AI VIETNAM
All-in-One Course

Machine Learning

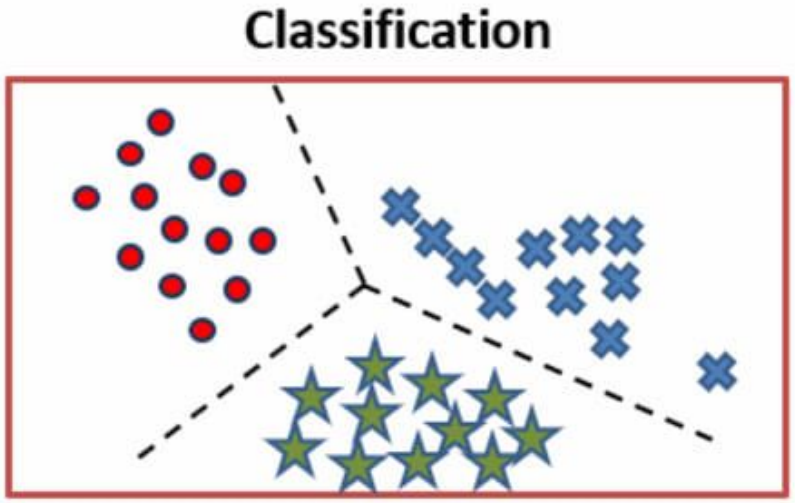
K-NEAREST NEIGHBORS

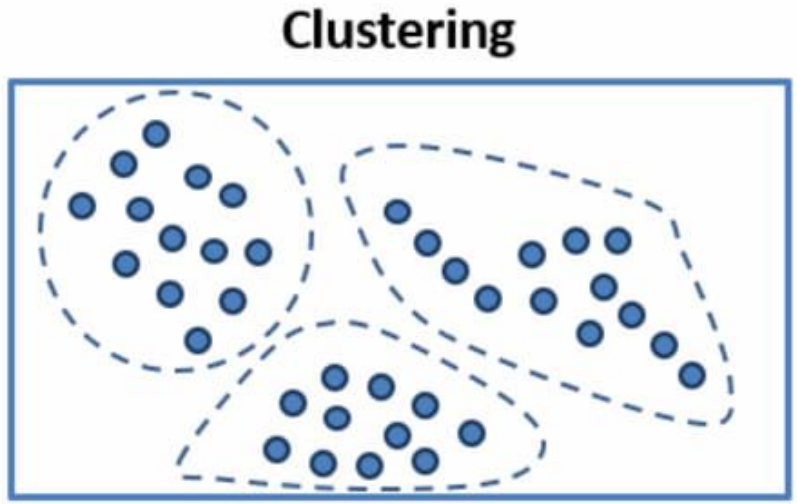
K-MEANS CLUSTERING

Nguyen Quoc Thai - Nguyen Tho Anh Khoa - Nguyen Dang Nha

Year 2024

Objectives





- ✔ Introduce and explain the basic principles of the KNN
- ✔ Demonstrate the application of KNN in classification tasks
- ✔ Illustrate how KNN can be utilized for regression problems
- ✔ Showcase examples of KNN
- ✔ Discuss various applications of KNN

- ✔ Define and describe the K-Means clustering algorithm
- ✔ Explain the operational steps of the K-Means algorithm
- ✔ Analyze the application of K-Means on the Iris dataset
- ✔ Explore the broad usage of K-Means.

Outline

- 1. K Nearest Neighbors
- 2. K-Mean Clustering
- 3. Summary

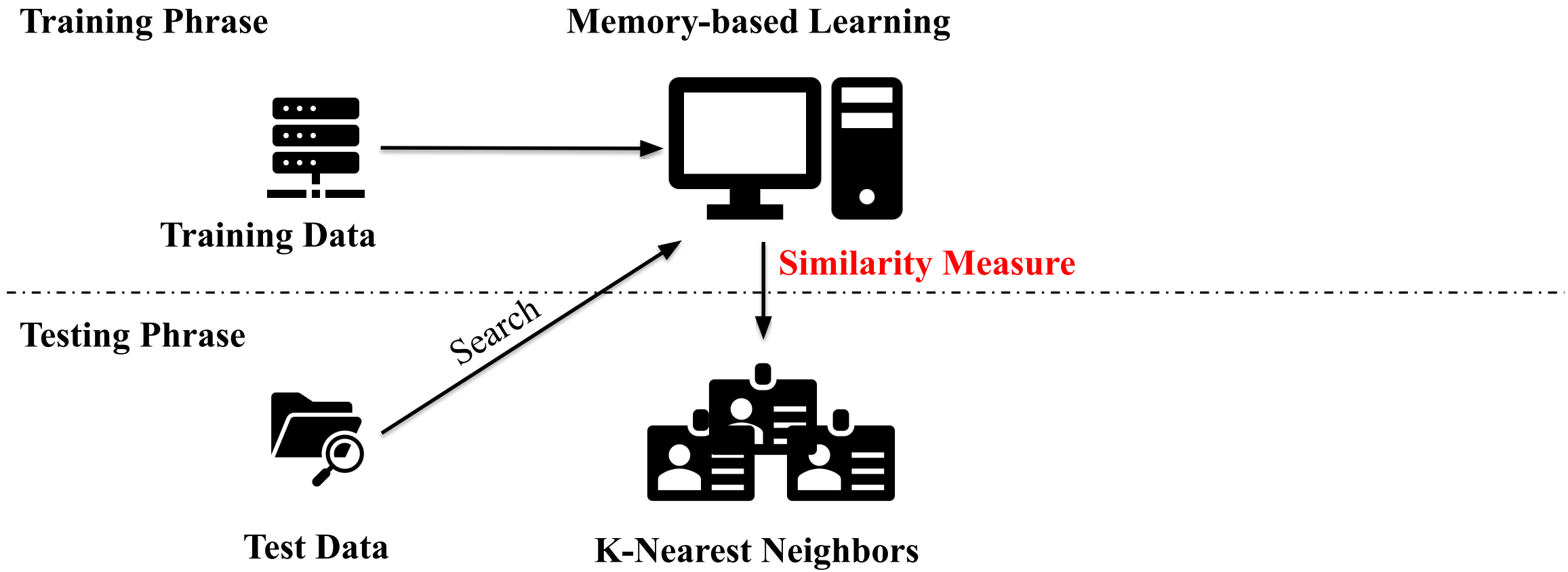
1 – K Nearest Neighbors

What is the KNN Algorithm?

- KNN is one of the simplest supervised machine learning algorithms
- Lazy Learning:
  - Does not “LEARN” until the test example is given
  - A new data is predicted based on K-Nearest Neighbors from the training data

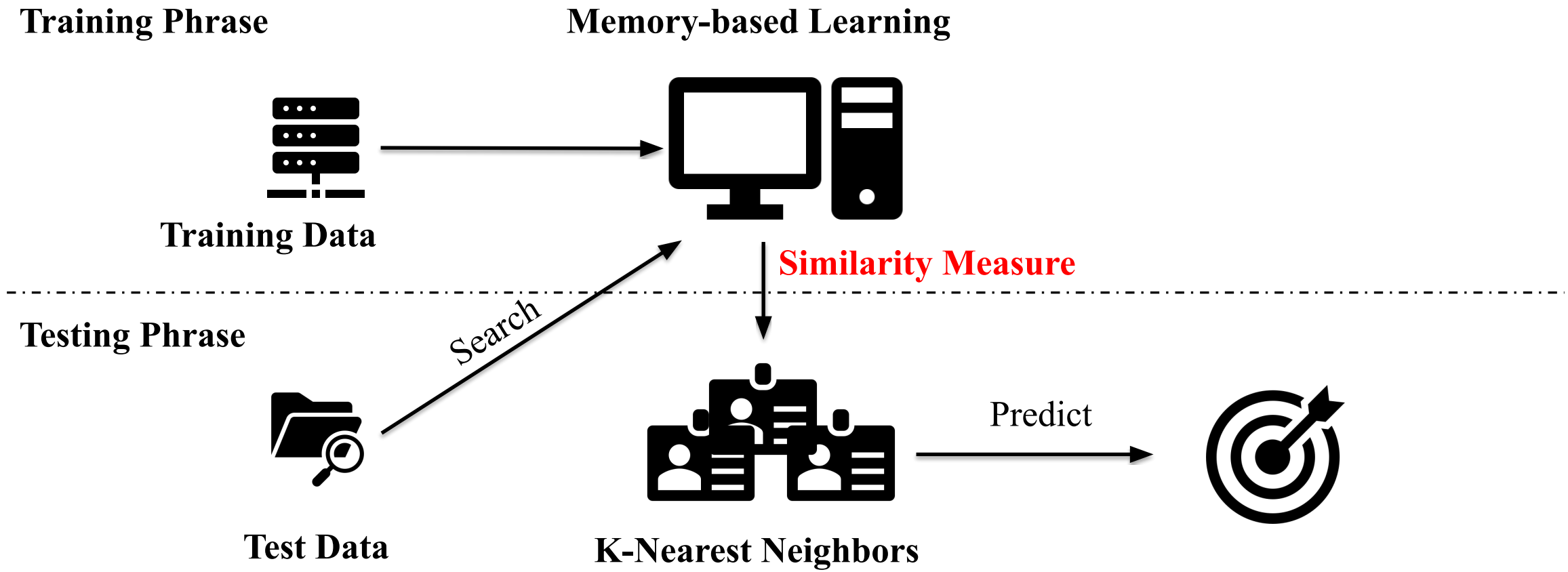
1 – K Nearest Neighbors

What is the KNN Algorithm?



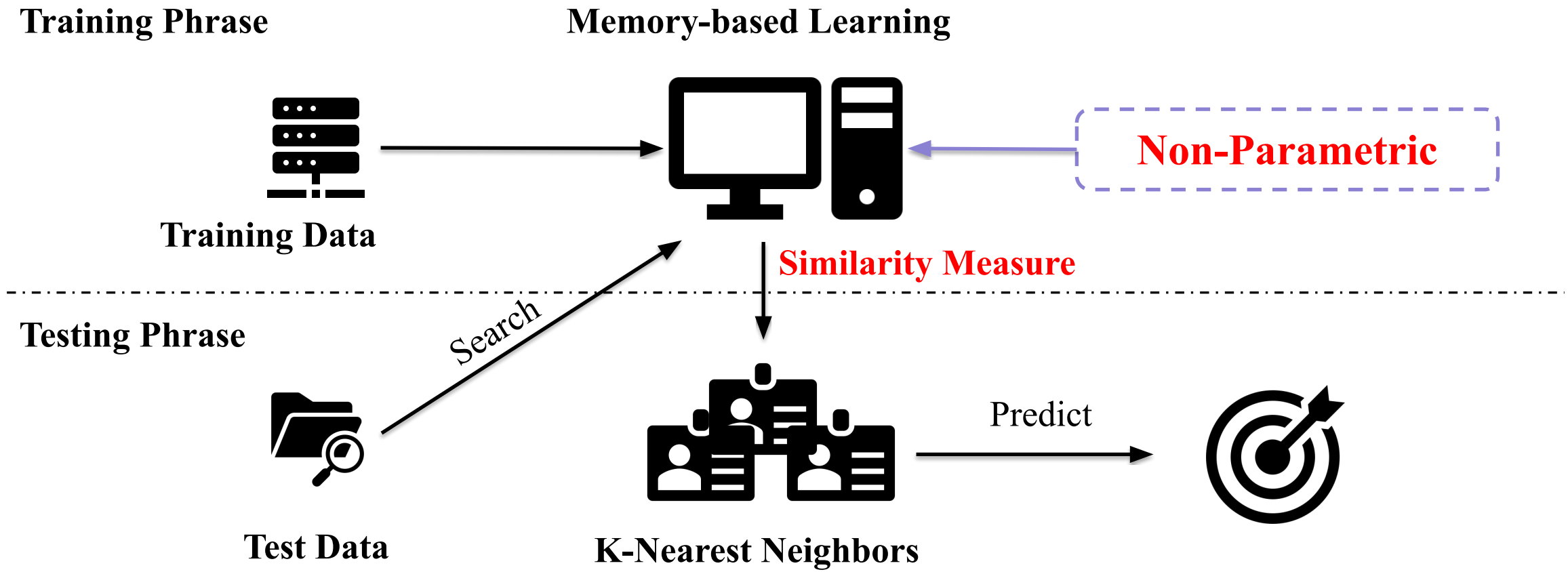
1 – K Nearest Neighbors

What is the KNN Algorithm?



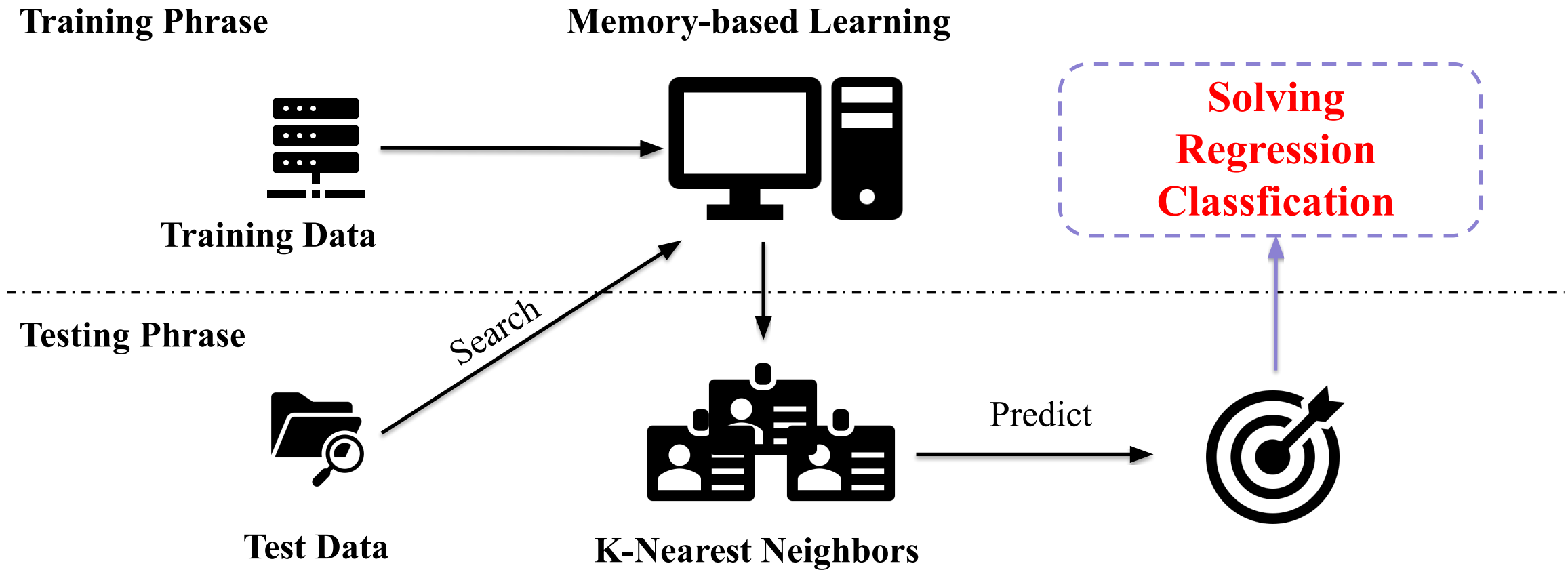
1 – K Nearest Neighbors

What is the KNN Algorithm?



