DIGITAL TALENT SCHOLARSHIP 2019

DIGITAL TALENT SCHOLARSHIP

digitalent.kominfo.go.id

Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

Clustering: Pendahuluan

Nama pembicara dengan gelar

Learning Objectives

- **In this lesson you will learn about:**
  - To understand the purpose and mechanism of recommendation systems.
  - To understand different types of recommender systems.
  - To implement recommender system on a real dataset.

Clustering for Segmentation

- The case is we have to segment the customer based on the characteristic of the customer
- So the company can effective to apply the specific business strategy to the customer or to allocated the marketing resources

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Clustering for Segmentation

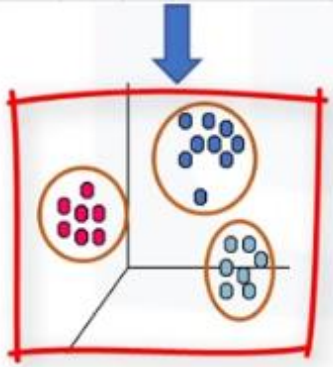
- The customer group could be generated based on the attributes or called as features of the data
- The clustering find the similarity of each customer based on the features

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Clustering for Segmentation

- One of the segmentation method is clustering
- Clustering works by unsupervised method based on the similarity of the customer

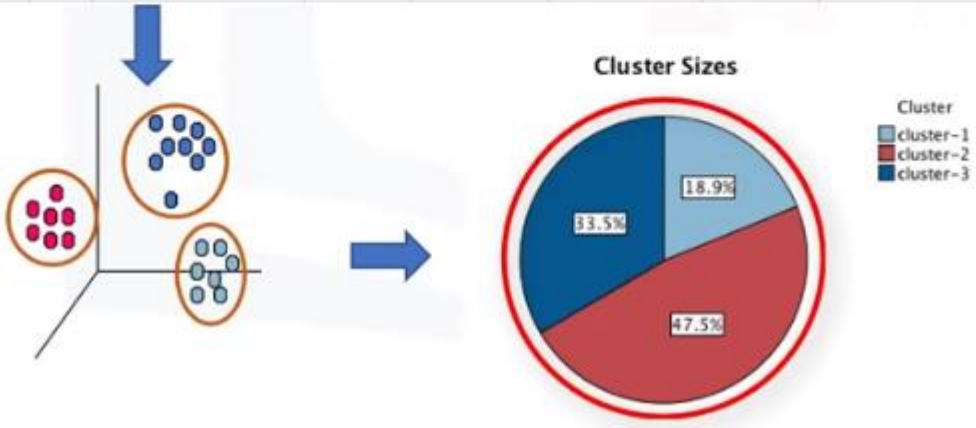
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Clustering for Segmentation

- For example the customer was clustered to be 3 groups
- Each groups has similar demographic

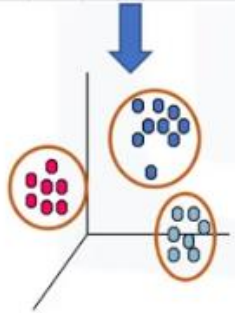
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

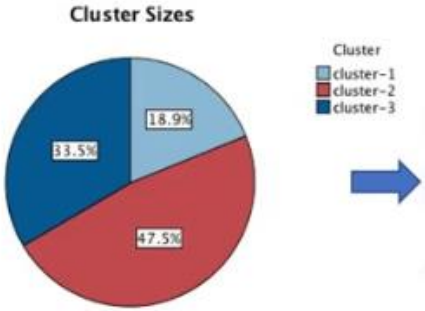


Clustering for Segmentation

- From the clustering result, we can create the profile to each group

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1





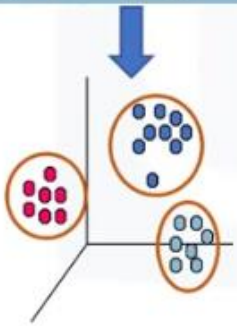
Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

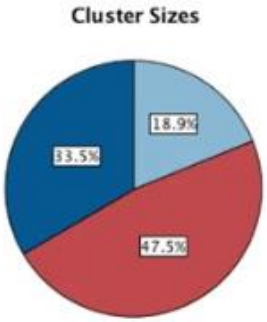
Clustering for Segmentation

- Finally we can assigned the individually data to one of the group

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFULUENT AND MIDDLE AGED
3	AFFULUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFULUENT AND MIDDLE AGED
6	AFFULUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFULUENT AND MIDDLE AGED





Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

Clustering for Segmentation

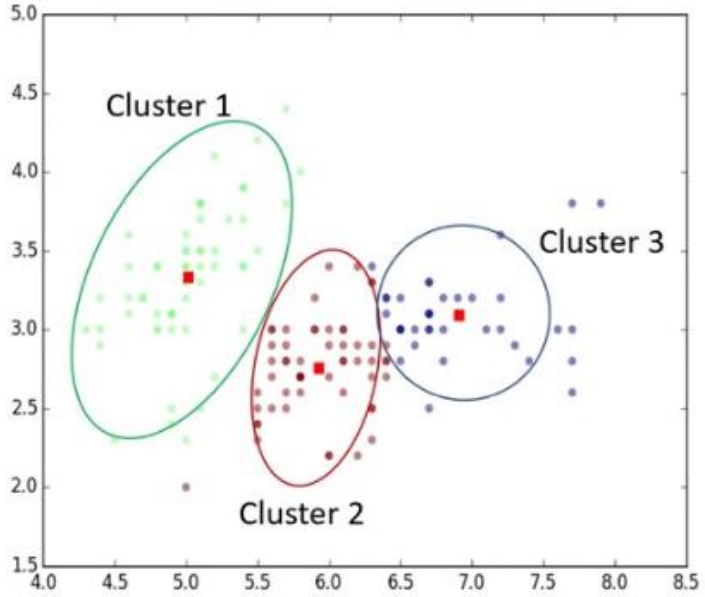
- So, from the clustering result we get
  - Individual customer preferences
  - Buying behaviour across various product
- We can develop the personal experience for each segment

What is clustering?

- Clustering is finding the clusters on the datasets unsupervised

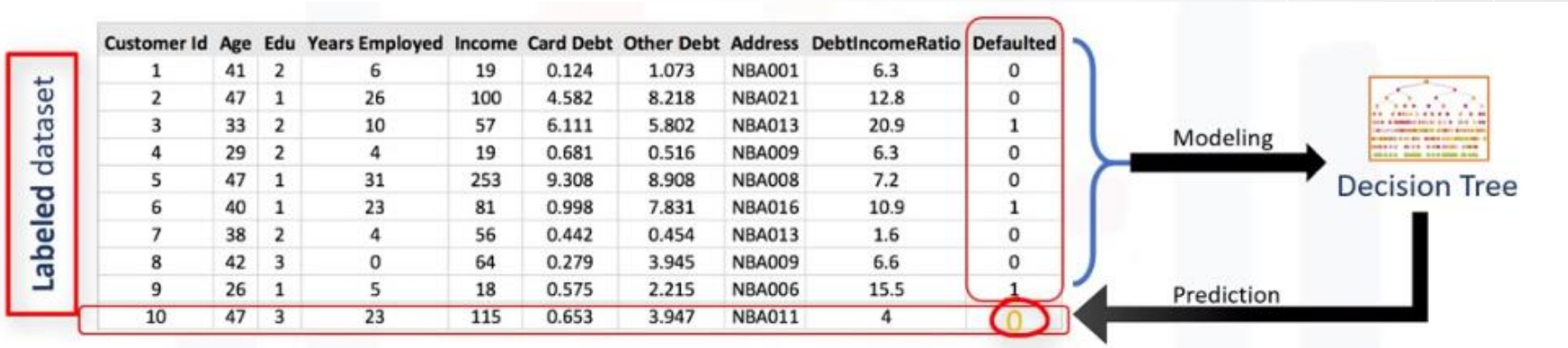
What is a cluster?

A group of objects that are **similar to other objects** in the cluster, and **dissimilar to data points** in other clusters.



