓系数

Useful techniques to evaluate the quality of clustering to determine the **optimal numbers of clusters.**

04

Choosing the value of K

problem : find clusters whose sum of squared deviations(离差平方和)within each cluster is minimum

μ_i is the center of $\mathcal{S}_i$

$$\text{Min}\sum_{i=1}^{\mathcal{K}}\sum_{x\in S_i}\|\mathcal{X}-\mu\|^2$$

where: $(x_{1,}x_2,\dots,x_n)$, K clusters: $(\mathcal{S}_1,\mathcal{S}_2,\dots,\mathcal{S}_{\mathcal{K}})$

- before clustering: for all observations, sum of Squares of Deviation of p variables

$$\text{totss} = \sum_{k=1}^{p} SS_{x_i}$$

- after clustering: sum of the Squares of Deviations of p variables in each resulting clustering

$$\text{tot. withness} = \sum_{k=1}^{\mathcal{K}}\sum_{i=1}^{p} SS'_{x_i} = \sum_{k=1}^{\mathcal{K}} \text{withness}$$

04

Choosing the value of K

- overall degree of discrete among clusters

$$\text{betweenss} = \text{toss} - \text{tot.withness}$$

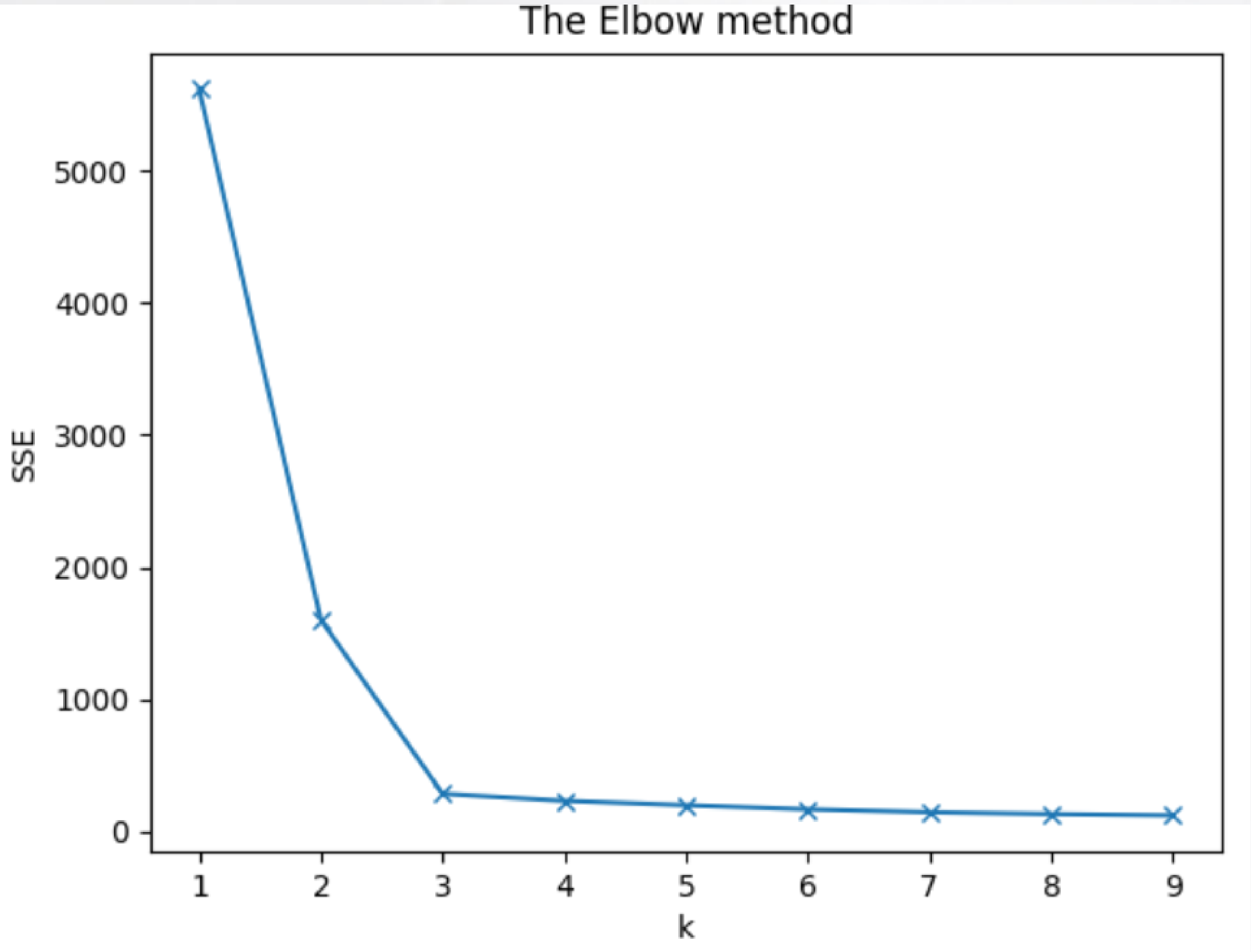
dissimilarity among clusters

- trade off

between/tot.withness $\dfrac{\text{betweenss}}{\mathcal{K}-1} \Big/ \dfrac{tot.withness}{n-\mathcal{K}}$

The larger, the better

Choosing the value of K



Choosing the value of K

Silhouette Coefficient(轮廓系数)

- **Mean intra-cluster distance(a):** Mean distance between the observation and all other data points in the same cluster.