1 – K Nearest Neighbors

What is the KNN Algorithm?



1 – K Nearest Neighbors

What is the KNN Algorithm?

Regression

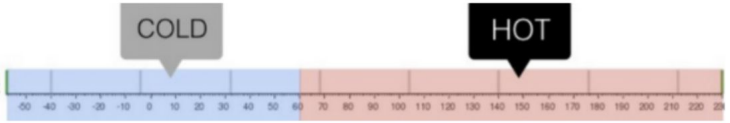
- Predict a continuous value based on the input variables

What will be the temperature tomorrow?

Classification

- Classify input variables to identify discrete output variables (labels, categories)

Will it be hot or cold tomorrow?



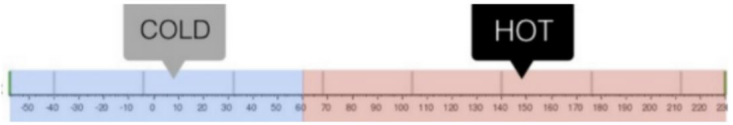
1 – K Nearest Neighbors

KNN: Classification Approach

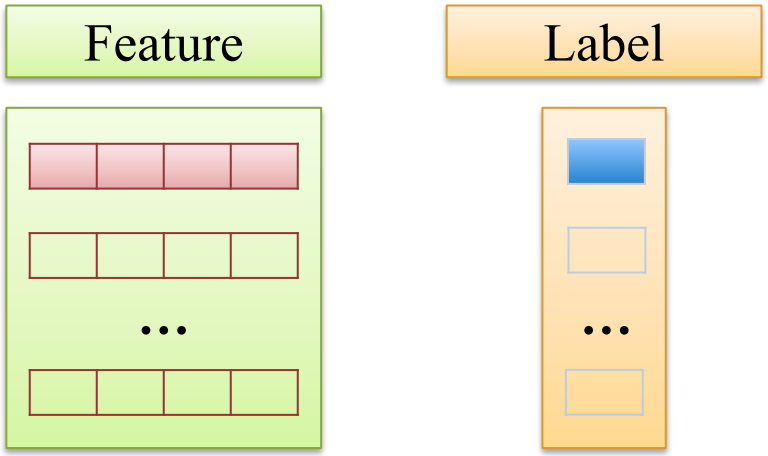
Classification

- Classify input variables to identify discrete output variables (labels, categories)

Will it be hot or cold tomorrow?



Training Data



Test Data

1 – K Nearest Neighbors

KNN: Classification Approach

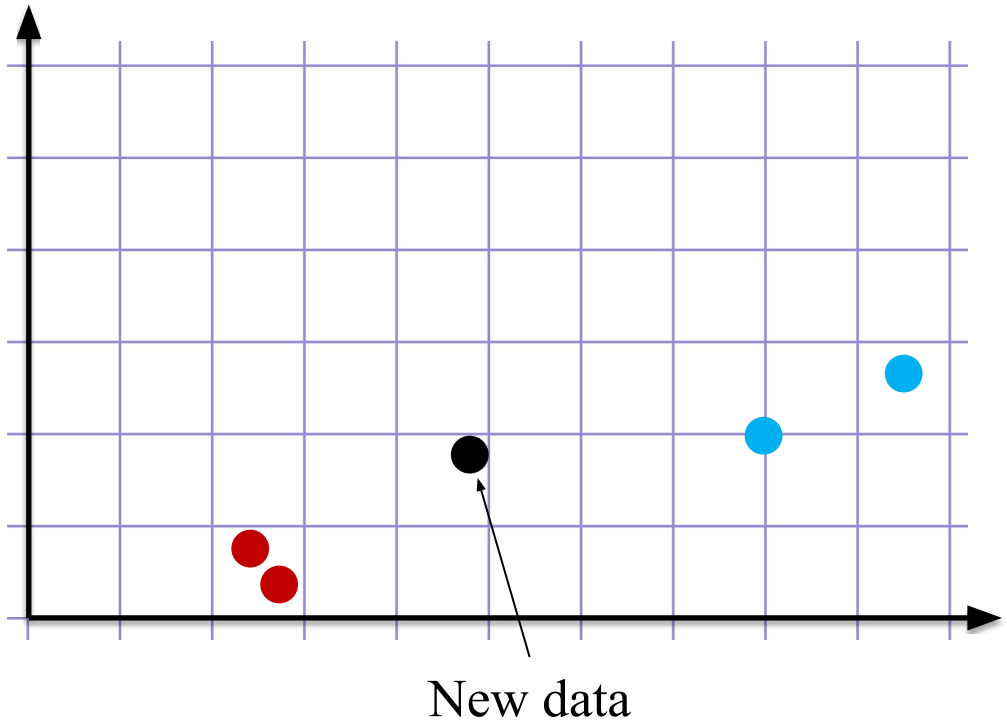
Step 1: Look at the data

Training Data

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.3	0.4	0
4	1	1
4.7	1.4	1

Test Data

2.4	0.8



1 – K Nearest Neighbors

KNN: Classification Approach

Step 2: Calculate distances

Training Data

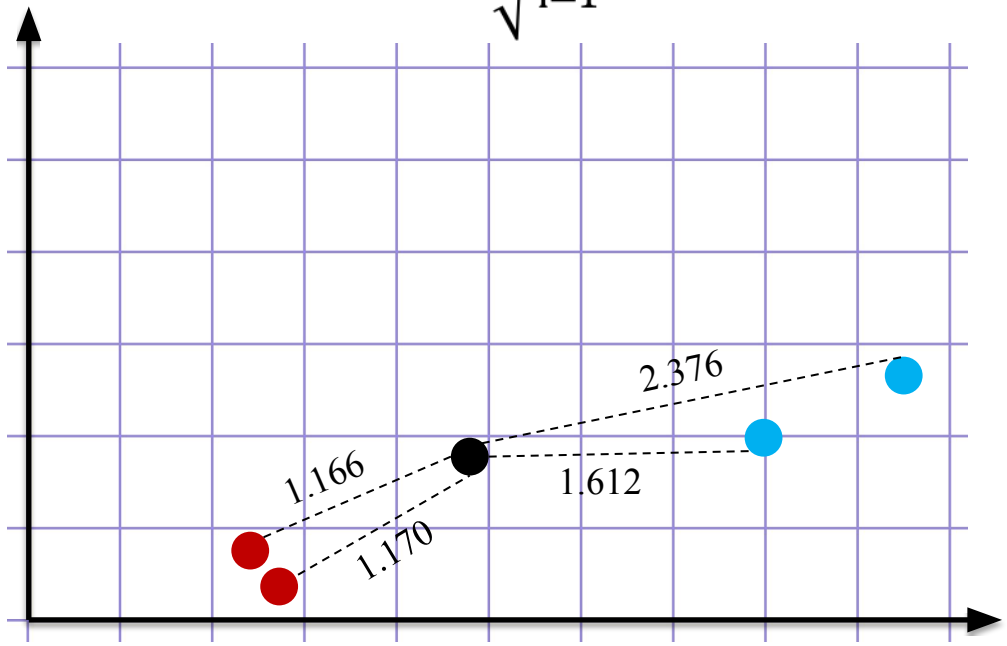
Petal_Length	Petal_Width	Label	Distance
1.4	0.2	0	1.166
1.3	0.4	0	1.170
4	1	1	1.612
4.7	1.4	1	2.376

Test Data

2.4	0.8

Euclidean Distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



1 – K Nearest Neighbors

KNN: Classification Approach

Step 3: Find neighbors

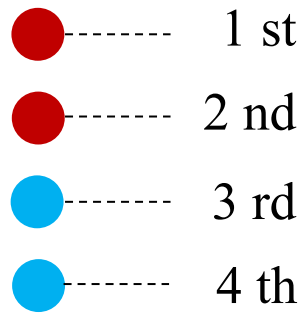
Training Data

Petal_Length	Petal_Width	Label	Distance
1.4	0.2	0	1.166
1.3	0.4	0	1.170
4	1	1	1.612
4.7	1.4	1	2.376

Test Data

2.4	0.8

Ranking points



Find the nearest neighbors by ranking points by increasing distance

1 – K Nearest Neighbors

KNN: Classification Approach

Step 4: Vote on labels

Training Data

Petal_Length	Petal_Width	Label	Distance
1.4	0.2	0	1.166
1.3	0.4	0	1.170
4	1	1	1.612
4.7	1.4	1	2.376

Test Data

2.4	0.8	→	0

K=3 Nearest neighbours

- 1 st
- 2 nd
- 3 rd

# of votes

- 2
- 1

Vote on the predicted class labels based on the class of the k nearest neighbors

1 – K Nearest Neighbors

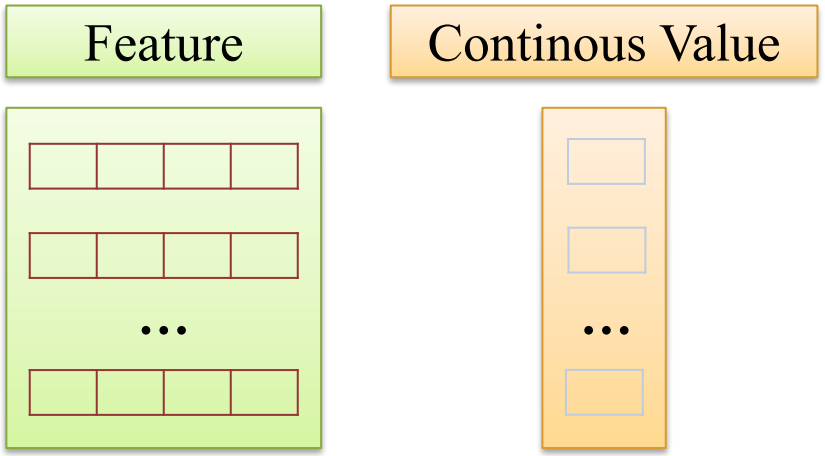
KNN: Regression Approach

Regression

- Predict a continuous value based on the input variables

What will be the temperature tomorrow?

Training Data



Test Data

1 – K Nearest Neighbors

KNN: Regression Approach

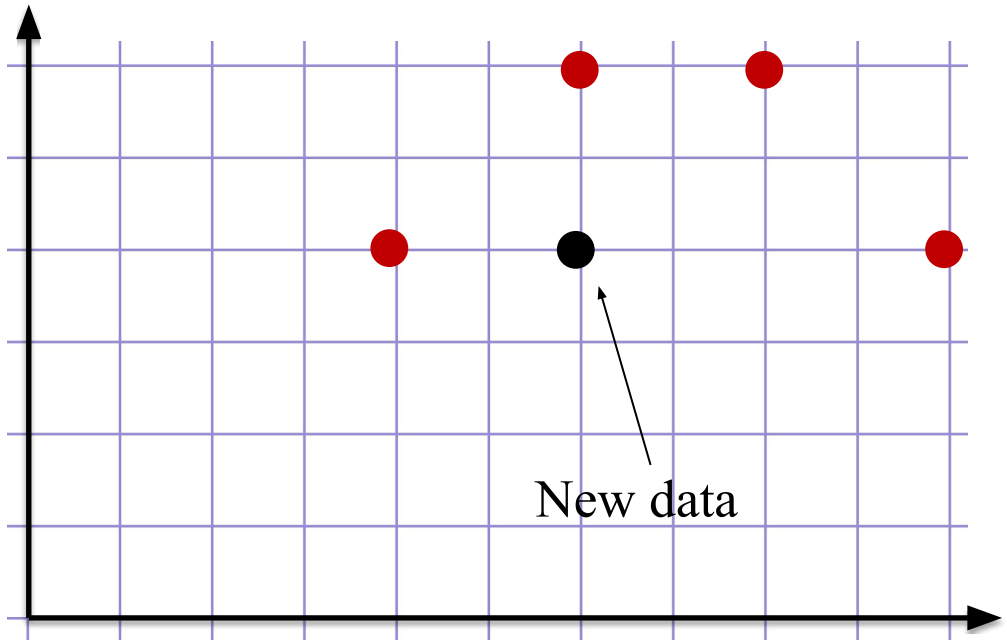
Step 1: Look at the data

Training Data

Length	Width	Price
2.0	2.0	2.0
3.0	3.0	2.5
4.0	3.0	3.5
5.0	2.0	5.0

Test Data

3.0	2.0



1 – K Nearest Neighbors

KNN: Regression Approach

Step 2: Calculate distances

Training Data

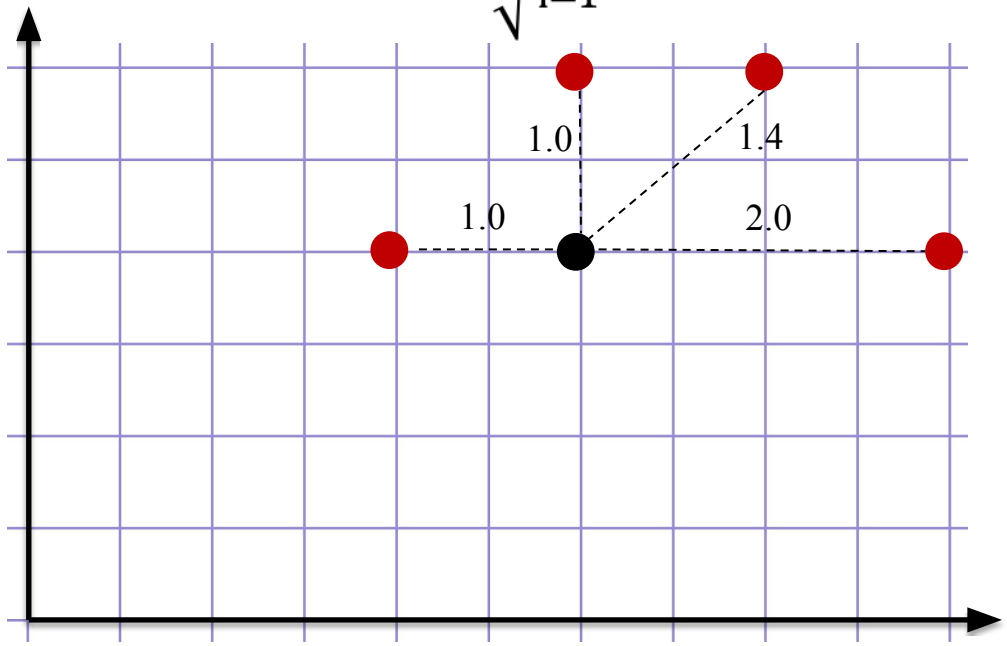
Length	Width	Price	Distance
2.0	2.0	2.0	1.0
3.0	3.0	2.5	1.0
4.0	3.0	3.5	1.4
5.0	2.0	5.0	2.0

Test Data

3.0	2.0

Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



1 – K Nearest Neighbors

KNN: Regression Approach

Step 3: Find neighbors

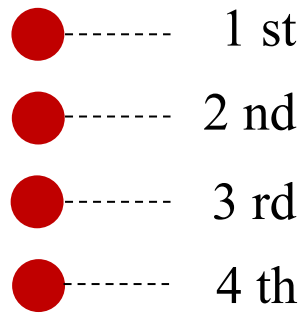
Training Data

Length	Width	Price	Distance
2.0	2.0	2.0	1.0
3.0	3.0	2.5	1.0
4.0	3.0	3.5	1.4
5.0	2.0	5.0	2.0