Clustering Vs. Classification

- Classification
  - Supervised using labelled datasets



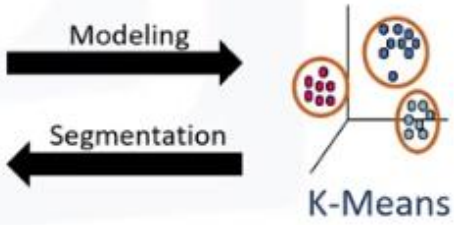
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0

Clustering Vs. Classification

- Clustering
  - Unsupervised using unlabelled datasets

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Clustering Vs. Classification

Labeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0

Modeling → Decision Tree

Prediction ←

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Modeling →

← Segmentation

K-Means

Clustering Application

- RETAIL MARKETING
  - Identifying buying patterns of customers
  - Recommending new books or movies to new customers
- BANKING
  - Fraud detection in credit card use
  - Identifying clusters of customers (e.g., loyal)
- INSURANCE
  - Fraud detection in claims analysis
  - Insurance risk of customers

Clustering Application

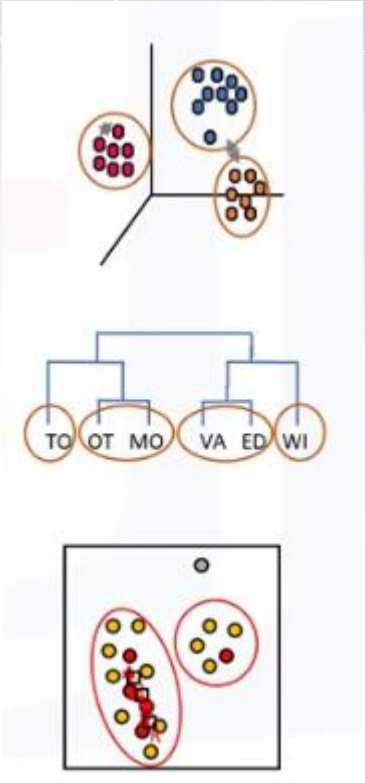
- PUBLICATION
  - Auto-categorizing news based on their content
  - Recommending similar news articles
- MEDICINE
  - Characterizing patient behaviour
- BIOLOGY
  - Clustering genetic markers to identify family ties

Why clustering?

- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

Clustering Algorithms

- Partitioned-based Clustering
  - Relatively efficient
  - E.g., k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
  - Produces trees of clusters
  - E.g. Agglomerative, Divisive
- Density-based Clustering
  - Produces arbitrary shaped clusters
  - E.g. DBSCAN



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

Clustering: K-Means

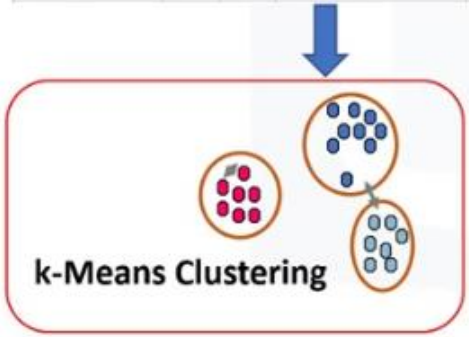
Nama pembicara dengan gelar

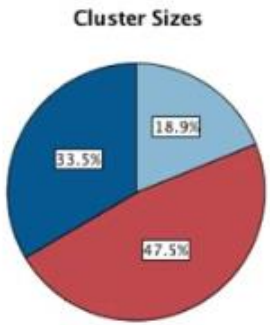
What is k-Means clustering?

- Clustering works on unsupervised data based on the similarity each datasets

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFULUENT AND MIDDLE AGED
3	AFFULUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFULUENT AND MIDDLE AGED
6	AFFULUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFULUENT AND MIDDLE AGED

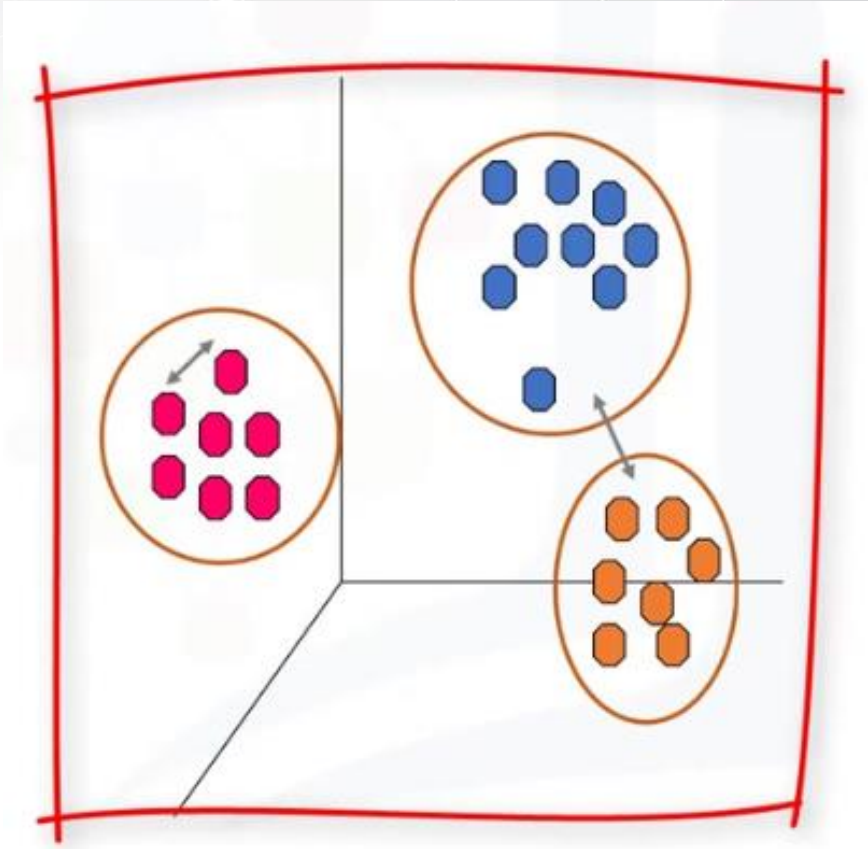




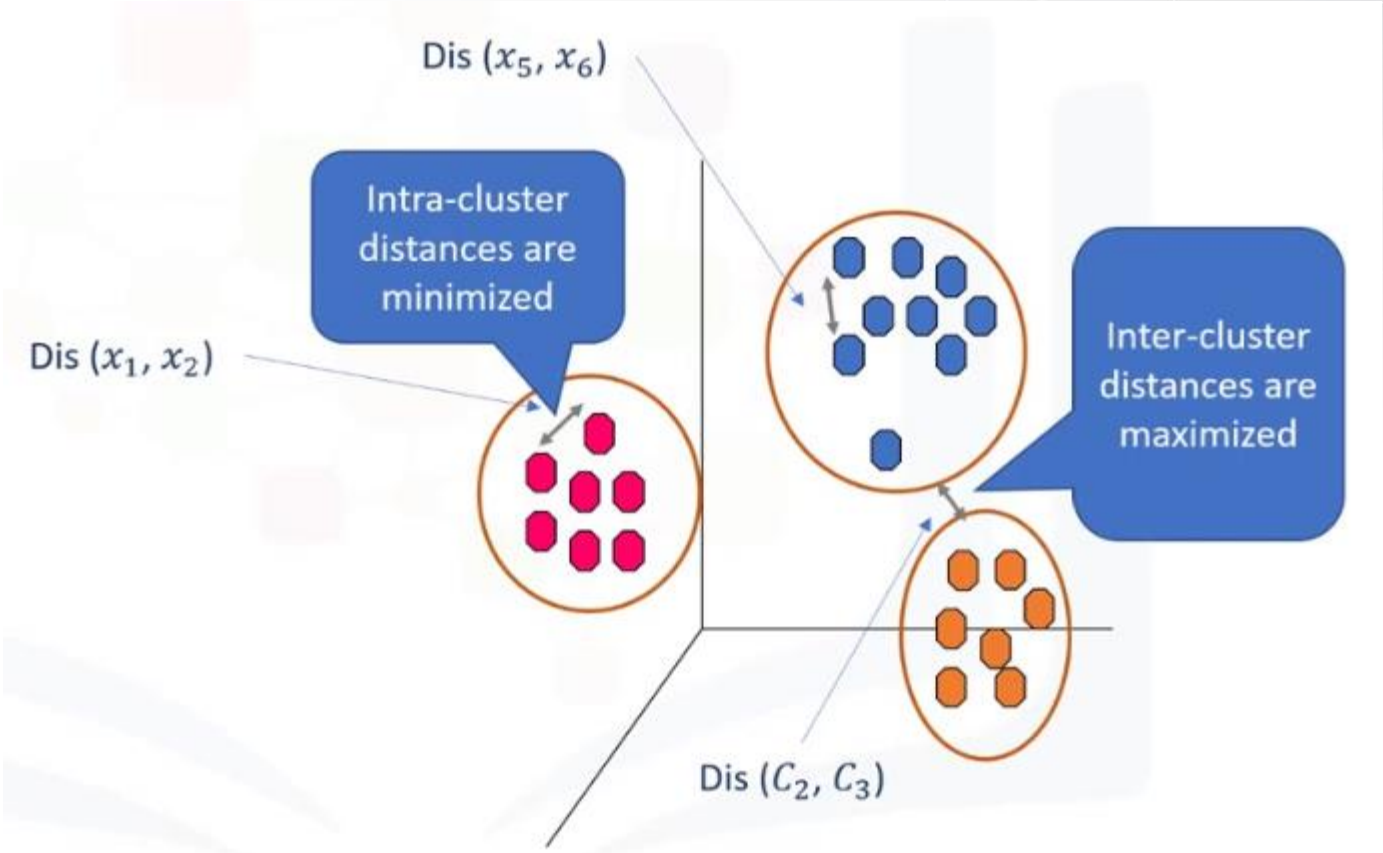
Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

K-Means algorithms

- Partitioning Clustering
- k-Means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar
- Examples across different clusters are very different