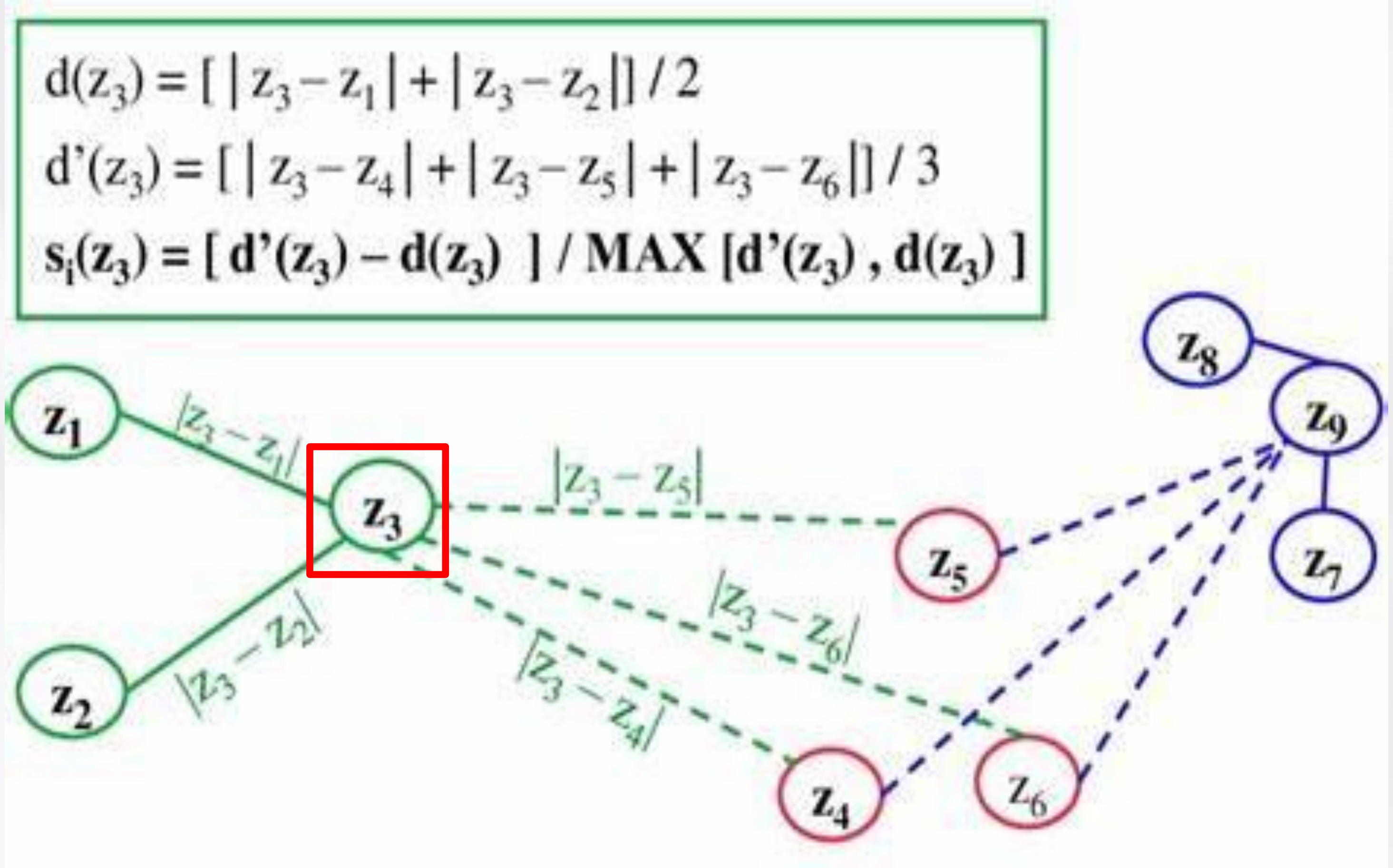
  Cohesion 内聚性

- **Mean nearest-cluster distance(b):** Mean distance between the observation and all other data points of the next nearest cluster.

  Separation 分离性

(b – a) / max(a, b)

a: 某个聚类内部的样本点到其他样本点的欧式距离平均
b: 对于每个样本点i，计算与其最近的其他聚类的样本点的欧式距离平均



$$d(z_3) = [\,|z_3 - z_1| + |z_3 - z_2|\,] / 2$$

$$d'(z_3) = [\,|z_3 - z_4| + |z_3 - z_5| + |z_3 - z_6|\,] / 3$$

$$s_i(z_3) = [\,d'(z_3) - d(z_3)\,] / \mathrm{MAX}\,[d'(z_3), d(z_3)]$$

For the single point **i**:

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

Global Silhouette Coefficient
is the average of S（i）

$$-1 \leq S(i) \leq 1$$

无序有重叠 Sprawling overlapped clusters

Tight, well-separated clusters

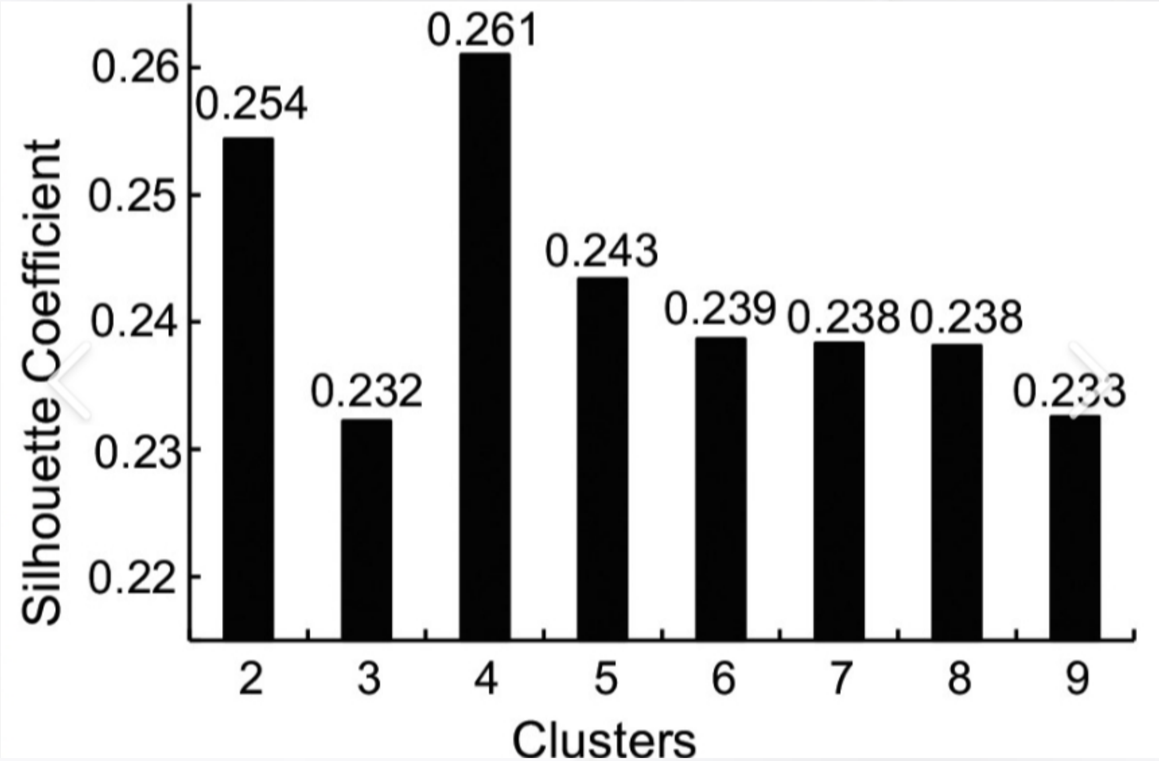
Choosing the value of K

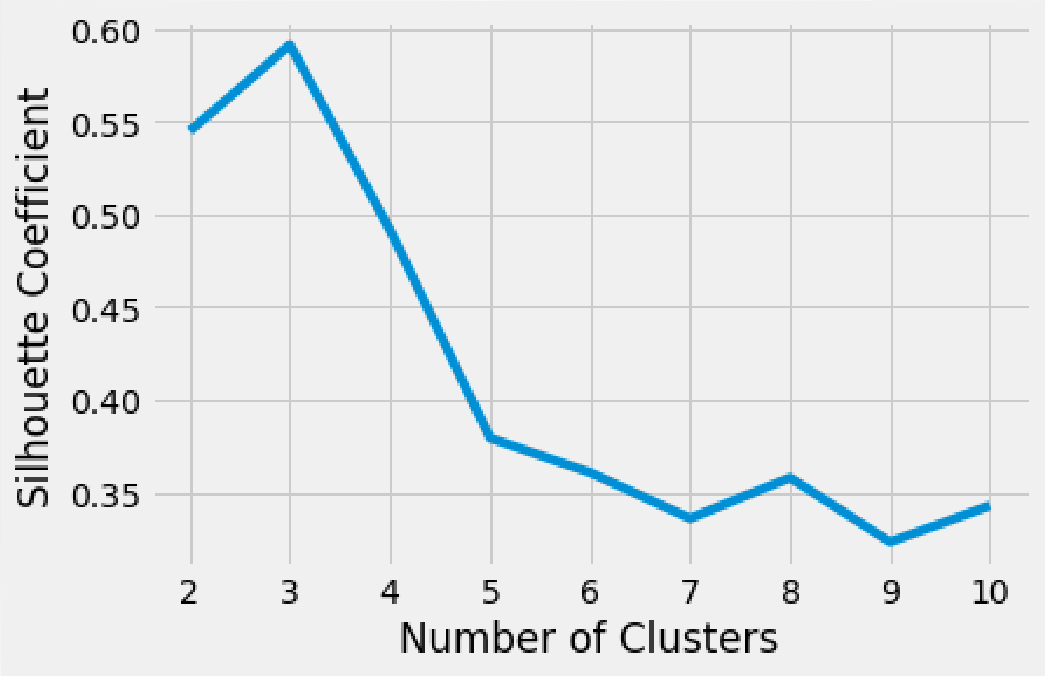
(b – a) / max(a, b)

Silhouette Coefficient(轮廓系数)

- It is used to measure how dense and well-separated the clusters are.
- The silhouette score falls within the range [-1, 1].
- The silhouette score of 1 means that the clusters are very dense and nicely separated.
- The score of 0 means that clusters are overlapping(only one sample). The score of less than 0 means that data belonging to clusters may be wrong/incorrect.
- If the average silhouette score is closer to -1, we say that the clusters are in bad shape and the data points within a cluster have no similarity to each other.

Choosing the value of K





05

Summary

If there is a significant difference in data point density, K-Means may lean towards clusters with higher density and ignore clusters with lower density, resulting in uneven clustering results

Prons	Cons
Simple and work well for regular disjoint clusters	requires apriori specification of the number of cluster centers， K.
Converge relatively fast	Noise and outliers in the data affect the accuracy of clustering.
Good for convex shapes	Bad performance with non-convex shapes
	Data points with different densities affect the accuracy of clustering. Since of convex shapes
	K-means is sensitive to cluster center initialization --May get stuck in local optima --poor convergence speed **Try multiple initializations**