Test Data

3.0	2.0

Ranking points



Find the nearest neighbours by ranking points by increasing distance

1 – K Nearest Neighbors

KNN: Regression Approach

Step 4: Compute the mean value

Training Data

Length	Width	Price	Distance
2.0	2.0	2.0	1.0
3.0	3.0	2.5	1.0
4.0	3.0	3.5	1.4
5.0	2.0	5.0	2.0

Test Data

3.0	2.0	→	2.67

$$Y_{pred} = \frac{1}{k} \sum_{x \in NB} y_x$$

K=3 Nearest neighbours

- 1 st
- 2 nd
- 3 rd
- 4 th

$$Y_{pred} = \frac{1}{3}(2.0 + 2.5 + 3.5) = 2.67$$

Compute the mean value of the k nearest neighbors

1 – K Nearest Neighbors

KNN: Summary

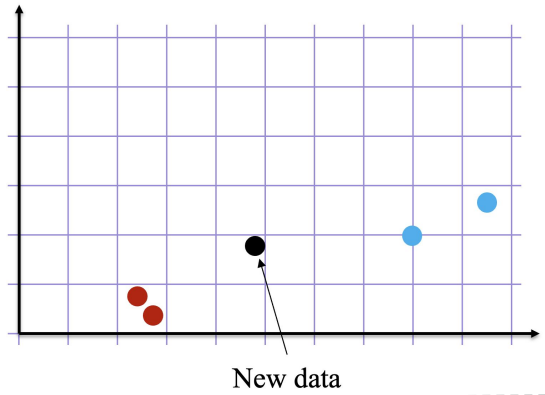
Step 1: Look at the data

Step 2: Calculate distances

Step 3: Find neighbours

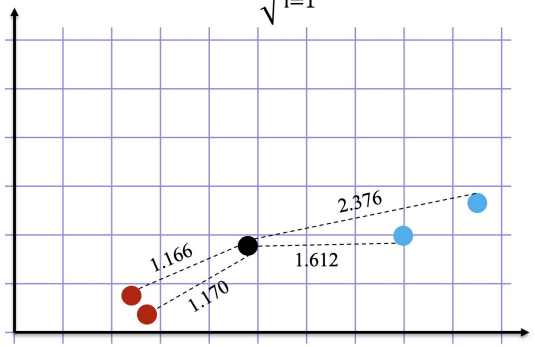
Step 4: Vote on labels

Classification



Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Ranking points

- 1 st
- 2 nd
- 3 rd
- 4 th

Find the nearest neighbors by ranking points by increasing distance

K=3 Nearest neighbours

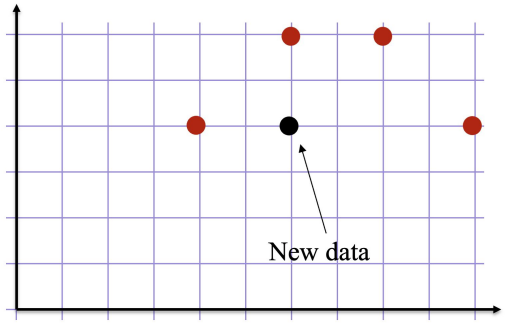
- 1 st
- 2 nd
- 3 rd

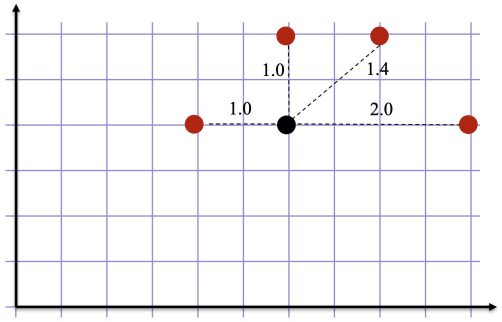
# of votes

- 2
- 1

Vote on the predicted class labels based on the class of the k nearest neighbors

Regression





Ranking points

- 1 st
- 2 nd
- 3 rd
- 4 th

Find the nearest neighbours by ranking points by increasing distance

K=3 Nearest neighbours

- 1 st
- 2 nd
- 3 rd
- 4 th

$$Y_{pred} = \frac{1}{k} \sum_{x \in NB} y_x$$

Compute the mean value of the k nearest neighbors

1 – K Nearest Neighbors

Geometry Distance Functions

- **Euclidean (p=2)**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Manhattan (p=1)**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Minkowski (p=norm)**

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- **Chebyshev (p=∞)**

$$d(x, y) = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_i |x_i - y_i|$$

1 – K Nearest Neighbors

Feature Scaling (Normalization)

- Standardize the range of independent variables (feature of data)

Training Data

Strong Influence

Petal_Length	L_Distance	Petal_Width	W_Distance	Label	Distance	Rank
1.4	1.8	0.2	0.4	0	1.844	3
1.3	1.9	0.4	0.2	0	1.910	4
4	0.8	1	0.4	1	0.894	1
4.7	1.3	1.4	0.8	1	1.526	2

Test Data

3.2	0.6

1 – K Nearest Neighbors

Feature Scaling (Normalization)

- Standardize the range of independent variables (feature of data)

Training Data

MinMaxScaler Normalization

Petal_Length	L_Distance	Petal_Width	W_Distance	Label	Distance	Rank
0.03	0.53	0.00	0.33	0	0.624	3
0.00	0.56	0.17	0.16	0	0.582	2
0.79	0.23	0.66	0.33	1	0.402	1
1.00	0.44	1.00	0.67	1	0.801	4

Test Data

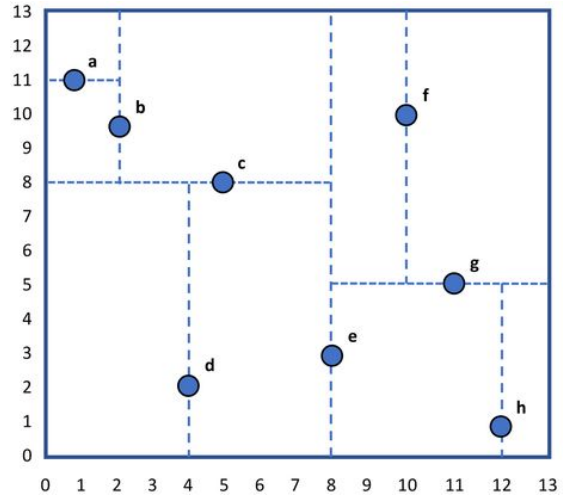
0.56	0.33

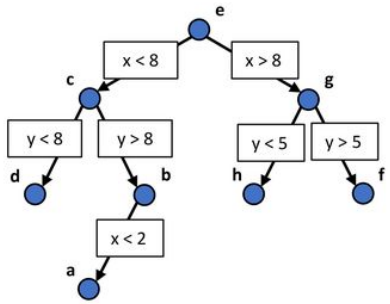
1 – K Nearest Neighbors

Searching in KNN

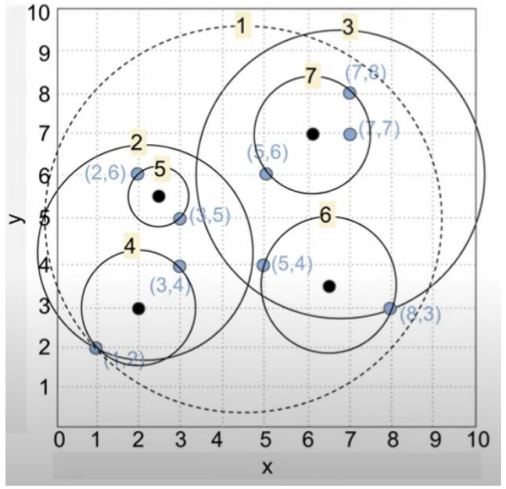
- Training dataset: N samples in D dimensions
- Brute Force: Naïve neighbor search – $O[DN^2]$

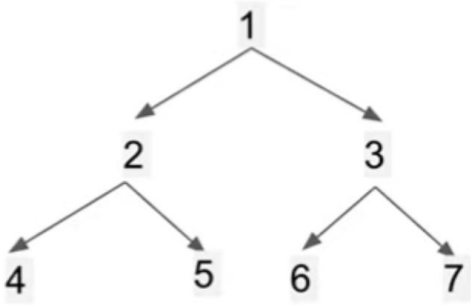
1 KD Tree O[NDlog(N)]





2 Ball Tree



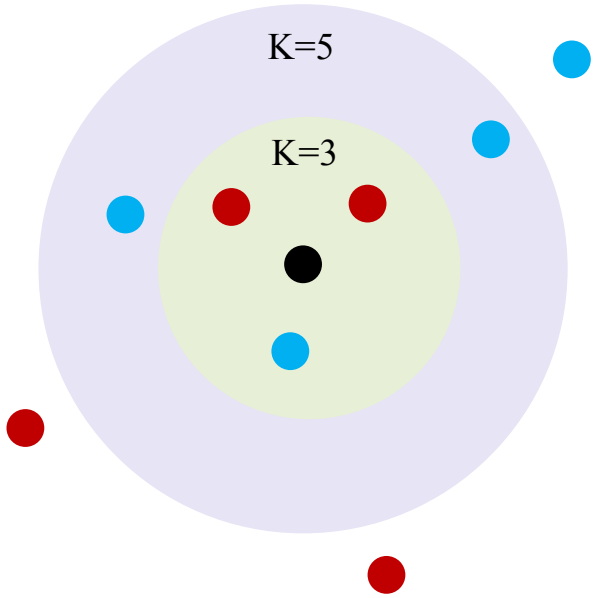


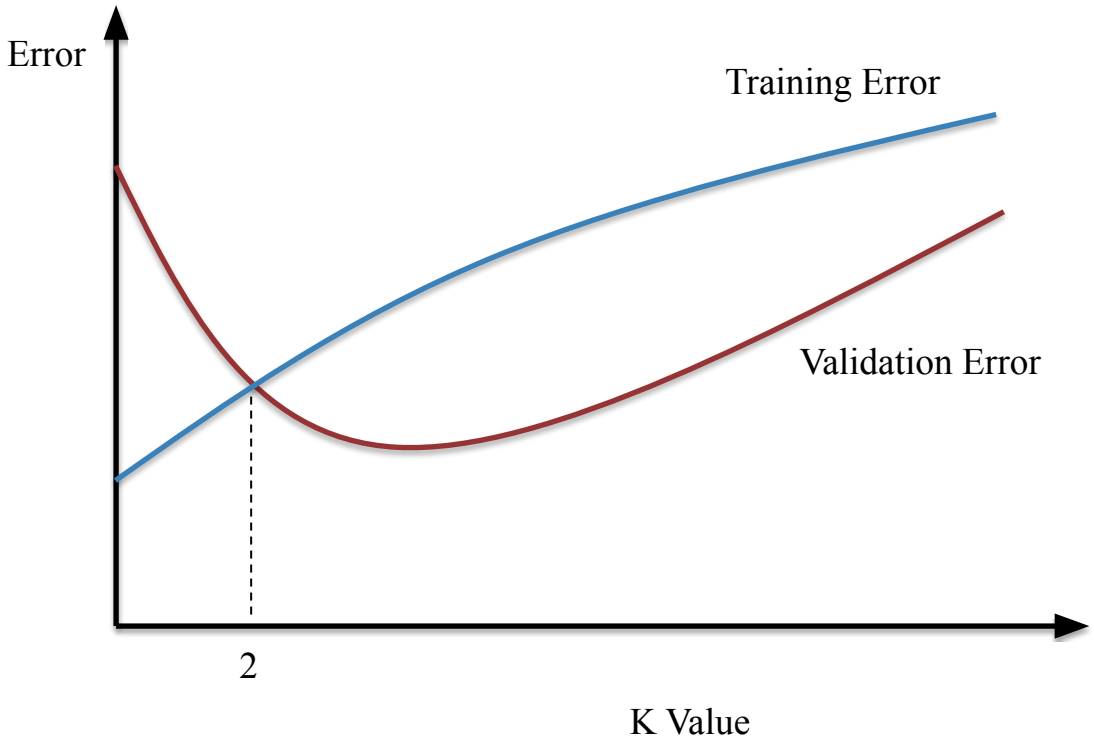
Source: https://varshasaini.in/kd-tree-and-ball-tree-knn-algorithm/

1 – K Nearest Neighbors

How to find the optimal value of K in KNN?

- Choose K based on the evaluation on the validation set (Accuracy, Error, F-Score,…)