Determine the similarity or dissimilarity



1-dimentional similarity/distance

Customer 1
Age
54

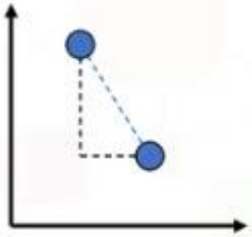
Customer 2
Age
50

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$\text{Dis}\,(x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

2-dimentional similarity/distance

Customer 1	
Age	Income
54	190



Customer 2	
Age	Income
50	200

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2} = 10.77$$

Multi-dimentional similarity/distance

Customer 1		
Age	Income	education
54	190	3

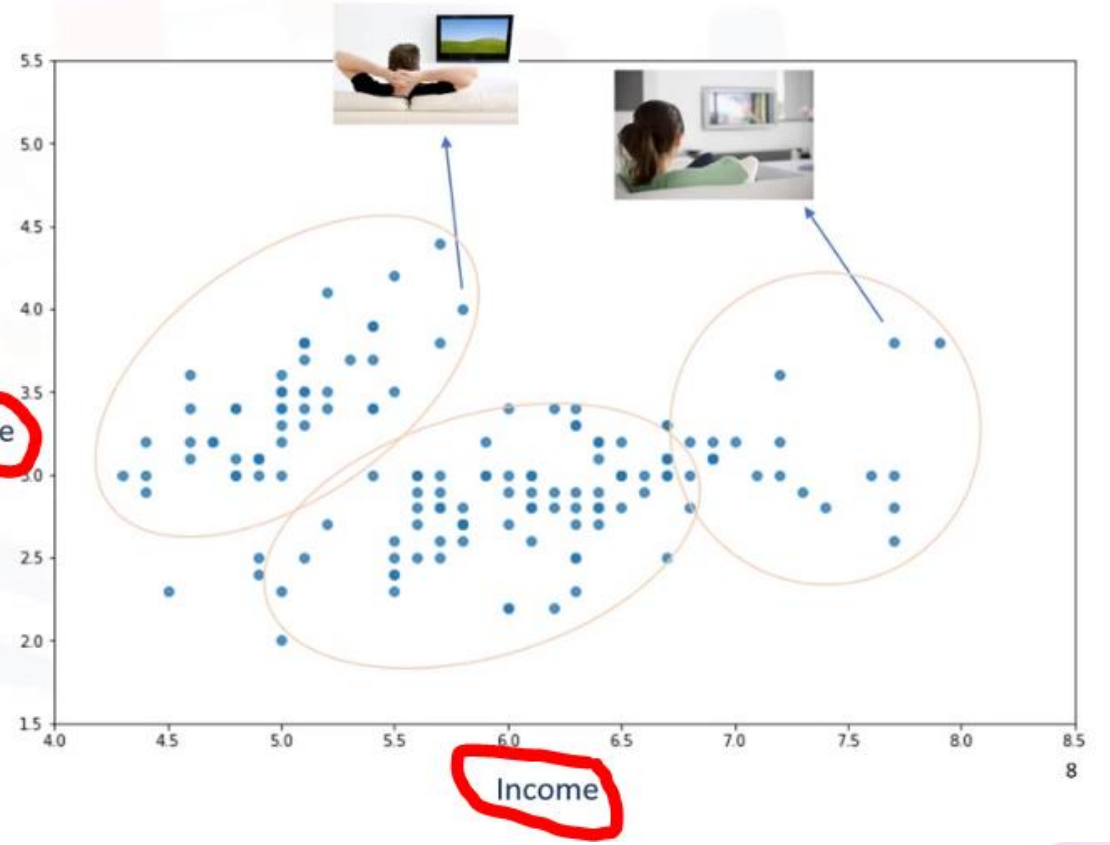
Customer 2		
Age	Income	education
50	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

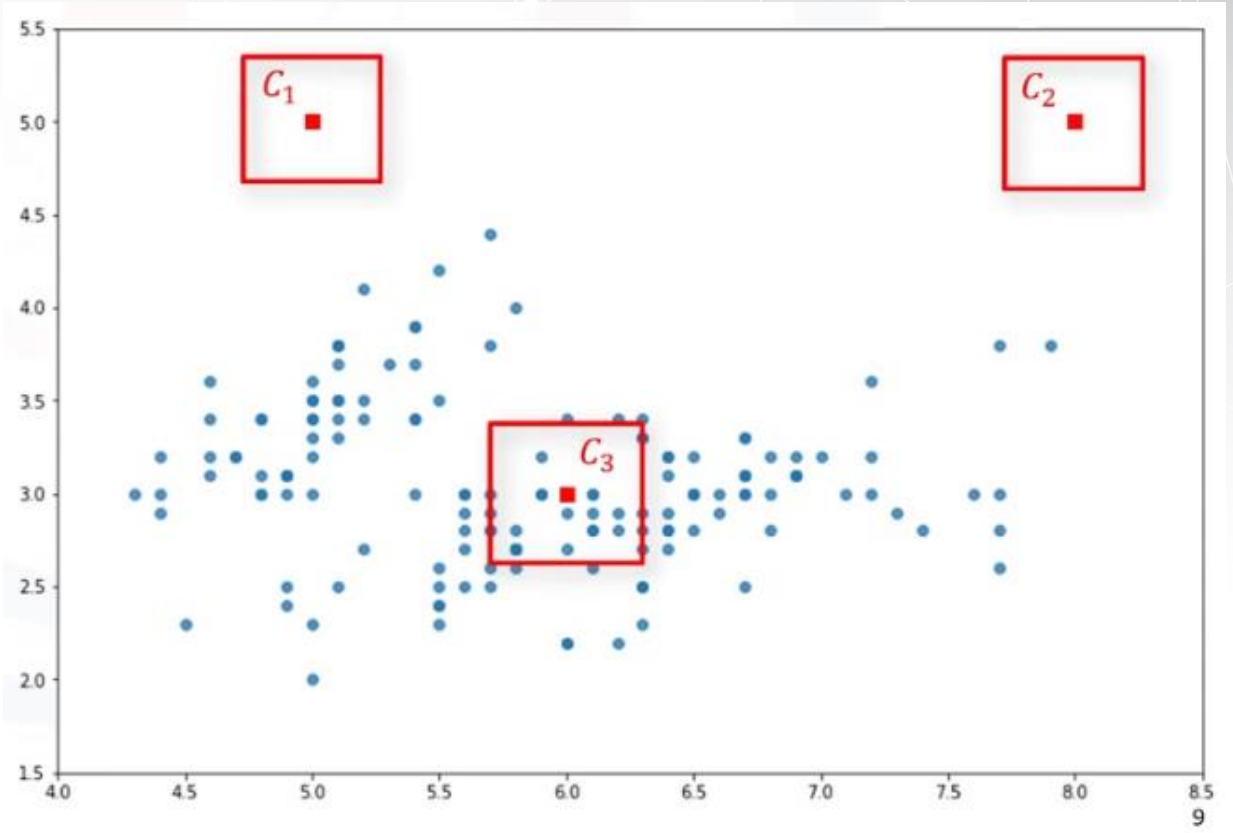
How does k-Means Clustering works?

Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...	...	..



k-Means clustering – initialize k

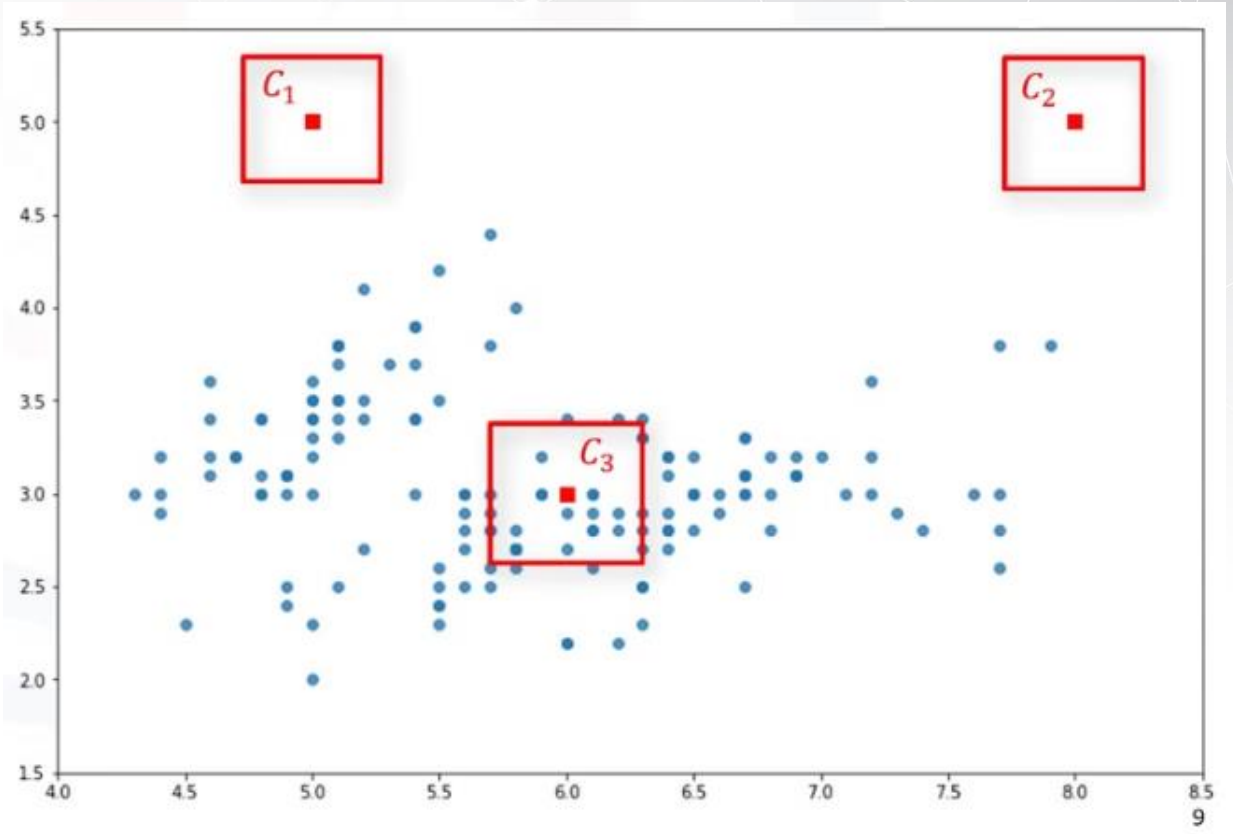
1. Initialize k=3 centroids randomly



k-Means clustering – initialize k

1. Initialize k=3 centroids randomly

C_1 = [8., 5.]
C_2 = [5., 5.]
C_3 = [6., 3.]

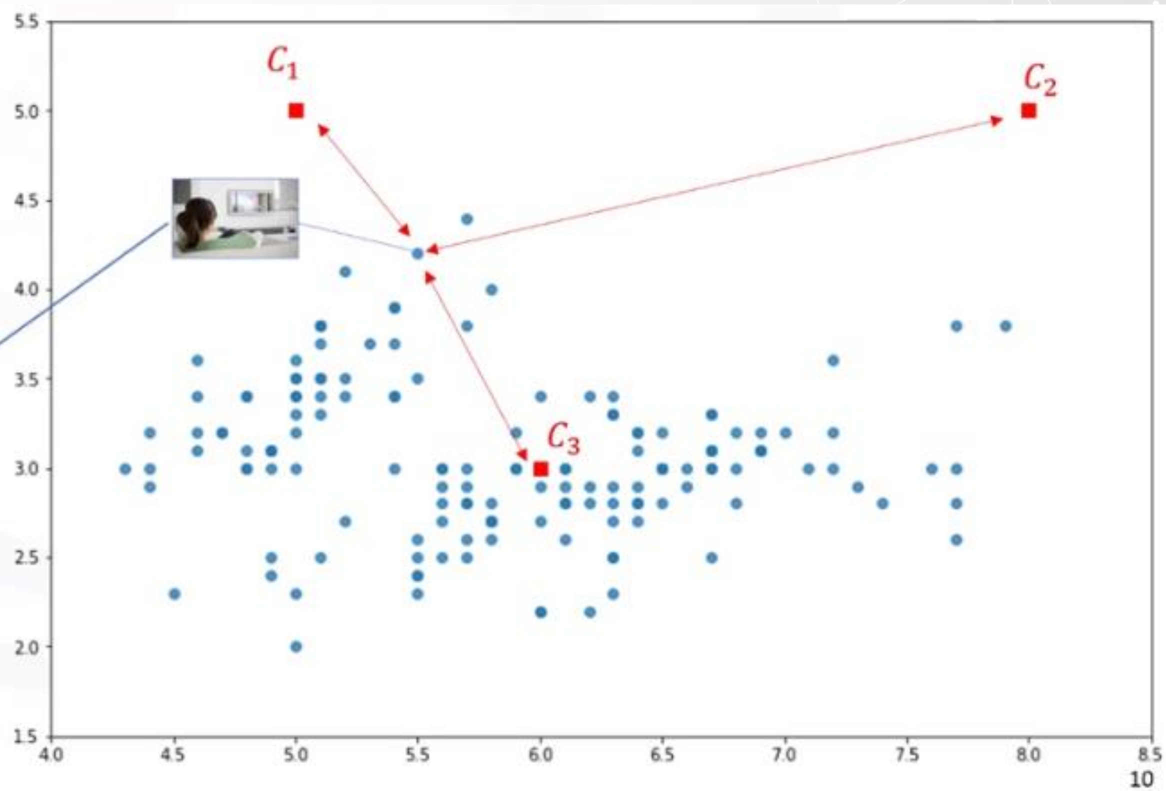


k-Means clustering – calculate the distance

2. Distance calculation

$$
\begin{array}{ccc}
C_1 & C_2 & C_3
\end{array}
$$

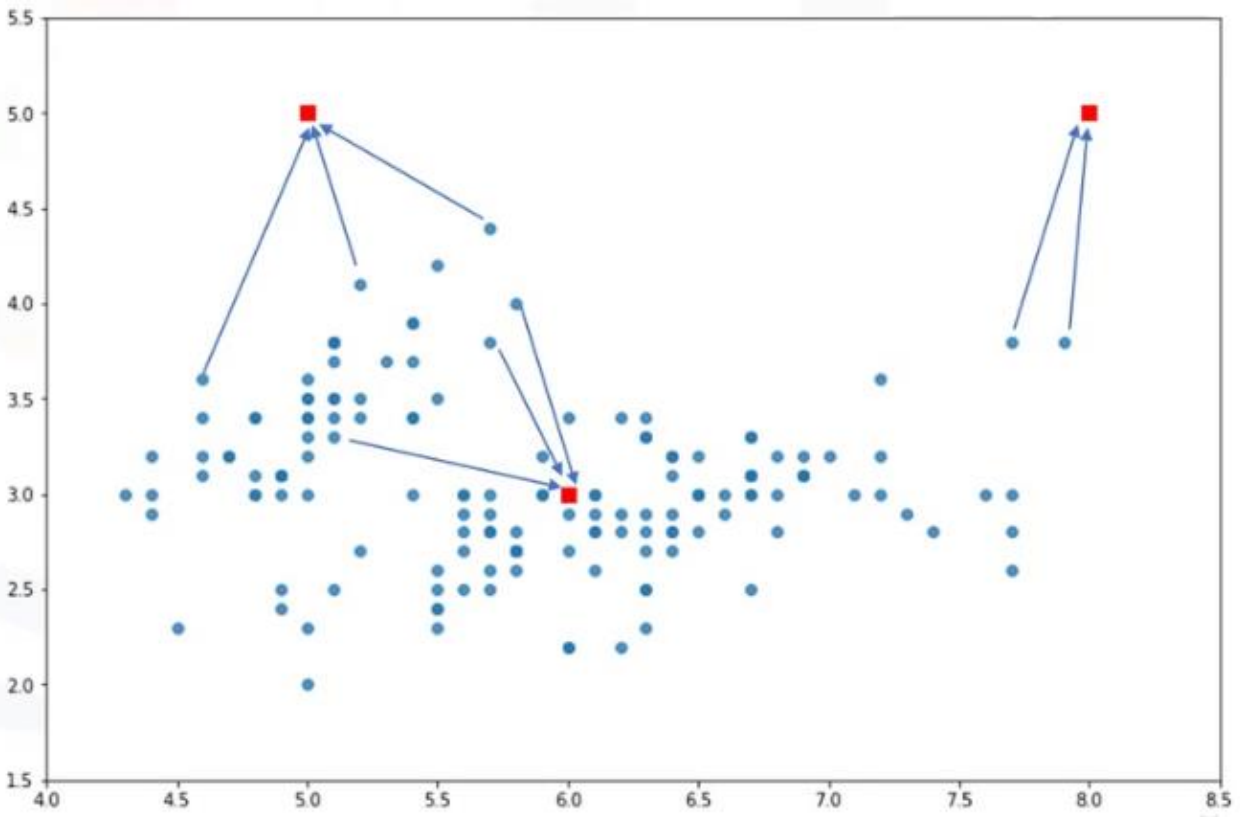
$$
\begin{bmatrix}
d(p1,c1) & d(p1,c2) & d(p1,c3) \\
d(p2,c1) & d(p2,c2) & d(p2,c3) \\
d(p3,c1) & d(p3,c2) & d(p3,c3) \\
d(p4,c1) & d(p4,c2) & d(p4,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3)
\end{bmatrix}
$$



k-Means clustering – assign to centroid

3. Assign each point to the closest centroid

$$\begin{matrix} C_1 & C_2 & C_3 \end{matrix}$$
$$\begin{bmatrix} d(p1,c1) & d(p1,c2) & d(p1,c3) \\ d(p2,c1) & d(p2,c2) & d(p2,c3) \\ d(p3,c1) & d(p3,c2) & d(p3,c3) \\ d(p4,c1) & d(p4,c2) & d(p4,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \end{bmatrix}$$