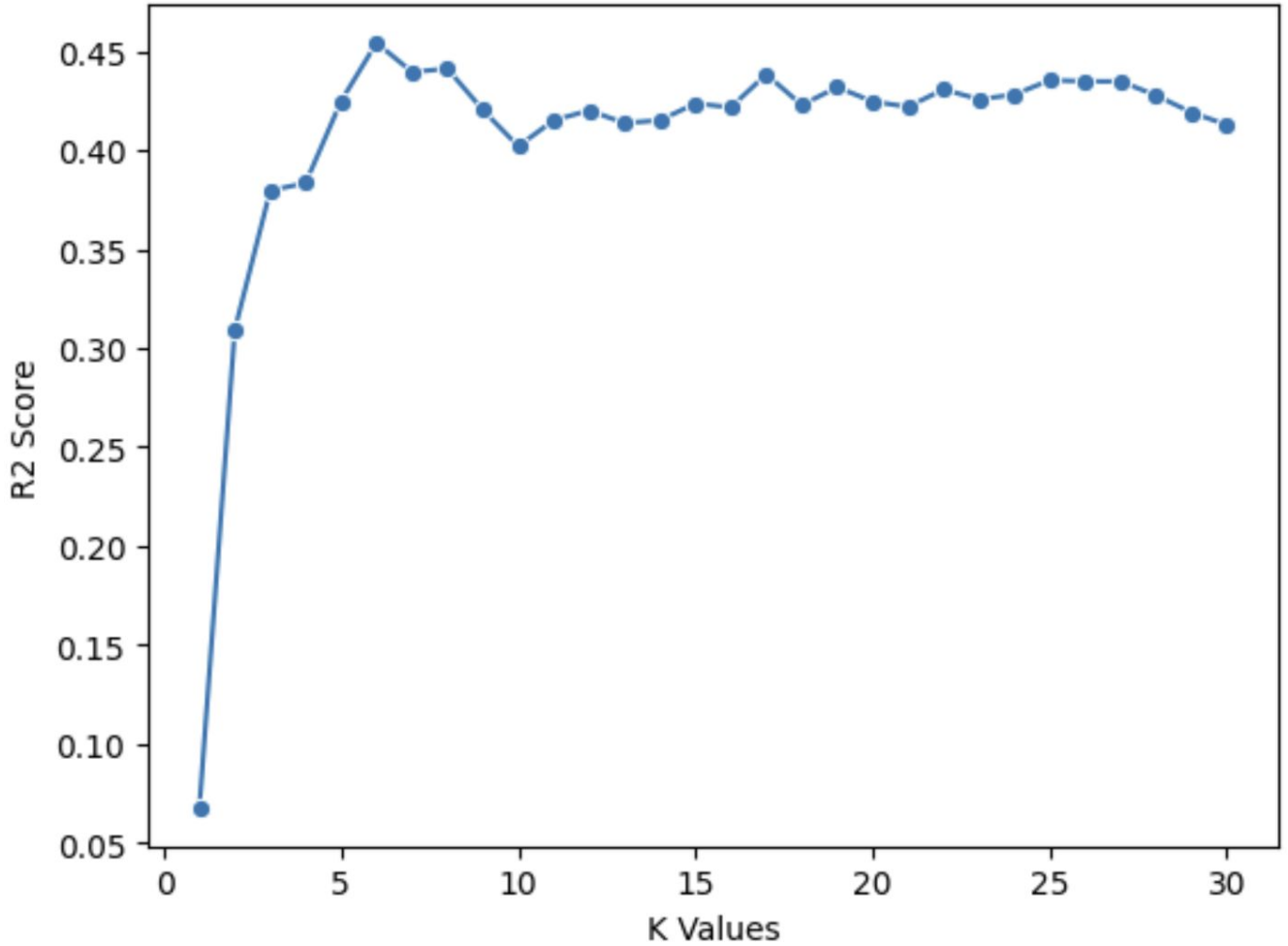




1 – KNN Applications

KNN for Regression: "Diabetes" Dataset

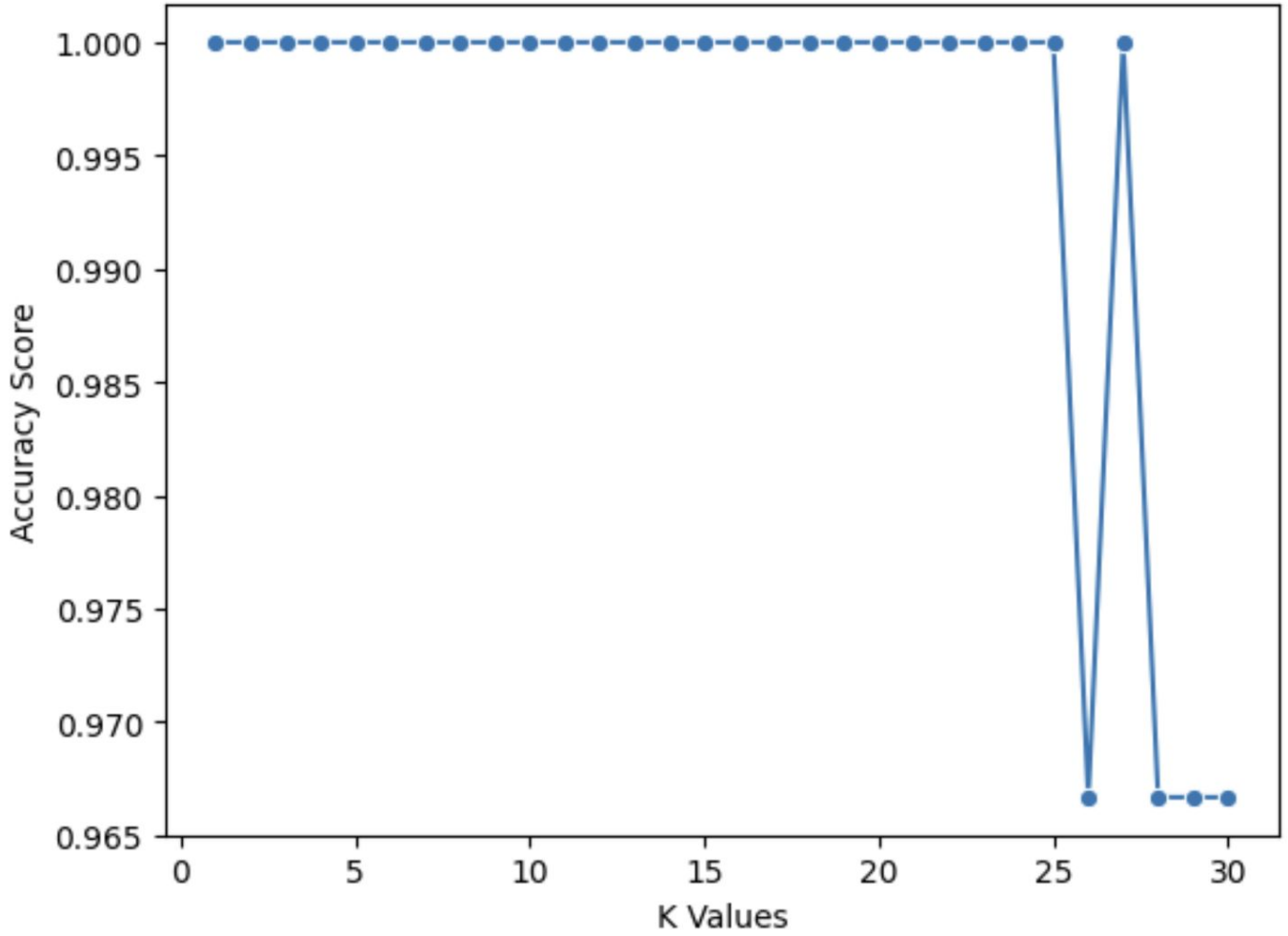
- Sample: 442
- Features: 10
- Target: 25 – 346
- R2-Score (Validation)



1 – KNN Applications

KNN for Classification: “Iris” Dataset

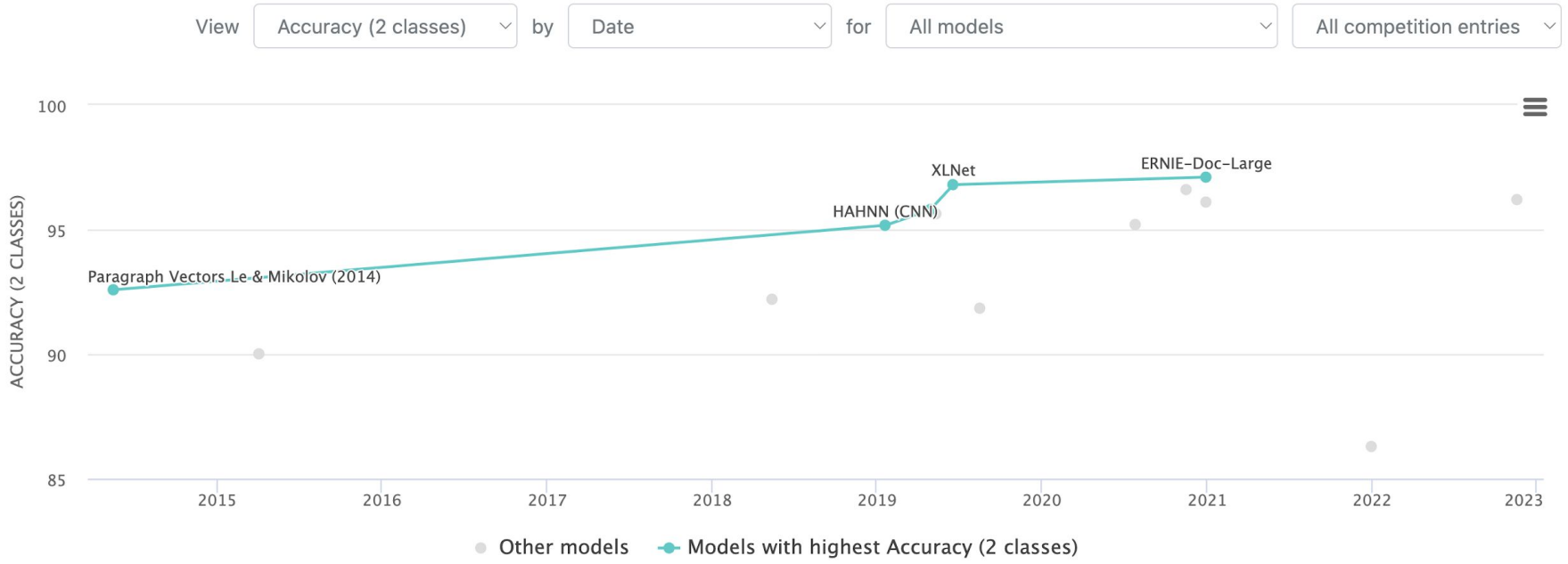
- Sample: 150
- Features: 4
- Classes: 3 (50 per class)
- Accuracy (Validation)



1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

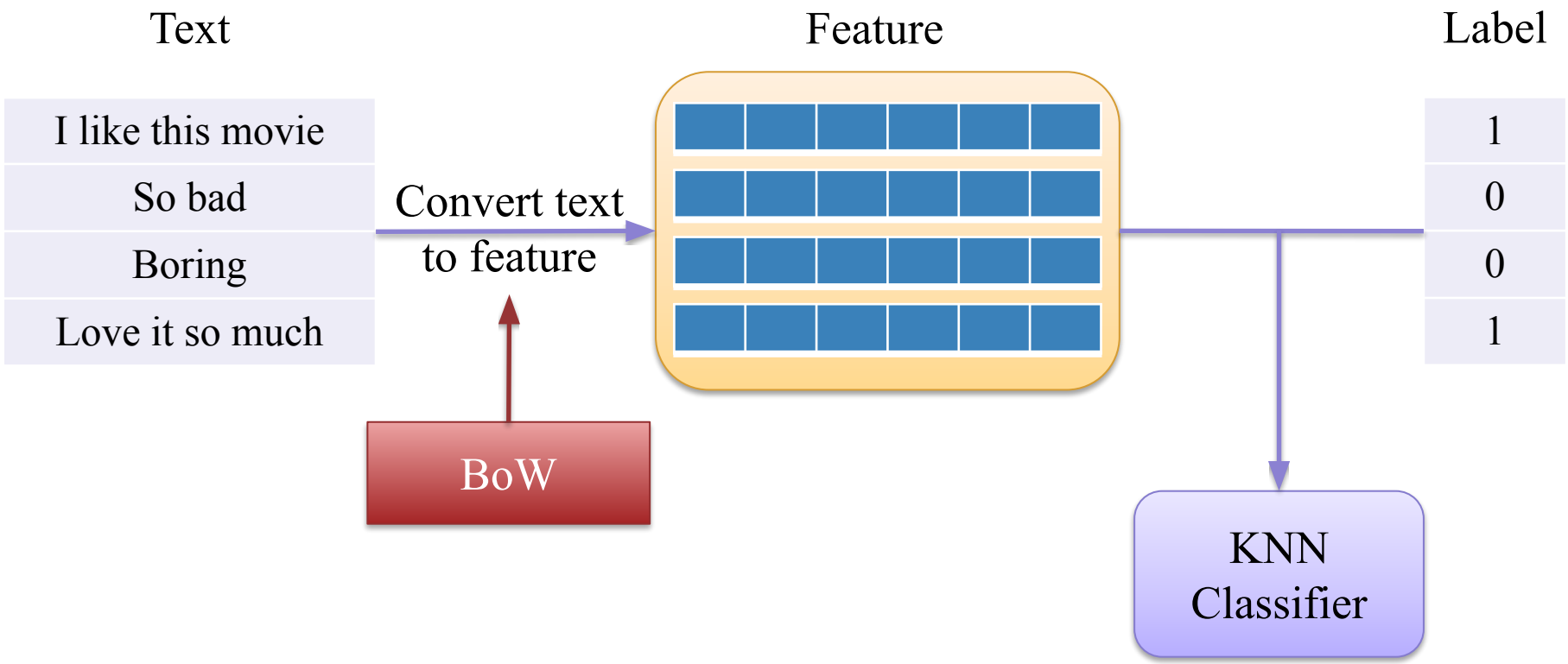
- Sample: 50.000 movie review
- Classes: 2 – Positive and Negative (25.000 per class)
- Accuracy, F1-Score



https://paperswithcode.com/sota/text-classification-on-imdb

1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

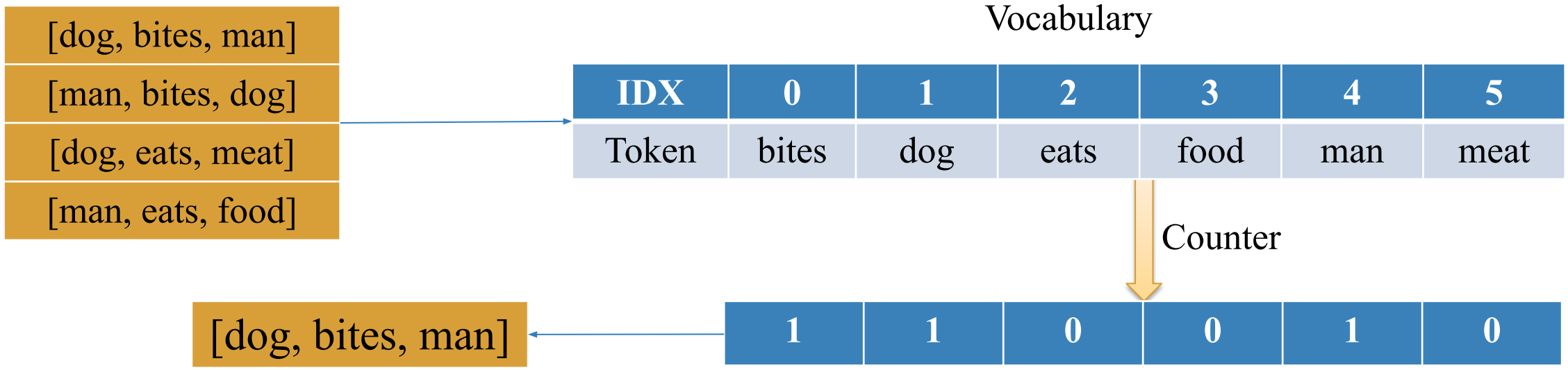


1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

- **Bag of Words (BoW)**
- **Document-Level:** Consider text as a bag (collection) of words
- **Represented by a V-dimensional**

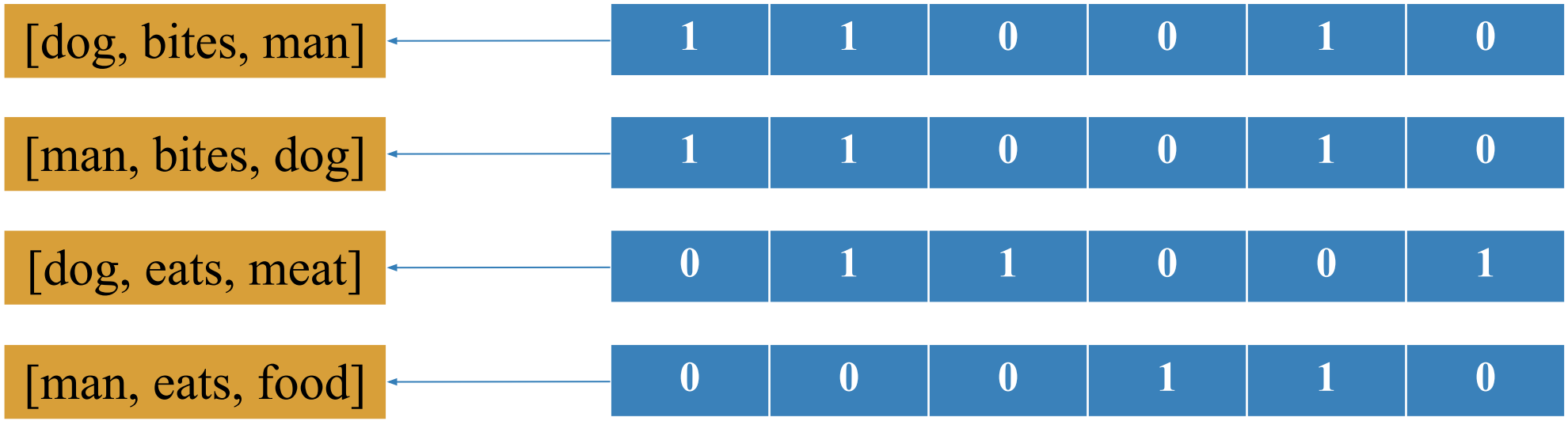
Use: the number of occurrences of the word in the document