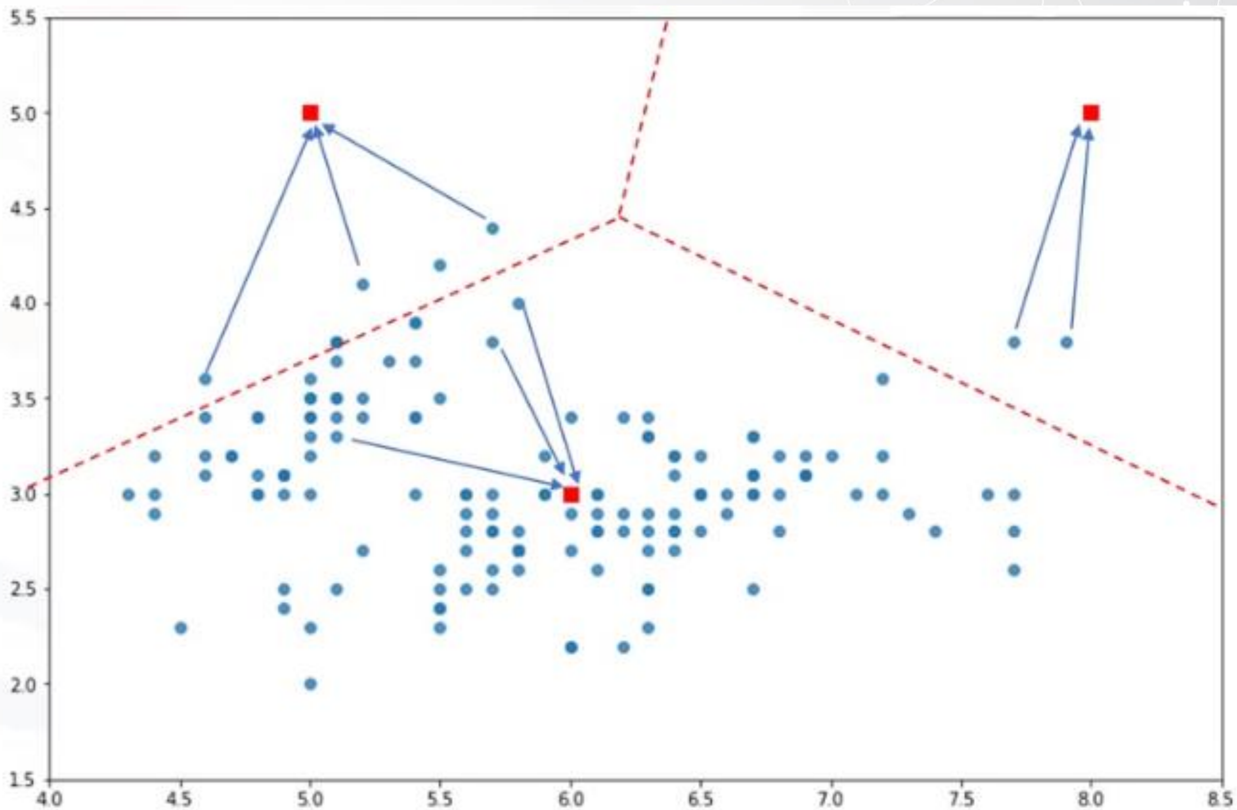


k-Means clustering – assign to centroid

3. Assign each point to the closest centroid

$$\begin{array}{ccc} C_1 & C_2 & C_3 \end{array}$$

$$\begin{bmatrix} d(p1,c1) & d(p1,c2) & d(p1,c3) \\ d(p2,c1) & d(p2,c2) & d(p2,c3) \\ d(p3,c1) & d(p3,c2) & d(p3,c3) \\ d(p4,c1) & d(p4,c2) & d(p4,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \\ d(p..,c1) & d(p..,c2) & d(p..,c3) \\ d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\ d(pn,c1) & d(pn,c2) & d(pn,c3) \end{bmatrix}$$

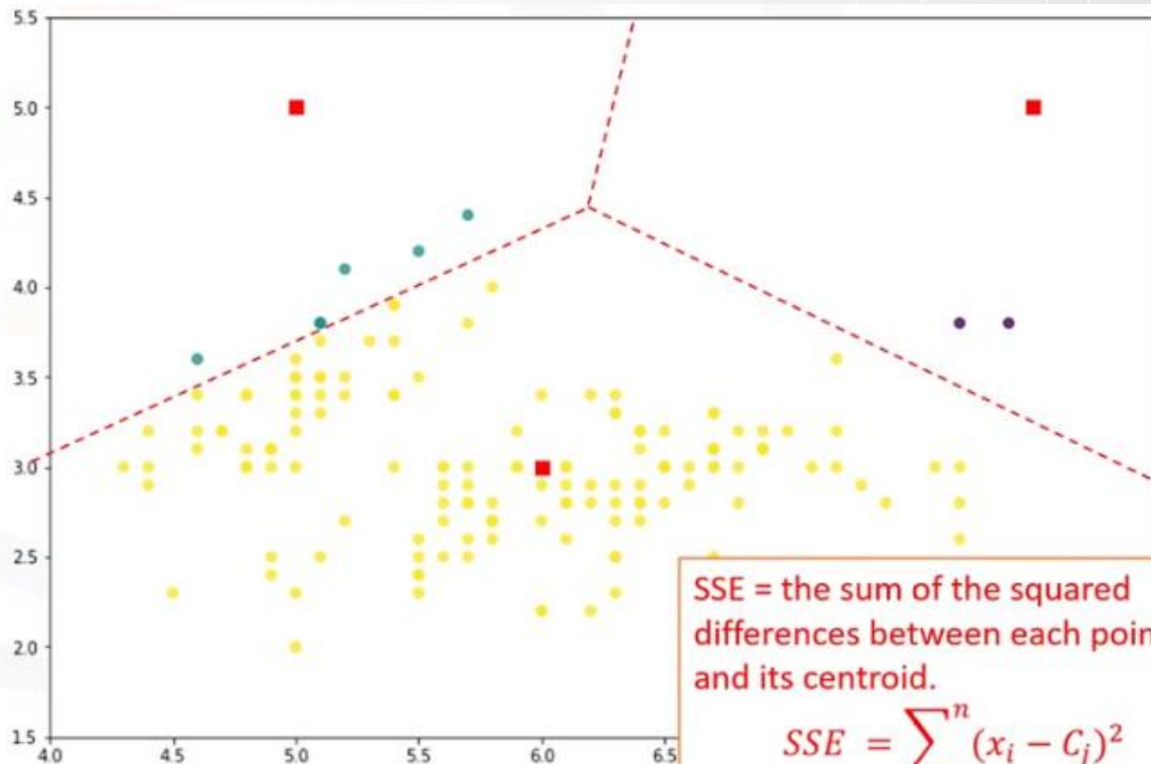


k-Means clustering – assign to centroid

3. Assign each point to the closest centroid

$$
\begin{array}{ccc}
C_1 & C_2 & C_3
\end{array}
$$

$$
\begin{bmatrix}
d(p1,c1) & d(p1,c2) & d(p1,c3) \\
d(p2,c1) & d(p2,c2) & d(p2,c3) \\
d(p3,c1) & d(p3,c2) & d(p3,c3) \\
d(p4,c1) & d(p4,c2) & d(p4,c3) \\
d(p \ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p \ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p \ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p \ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3)
\end{bmatrix}
$$



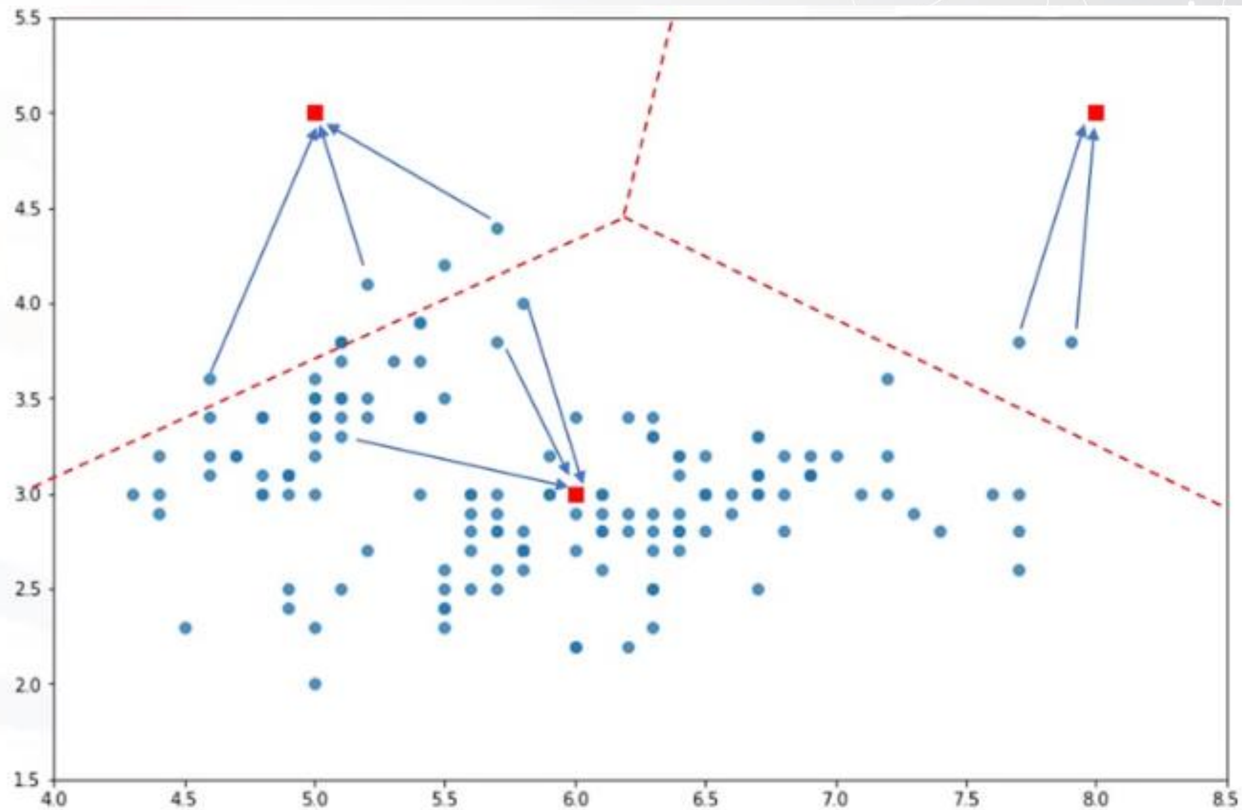
SSE = the sum of the squared differences between each point and its centroid.

$$SSE = \sum_{1}^{n} (x_i - C_j)^2$$

k-Means clustering – assign to centroid

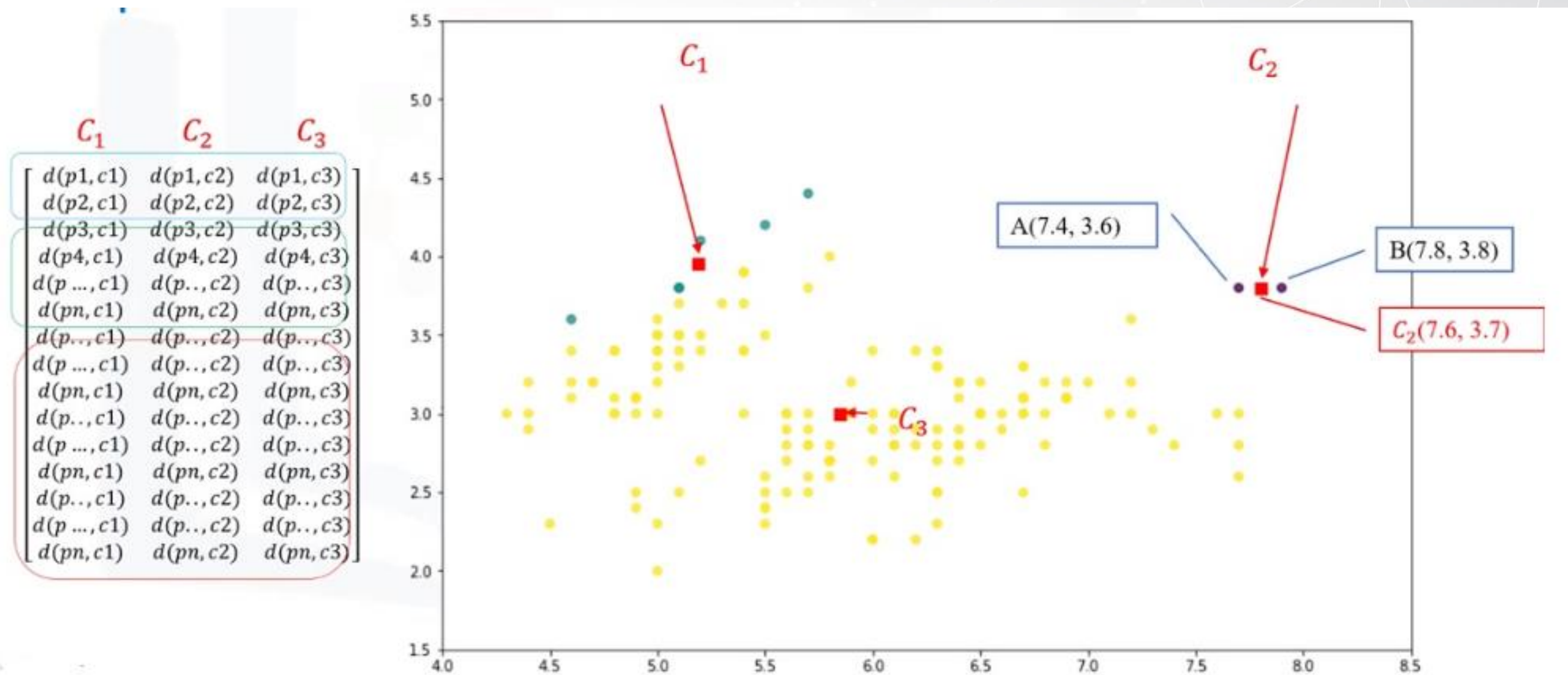
3. Assign each point to the closest centroid

C_1	C_2	C_3
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p.., c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(p \ldots, c1)$	$d(p.., c2)$	$d(p.., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$



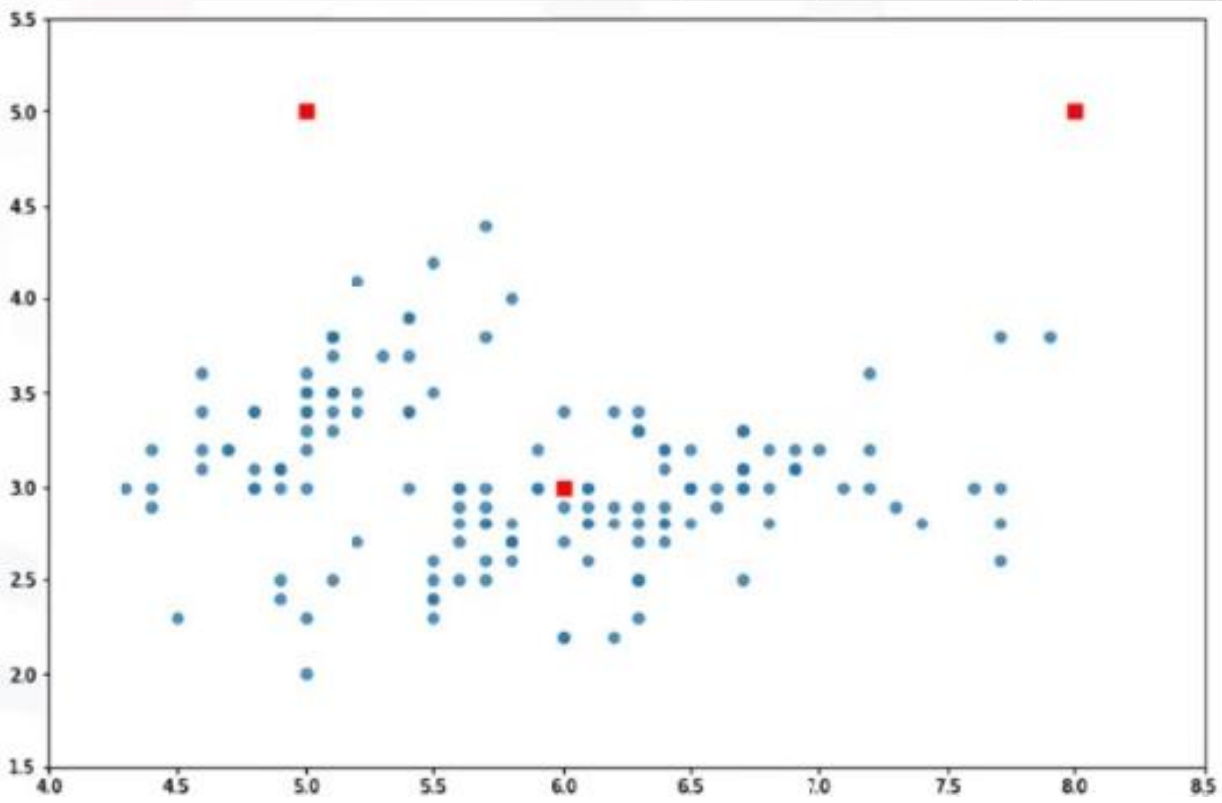
k-Means clustering – compute new centroids