1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

- **Bag of Words (BoW)**
- **Document-Level:** Consider text as a bag (collection) of words
- **Represented by a V-dimensional**



1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

- **Bag of Words (BoW)**
- **Out of vocabulary (OOV)**

Vocabulary

IDX	0	1	2	3	4	5
Token	bites	dog	eats	food	man	meat

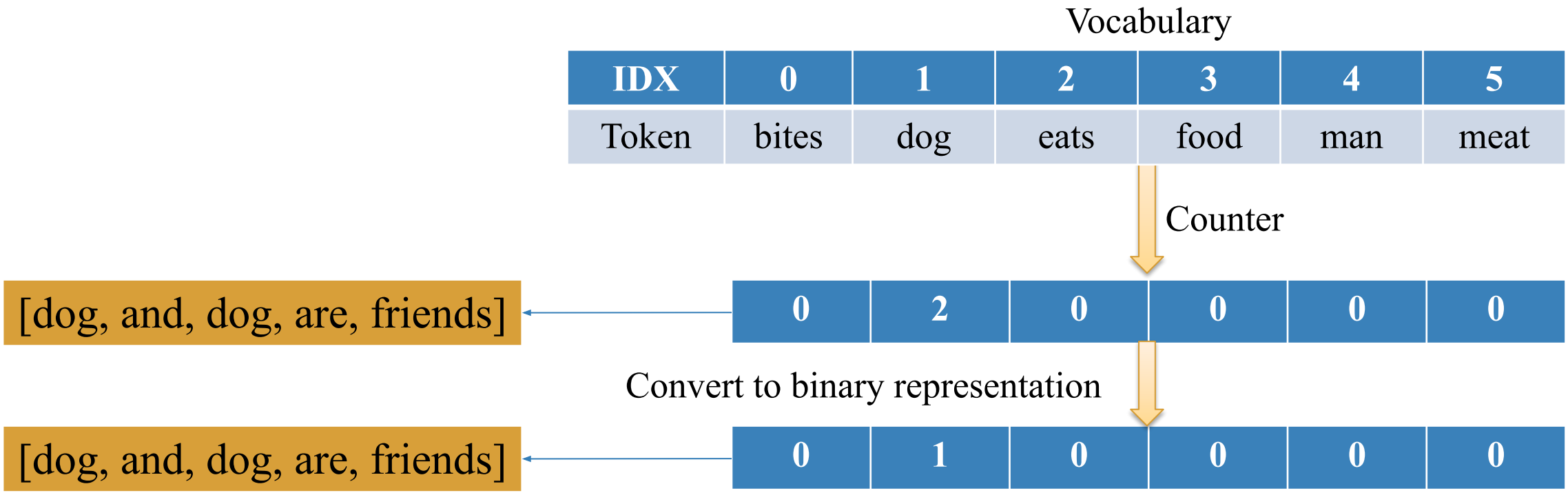
Counter

[dog, bites, man]	1	1	0	0	1	0
[dog, and, dog, are, friends]	0	2	0	0	0	0

1 – KNN Applications

KNN for Text Classification: "IMDB" Dataset

- **Bag of Words (BoW)**
- Representation without considering frequency (Binary Representation)



1 – Quiz

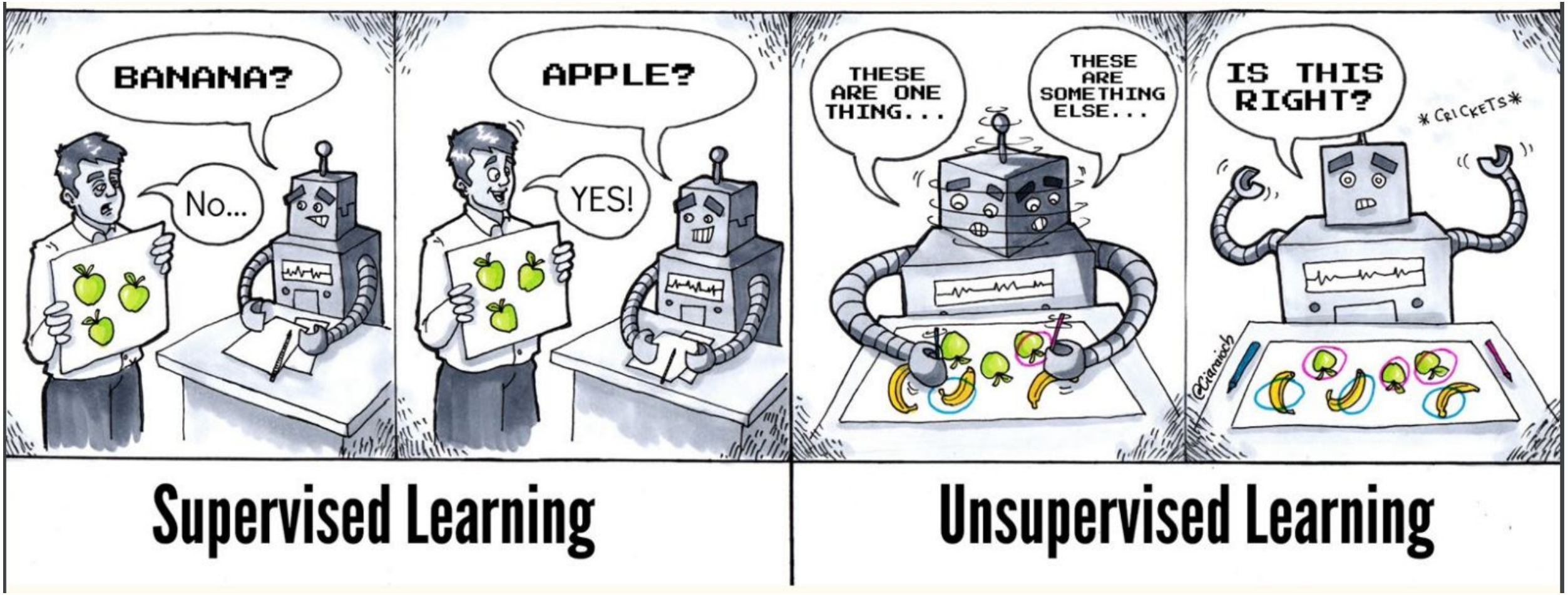
Một nhược điểm lớn của KNN là gì?

A) Cần nhiều dữ liệu huấn luyện	B) Không thể áp dụng cho dữ liệu phân loại
C) Không thể sử dụng khoảng cách Euclidean	D) Tốc độ dự đoán chậm nếu dữ liệu lớn

AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

What is K-Means Clustering?



2– K-means clustering

What is K-Means Clustering?

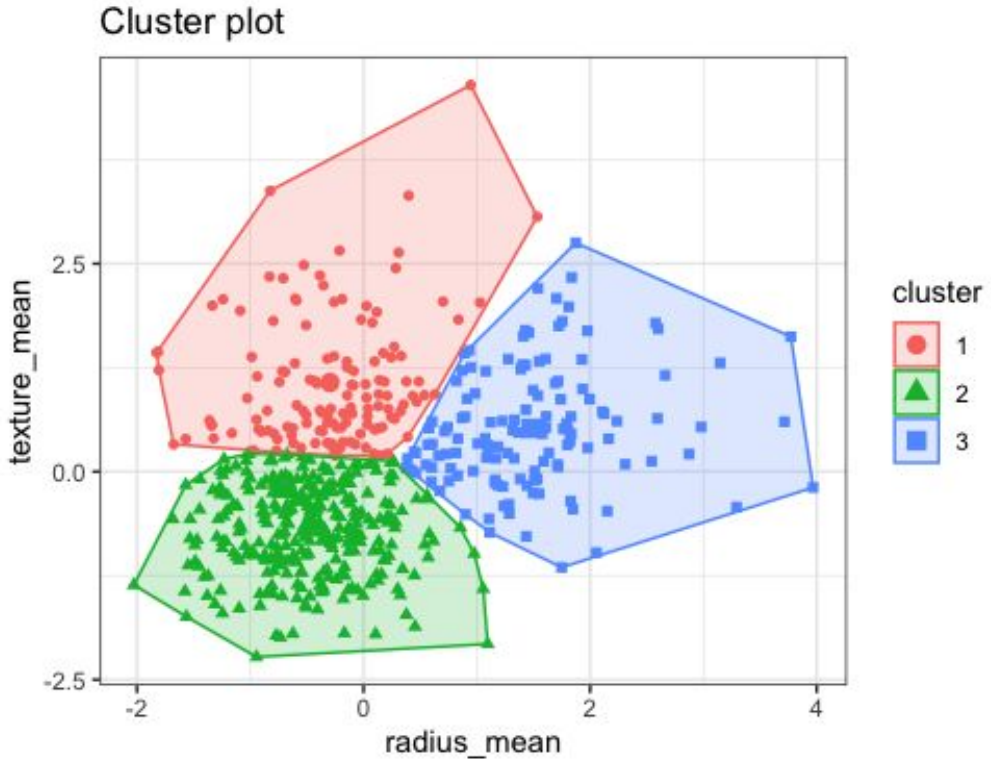
- Clustering is an **unsupervised machine learning technique** designed to **group unlabeled examples** based on their similarity to each other. (If the examples are labeled, this kind of grouping is called classification.)
- Key Concepts:
  - **Similarity Measurement:** Clustering algorithms group data points based on similarity metrics. These metrics, such as Euclidean distance for numerical data or cosine similarity for text data, determine how "close" or "similar" data points are to one another
  - **Grouping:** The primary goal is to group data points so that those within the same cluster are more similar to each other than to those in other clusters. This is done without prior knowledge of the group labels.

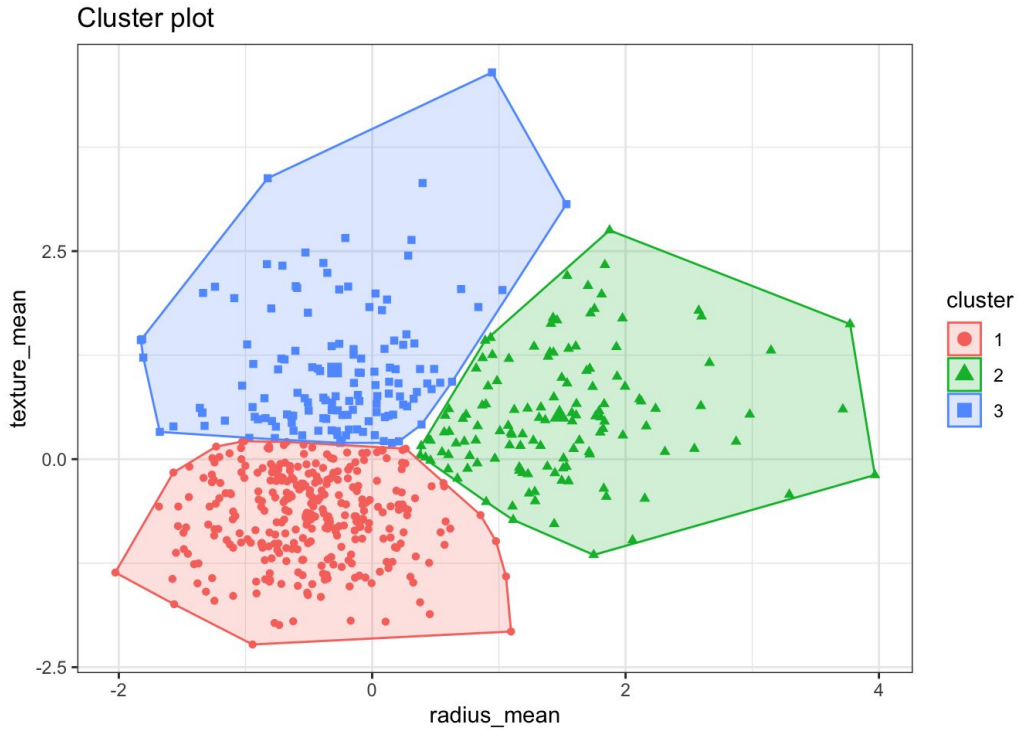
AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

What is K-Means Clustering?

- Clustering is an **unsupervised machine learning technique** designed to **group unlabeled examples** based on their similarity to each other. (If the examples are labeled, this kind of grouping is called classification.)





AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

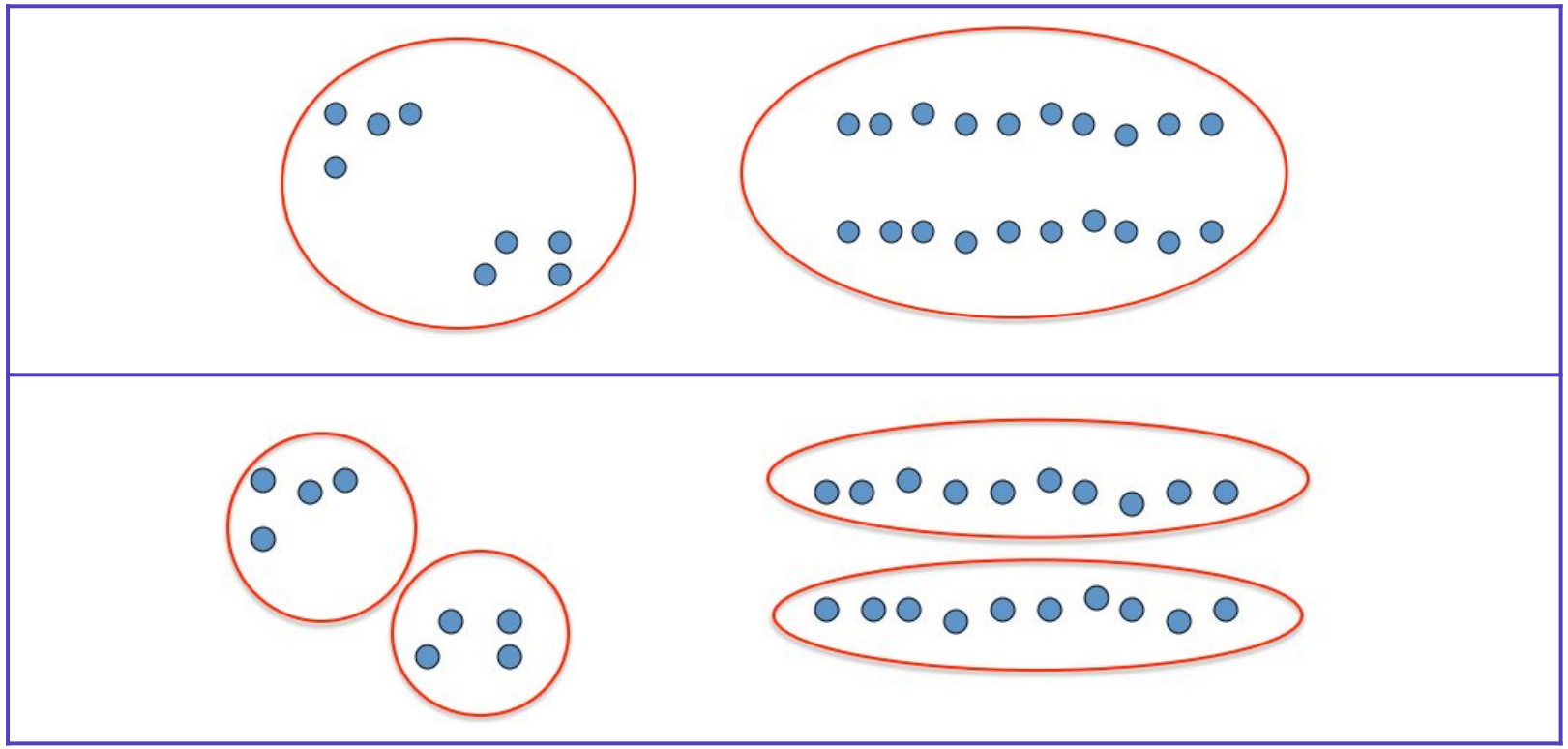
What is K-Means Clustering?

- Clustering is an **unsupervised machine learning technique** designed to **group unlabeled examples** based on their similarity to each other. (If the examples are labeled, this kind of grouping is called classification.)
- Clustering vs. Classification:
  - **Unsupervised vs. Supervised:** Clustering is unsupervised and does not use output labels in training. Classification is supervised and relies on pre-labeled data to learn the mapping from inputs to outputs.
  - **Discovery vs. Prediction:** Clustering is used to discover the inherent groupings in the data, such as grouping customers by purchasing behavior. Classification predicts the category of new observations based on past data.

2– K-means clustering

What is K-Means Clustering?

- Clustering is an **unsupervised machine learning technique** designed to **group unlabeled examples** based on their similarity to each other. (If the examples are labeled, this kind of grouping is called classification.)



2– K-means clustering

What is K-Means Clustering?

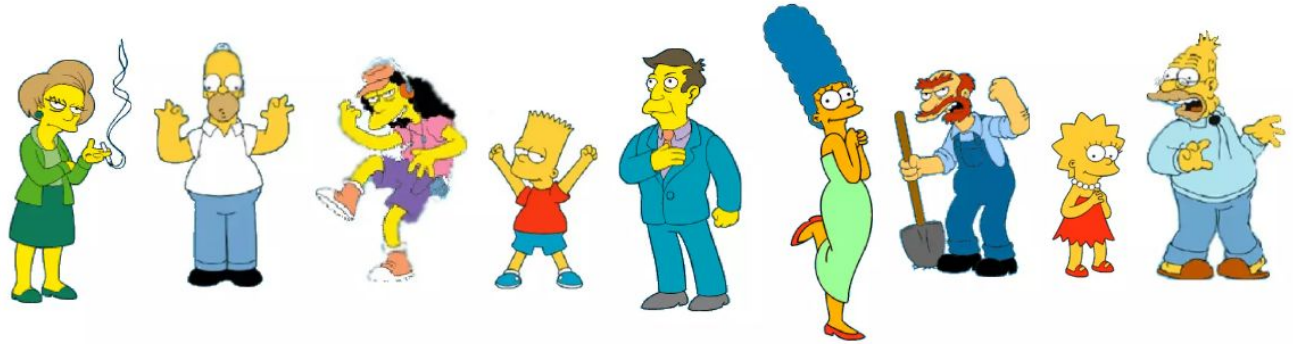
- Clustering is an **unsupervised machine learning technique** designed to **group unlabeled examples** based on their similarity to each other. (If the examples are labeled, this kind of grouping is called classification.)
- Applications of Clustering:
  - **Market Segmentation:** Businesses use clustering to group customers based on purchasing patterns, demographics, and interests to tailor marketing strategies.
  - **Anomaly Detection:** Clustering can identify unusual data points that do not fit into any group. These anomalies can indicate fraudulent activity, mechanical faults, or errors in data collection.
  - **Image Segmentation:** In image processing, clustering is used to partition an image into segments that represent different objects or regions for further analysis.

AI VIET NAM
@aivietnam.edu.vn

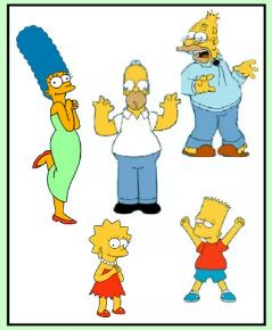
2– K-means clustering

What is K-Means Clustering?

What is a natural grouping among these objects?

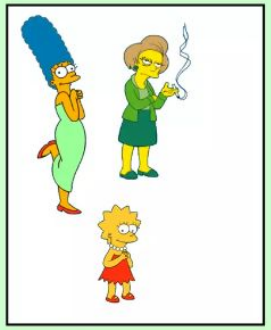


Clustering is subjective



Simpson's Family

School Employees



Females

Males

AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

What is K-Means Clustering?

Doc1 : Health , Medicine, Doctor
Doc 2 : Machine Learning, Computer
Doc 3 : Environment, Planet
Doc 4 : Pollution, Climate Crisis
Doc 5 : Covid, Health , Doctor

Doc1 : Health , Medicine, Doctor
Doc 2 : Machine Learning, Computer
Doc 3 : Environment, Planet
Doc 4 : Pollution, Climate Crisis
Doc 5 : Covid, Health , Doctor

Doc1 : Health , Medicine, Doctor
Doc 5 : Covid, Health , Doctor

Doc 2 : Machine Learning, Computer

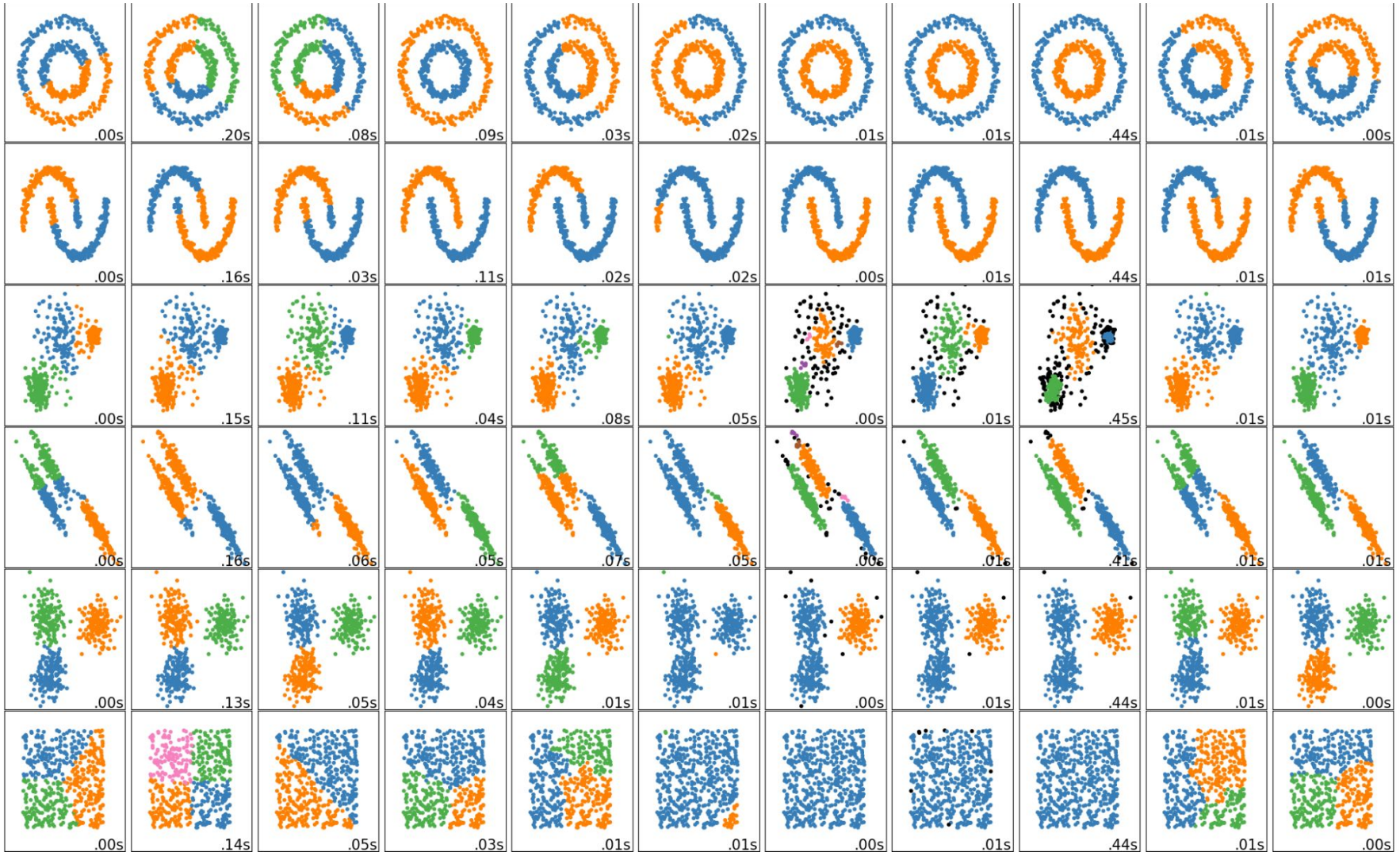
Doc 3 : Environment, Planet

Doc 4 : Pollution, Climate Crisis

AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

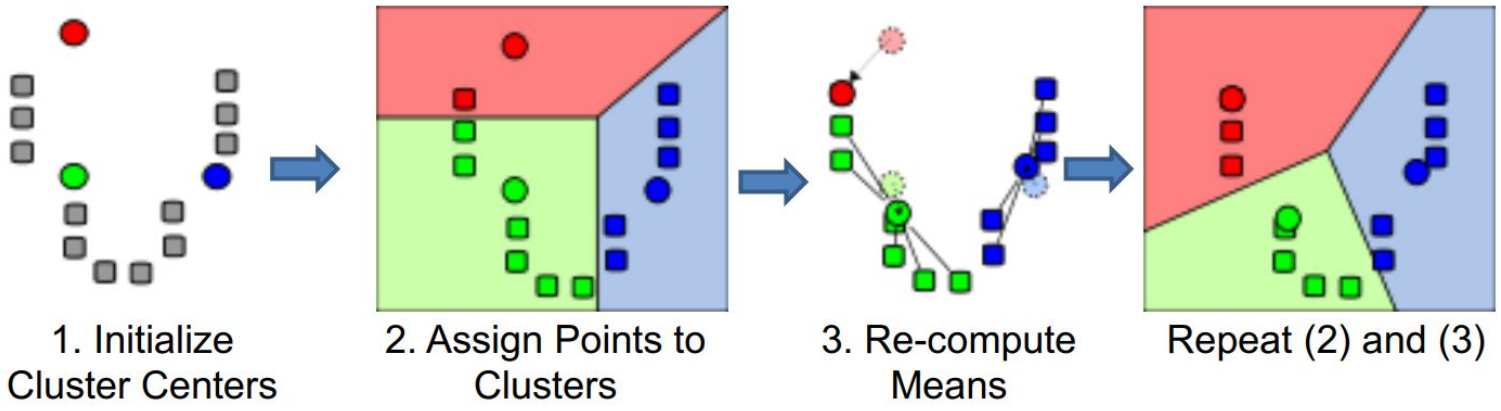
What is K-Means Clustering?



2– K-means clustering

How K-Means Works

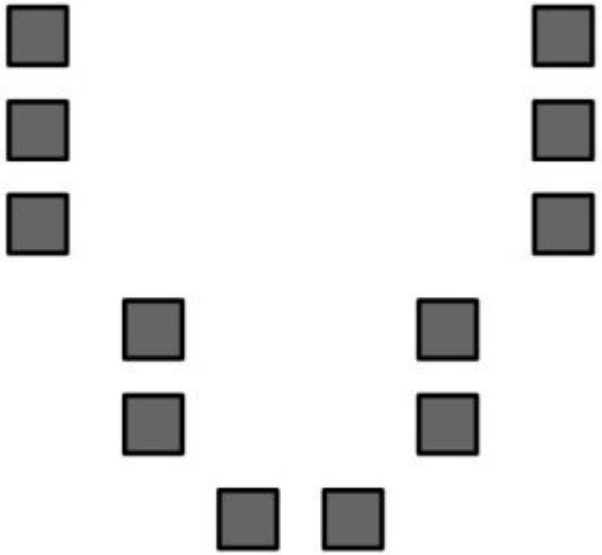
- **The k-means algorithm**
  - 1. randomly (or with another method) pick k cluster centers {c1,...,ck}
  - 2. for each j, set the cluster Xj to be the set of points in X that are the closest to center cj
  - 3. for each j let cj be the center of cluster Xj (mean of the vectors in Xj)
  - 4. repeat (go to step 2) until convergence