4. Compute the new centroids for each cluster



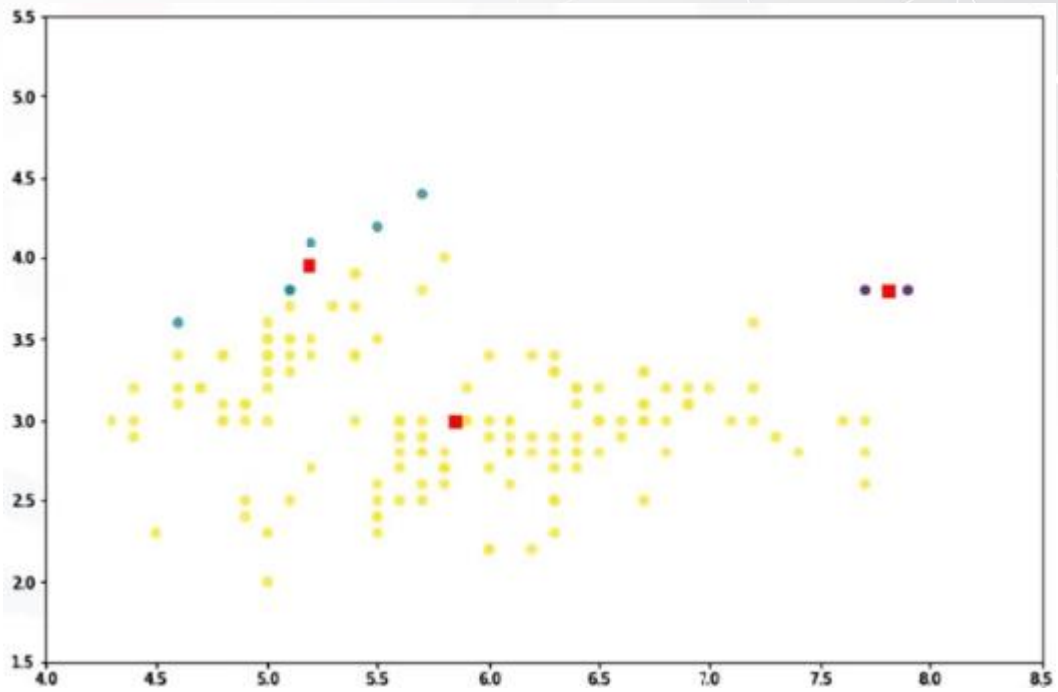
k-Means clustering – repeat

5. Repeat until there are no more changes



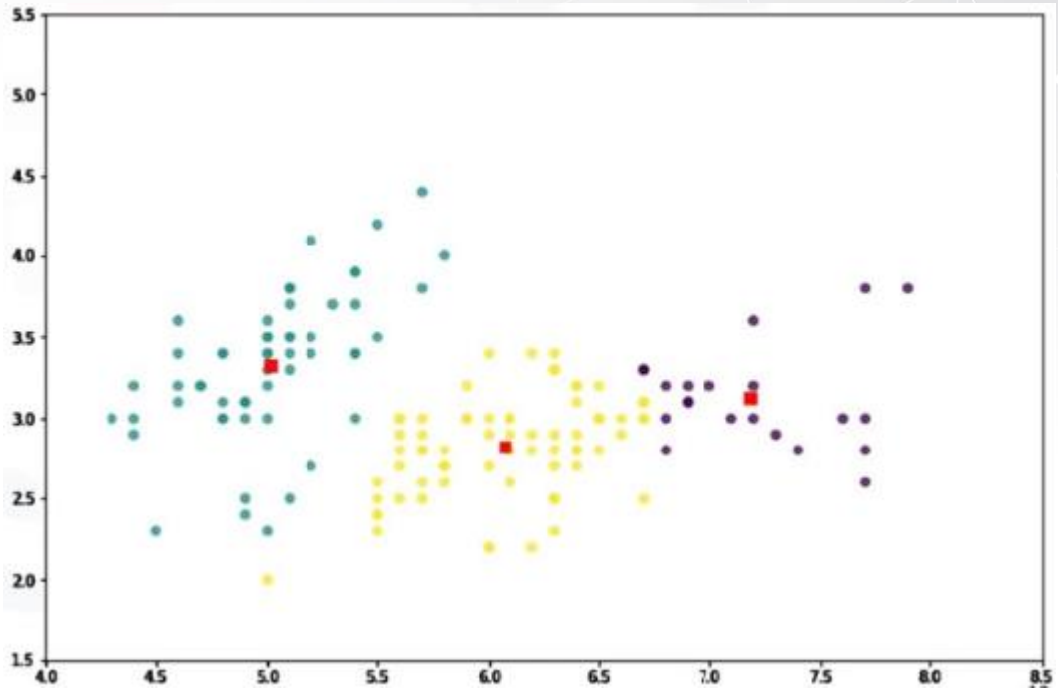
k-Means clustering – repeat

5. Repeat until there are no more changes



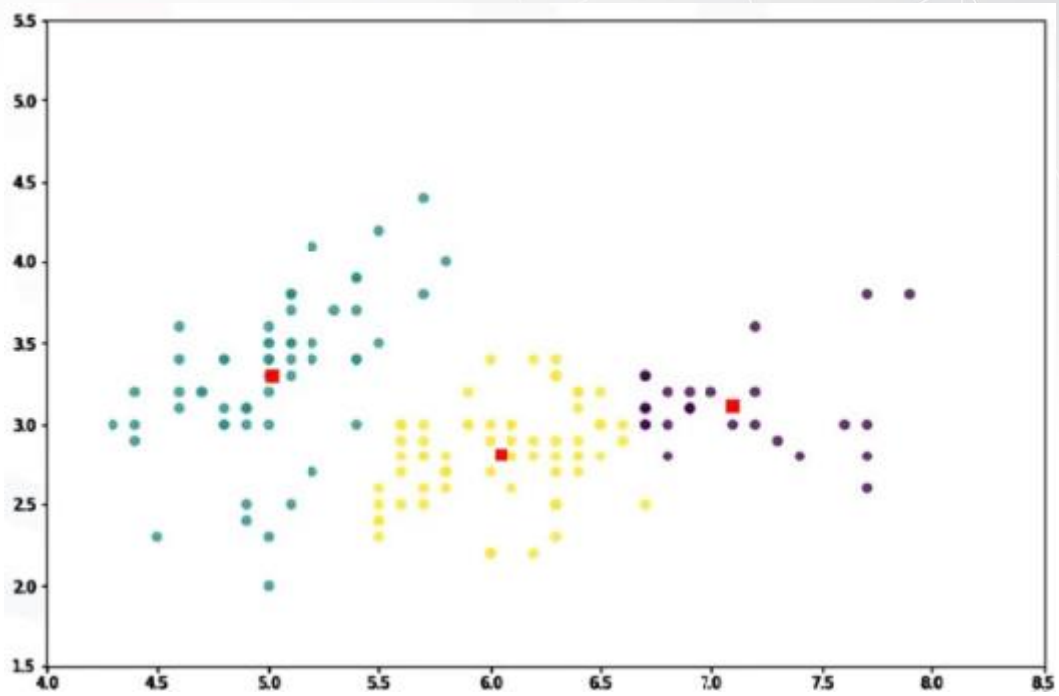
k-Means clustering – repeat

5. Repeat until there are no more changes



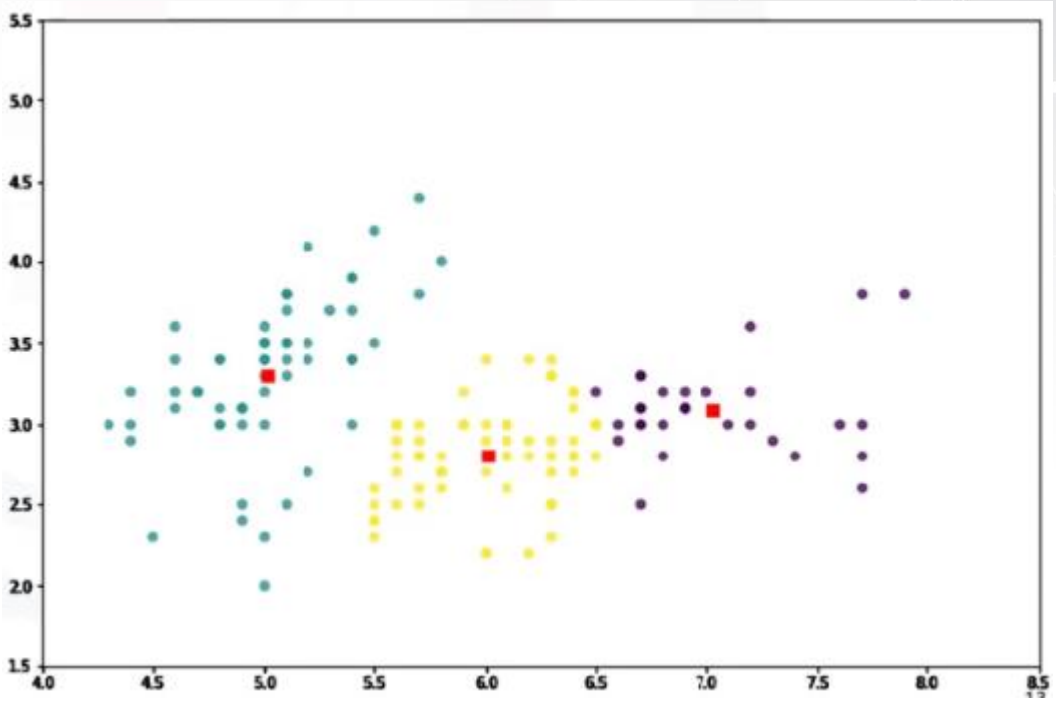
k-Means clustering – repeat

5. Repeat until there are no more changes



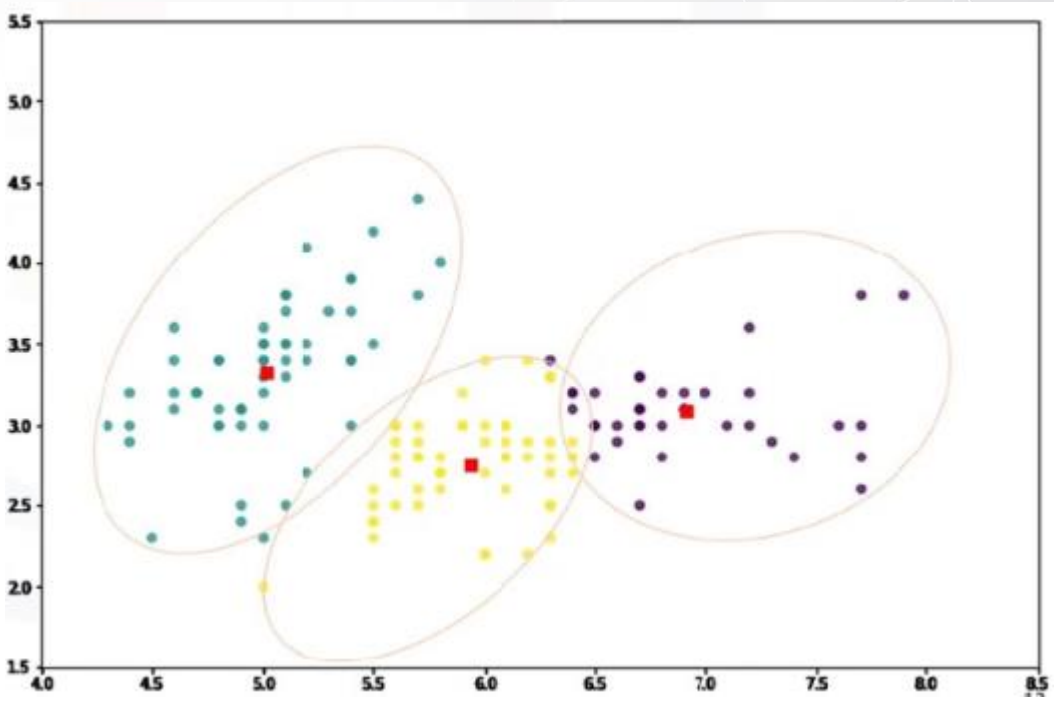
k-Means clustering – repeat

5. Repeat until there are no more changes



k-Means clustering – repeat

5. Repeat until there are no more changes

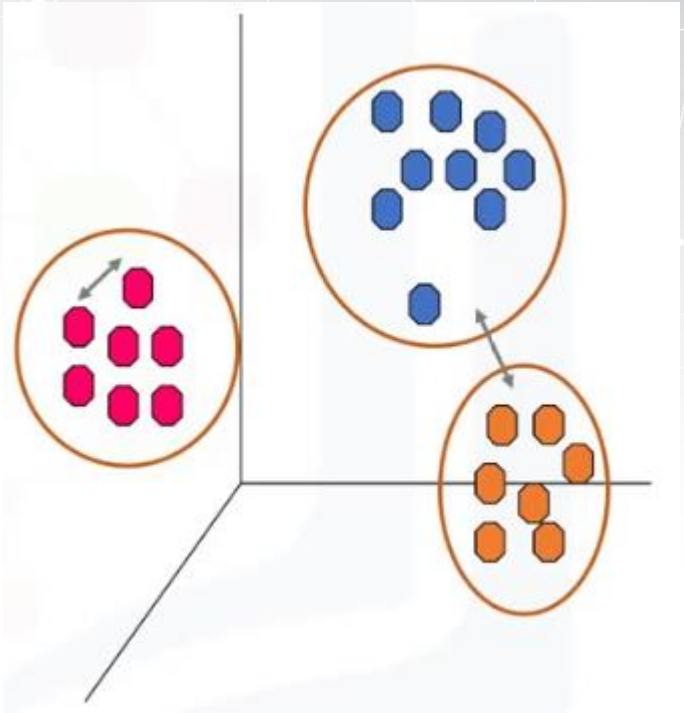


k-Means clustering algorithm

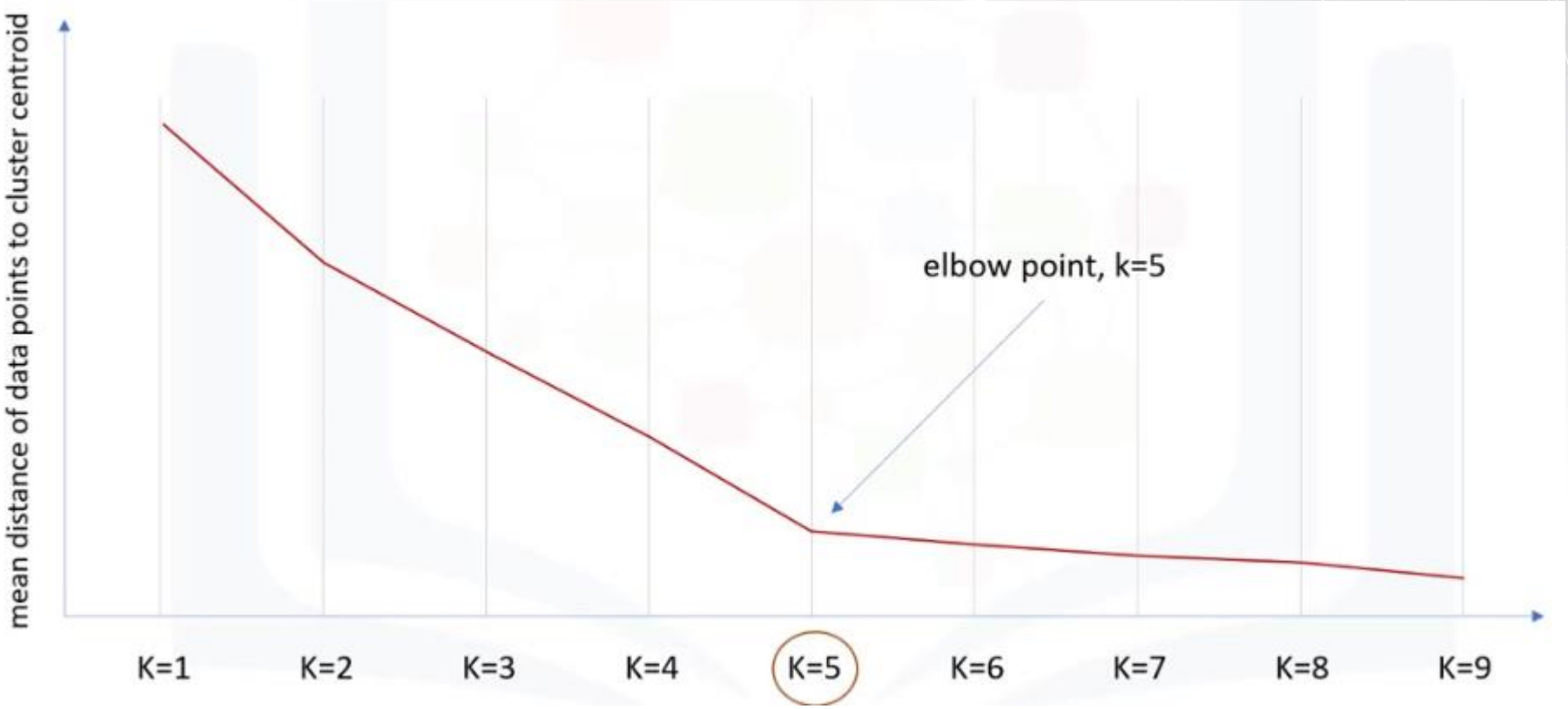
1. Randomly placing k centroids, one for each cluster
2. Calculate the distance of each point from each centroid
3. Assign each data point (object) to its closest centroid, creating cluster
4. Recalculate the position of the k centroids
5. Repeat the steps 2-4, until the centroids no longer move

k-Means accuracy

- External approach
  - Compare the clusters with the ground truth, if it is available
- Internal approach
  - Average the distance between data points within a cluster



Choosing k


mean distance of data points to cluster centroid
elbow point, k=5
K=1
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9

k-Means recap

- Med and Large sized databases (Relatively efficient)
- Produces sphere-like clusters
- Needs number of clusters (k)

IKUTI KAMI

DIGITAL TALENT SCHOLARSHIP

digitalent.kominfo
digitalent.kominfo
DTS_kominfo
Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110