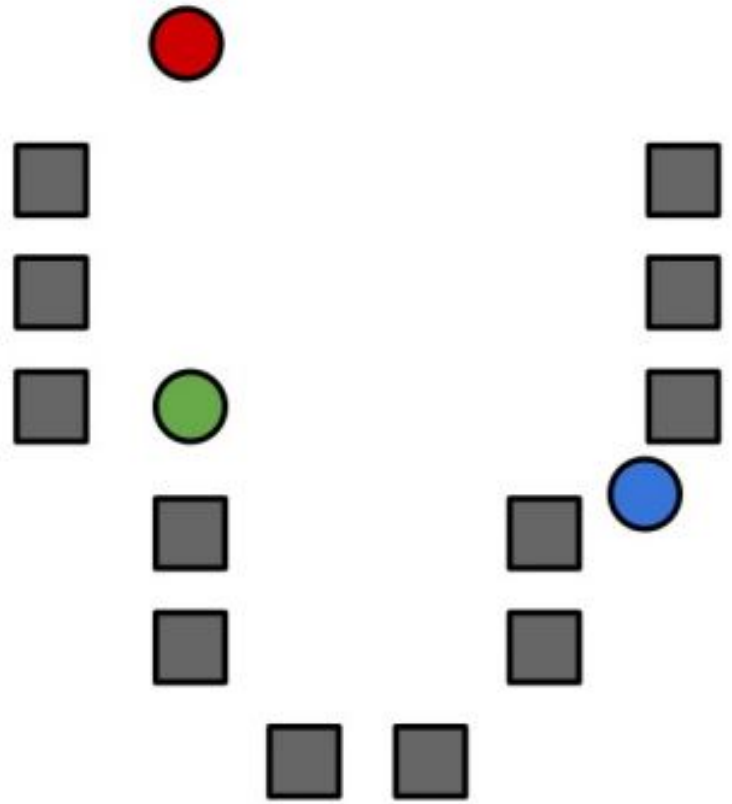


2– K-means clustering

How K-Means Works

- 1. randomly (or with another method) pick k cluster centers {c1,...,ck}

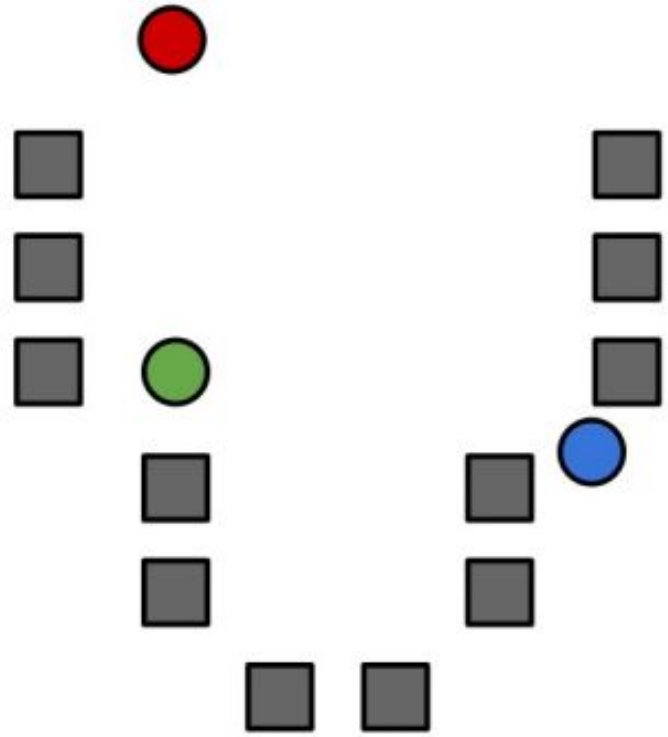




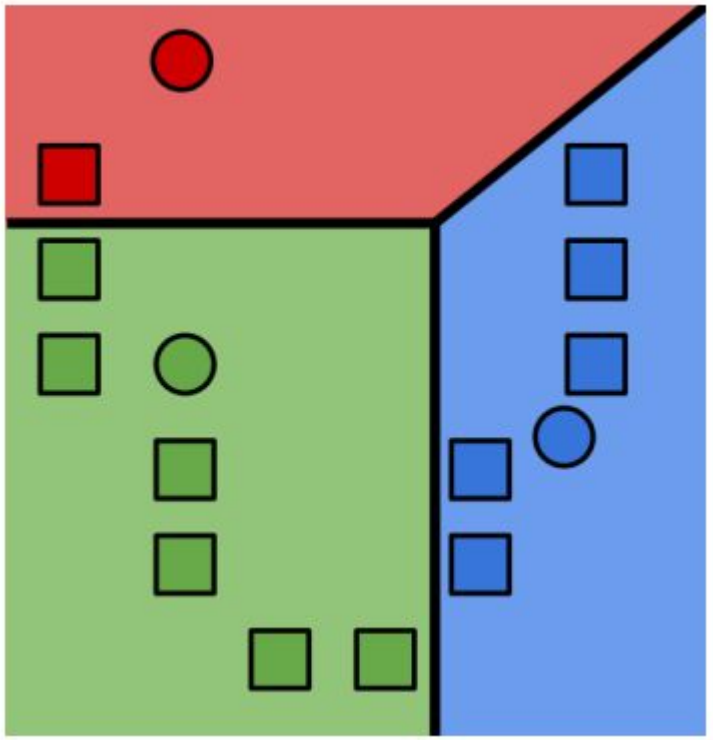
2– K-means clustering

How K-Means Works

- 2. for each j, set the cluster Xj to be the set of points in X that are the closest to center cj



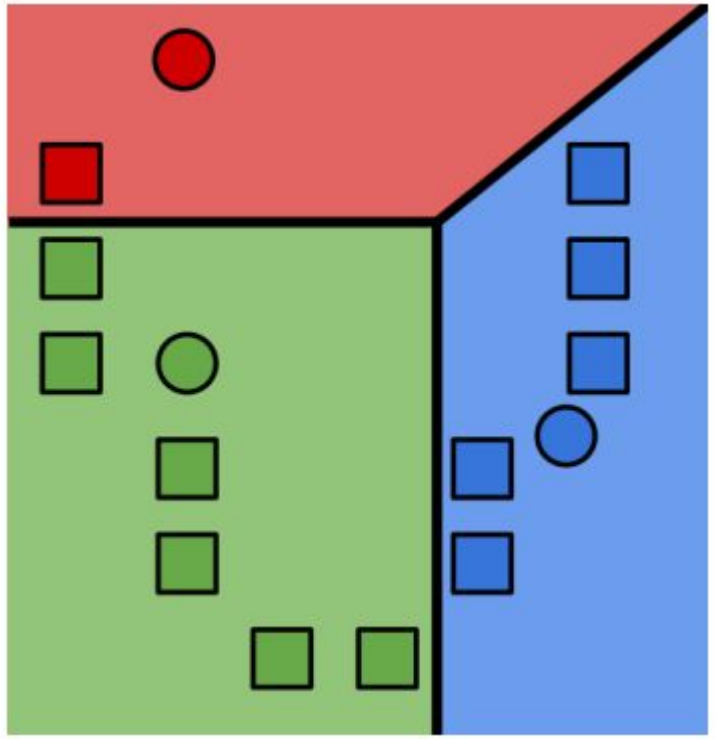
$$d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - c_{jl})^2}$$



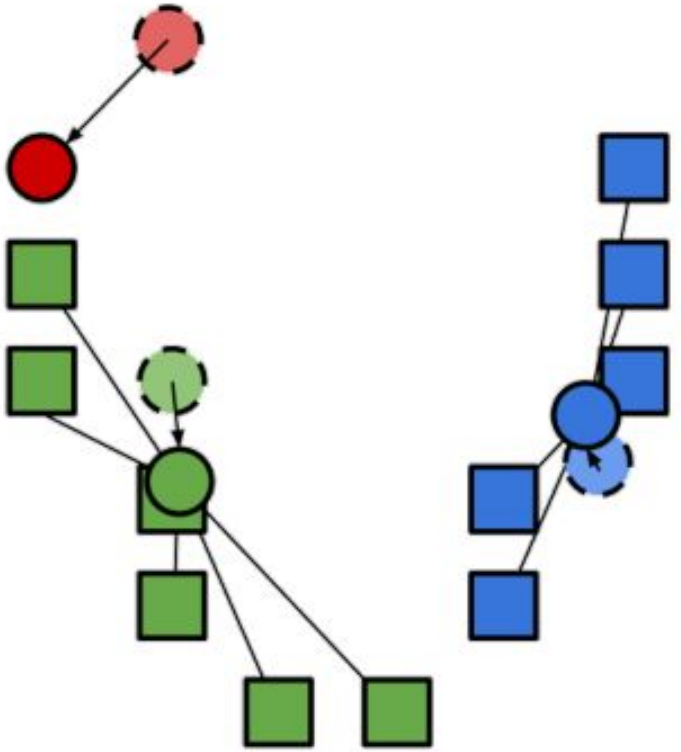
2– K-means clustering

How K-Means Works

- for each j let cj be the center of cluster Xj (mean of the vectors in Xj)



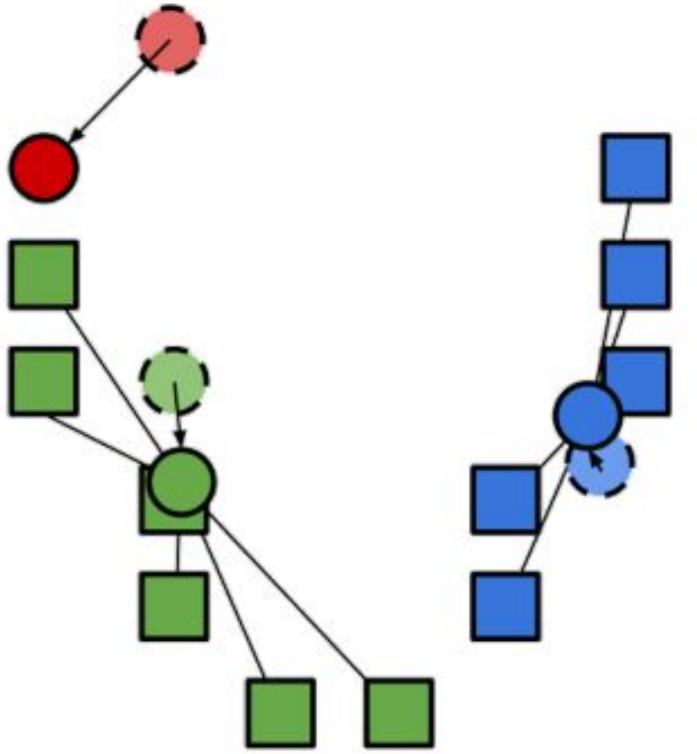
$$\mathbf{c}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i$$

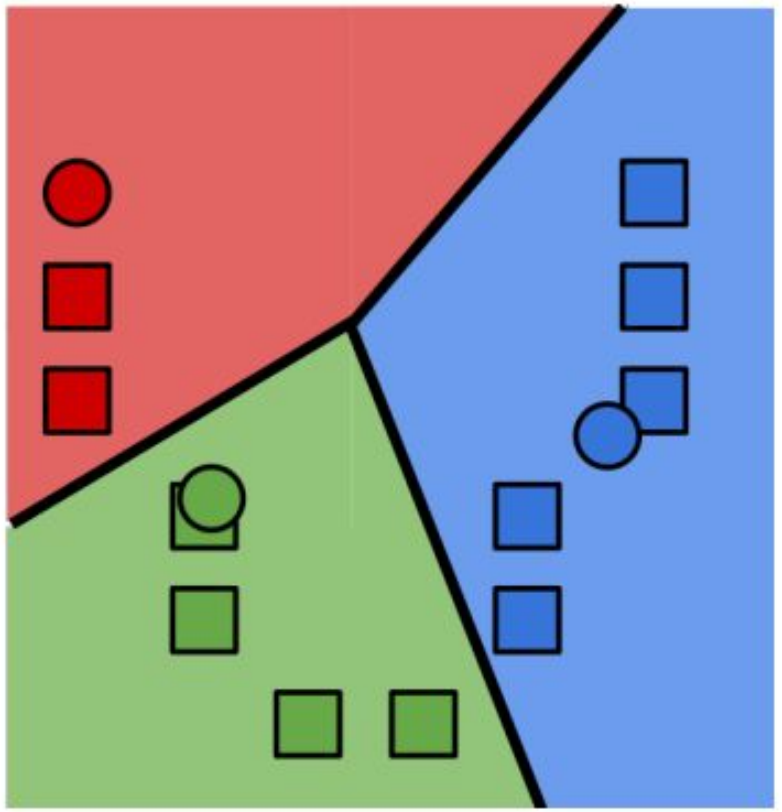


2– K-means clustering

How K-Means Works

- 4. repeat (go to step 2) until convergence





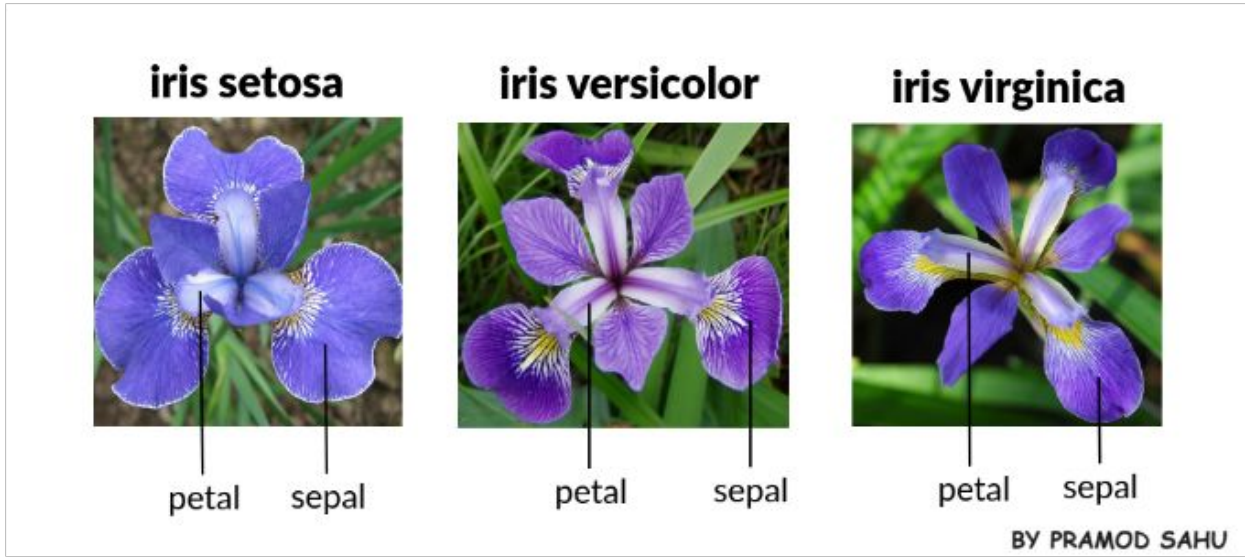
AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

Example - Iris dataset case study

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

150 rows × 4 columns



Iris Dataset 150 samples, 4 features

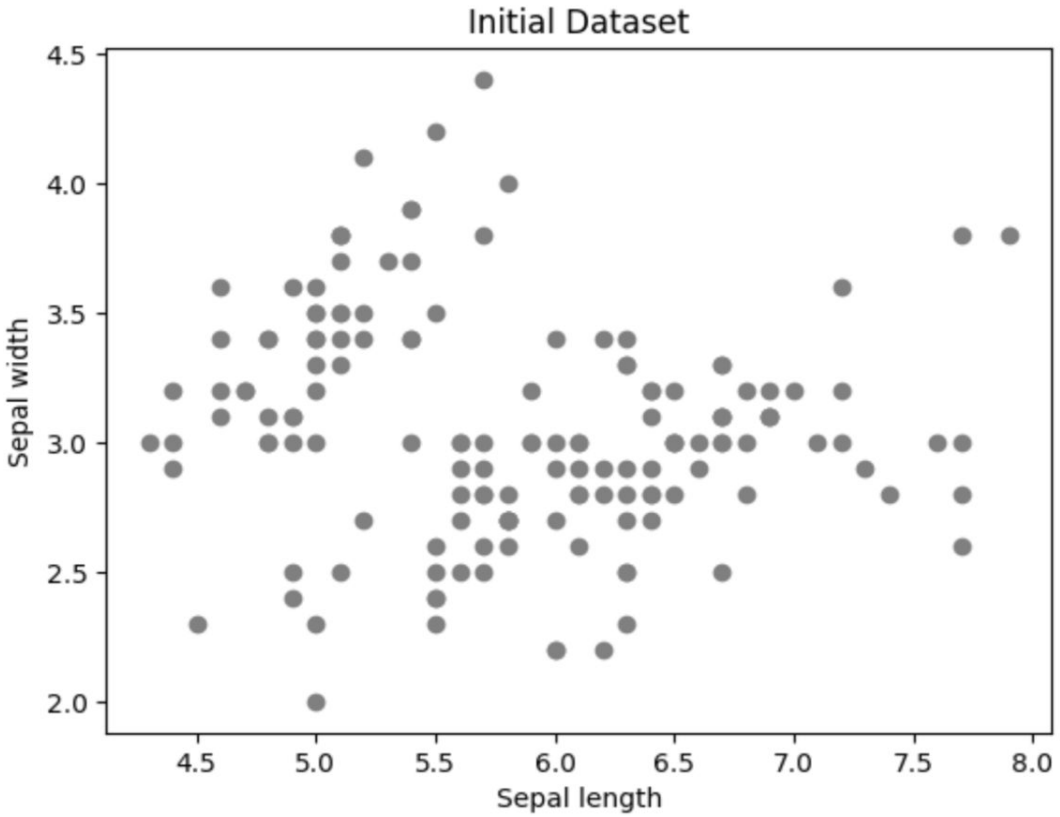
Iris classification

2– K-means clustering

Example - Iris dataset case study

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

150 rows × 4 columns

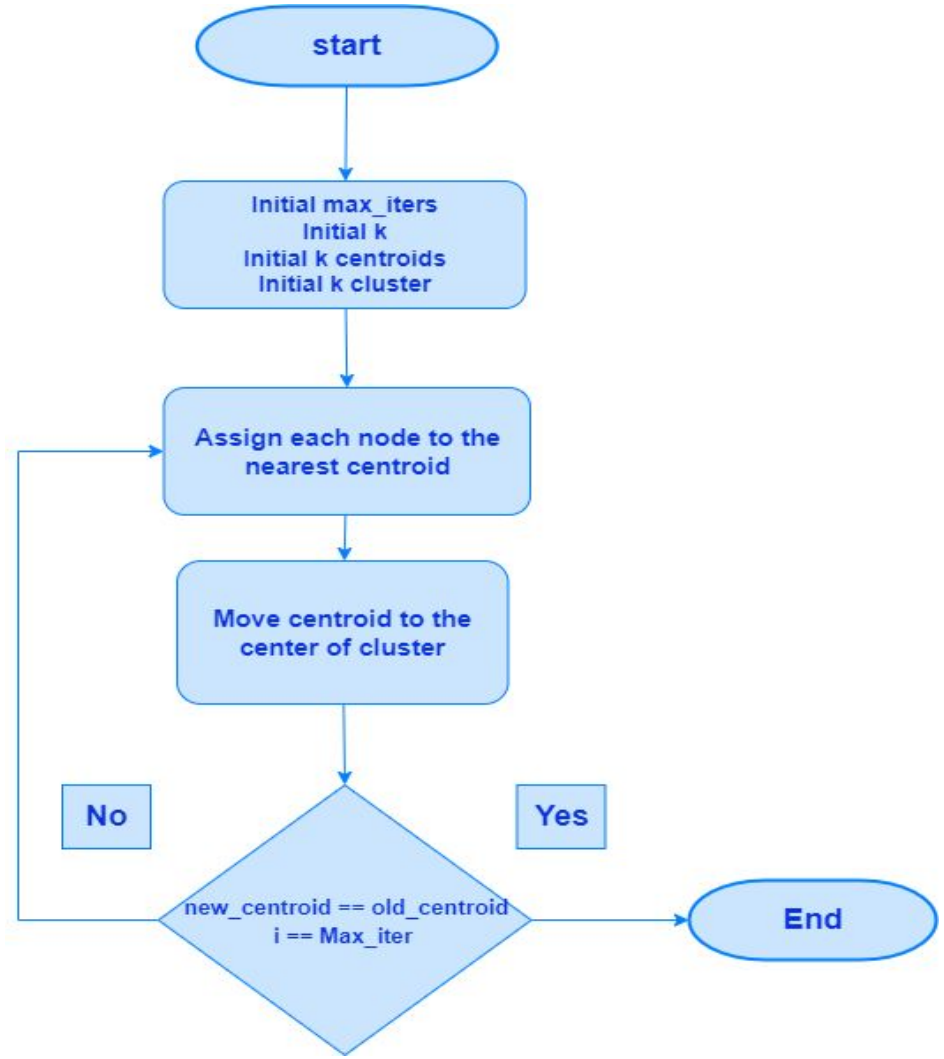


Iris Dataset with 4 features

Plotting Iris Dataset with the first two features

2– K-means clustering

Example - Iris dataset case study



Euclid Distance

We will allocate the data point to the centroid by calculating the distance with Euclid function

$$d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - c_{jl})^2}$$

Calculate mean of all data point in one cluster and move centroid to this position

$$\mathbf{c}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i$$

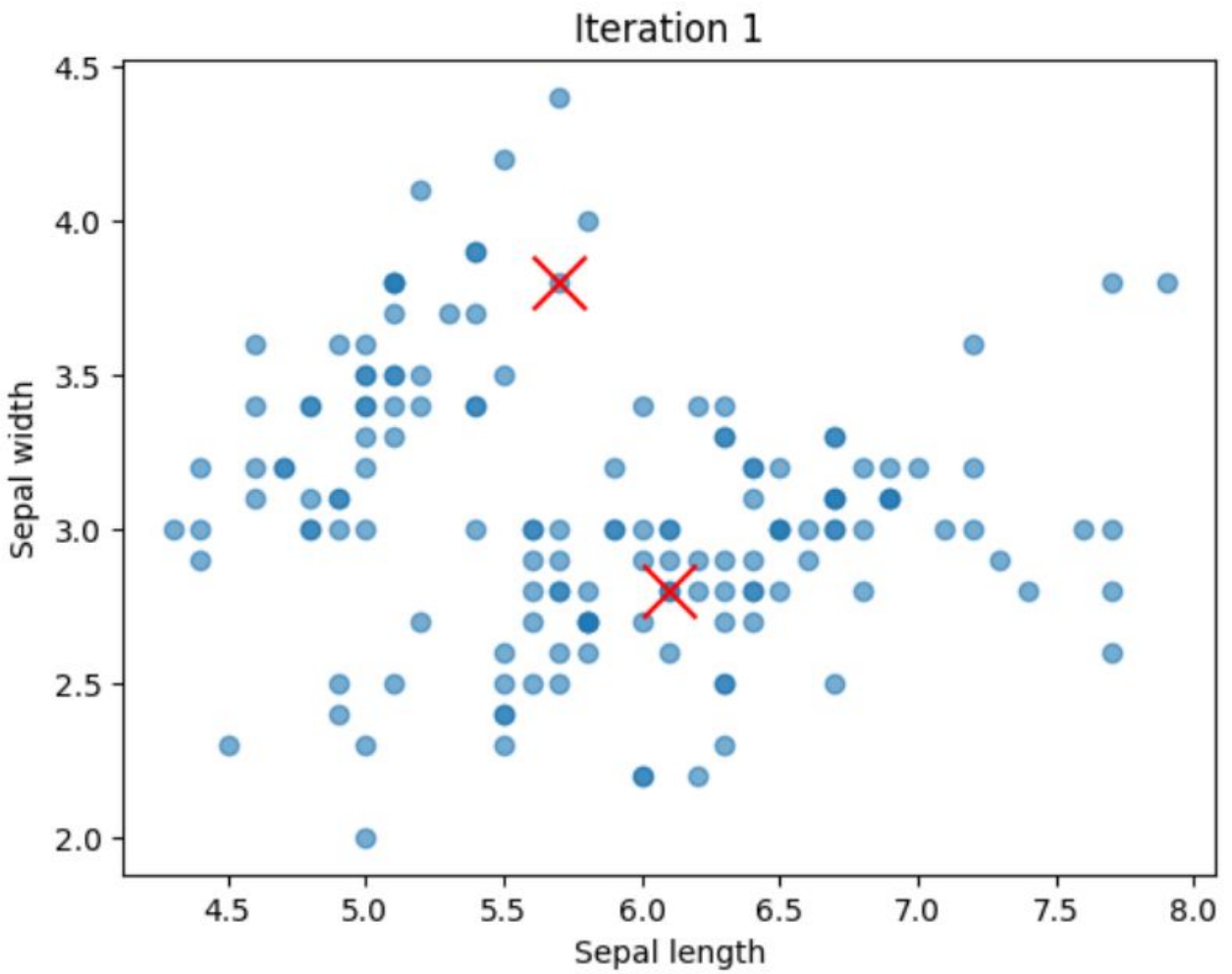
AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

Example - Iris dataset case study

Step 1: Initialization

- K = 2 (2 cluster)
- Max-iter: 100
- 2 Centroid: Choose randomly two data point in dataset: e.g (5.7, 3.8), (6.1, 2.8)



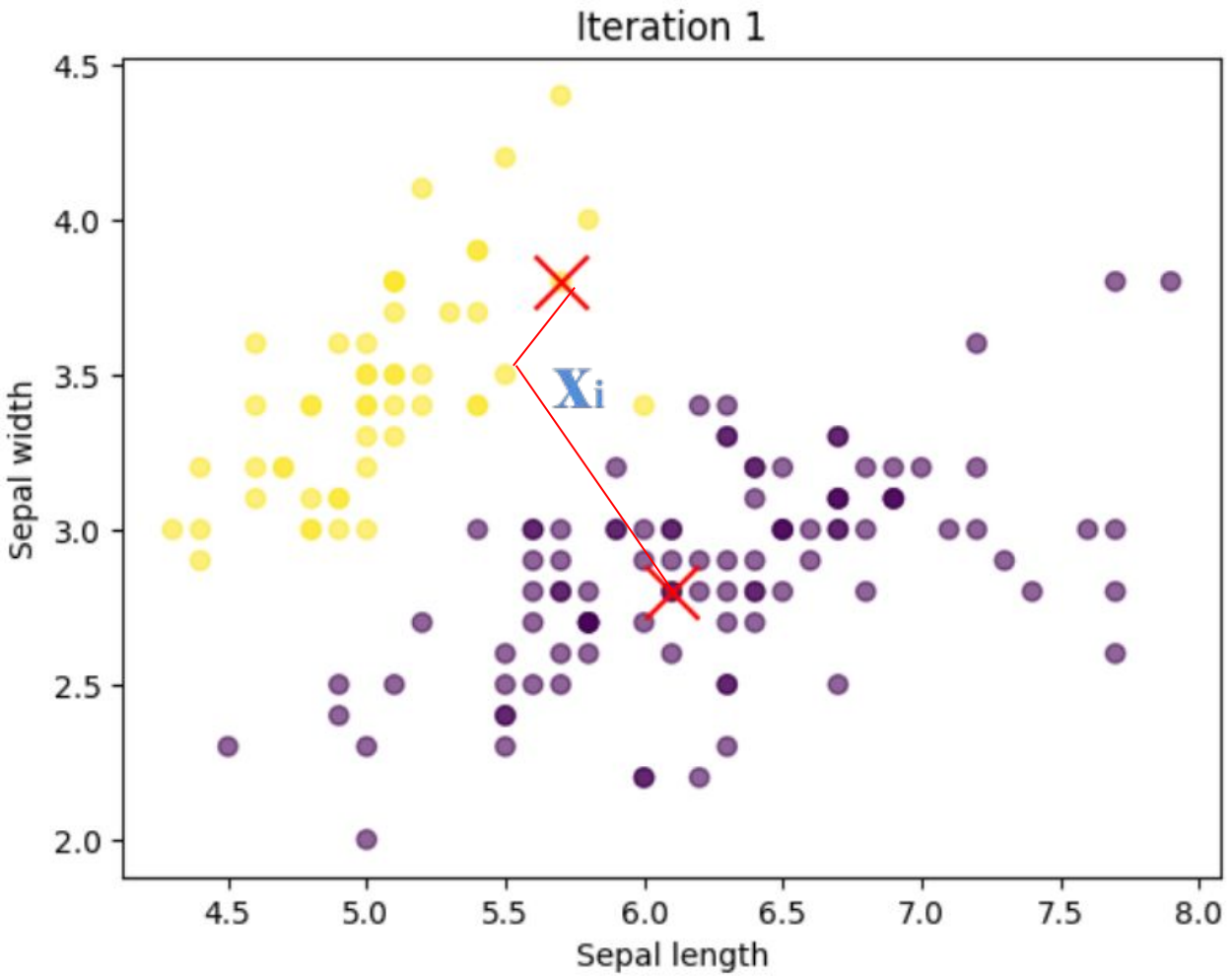
2– K-means clustering

Example - Iris dataset case study

Step 2: Allocation

- With each data point $\mathbf{x}_i$, calculate the Euclidean Distance of it with 2 centroids. If $\mathbf{x}_i$ near centroid 1 than centroid 2, assign $\mathbf{x}_i$ to the controid 1. And vice versa.

$$d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - c_{jl})^2}$$



2– K-means clustering

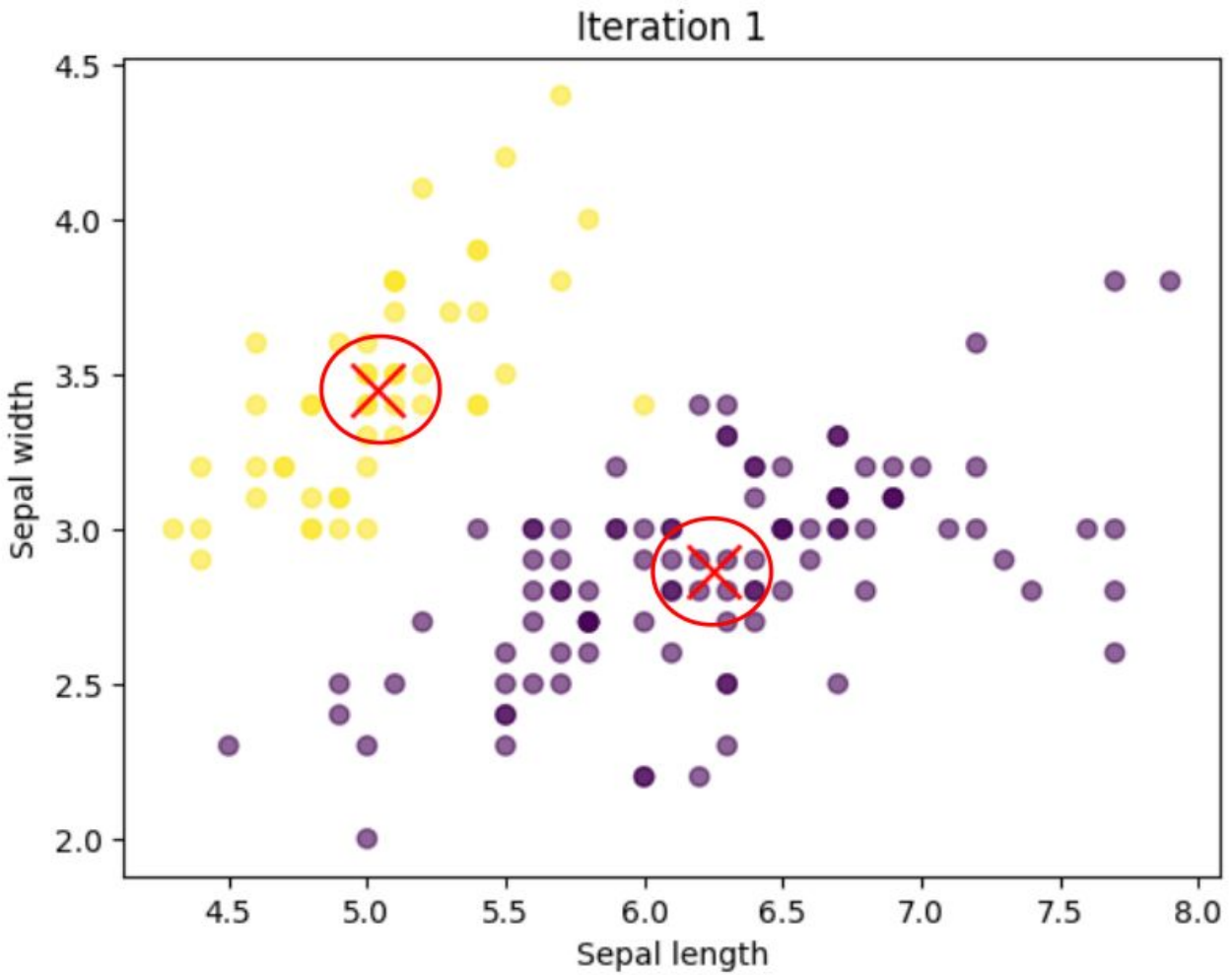
Example - Iris dataset case study

Step 3: Move centroid

- With each cluster, we calculate the mean position of all data points.

$$\mathbf{c}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i$$

- Then this position will become the new centroid of this cluster.



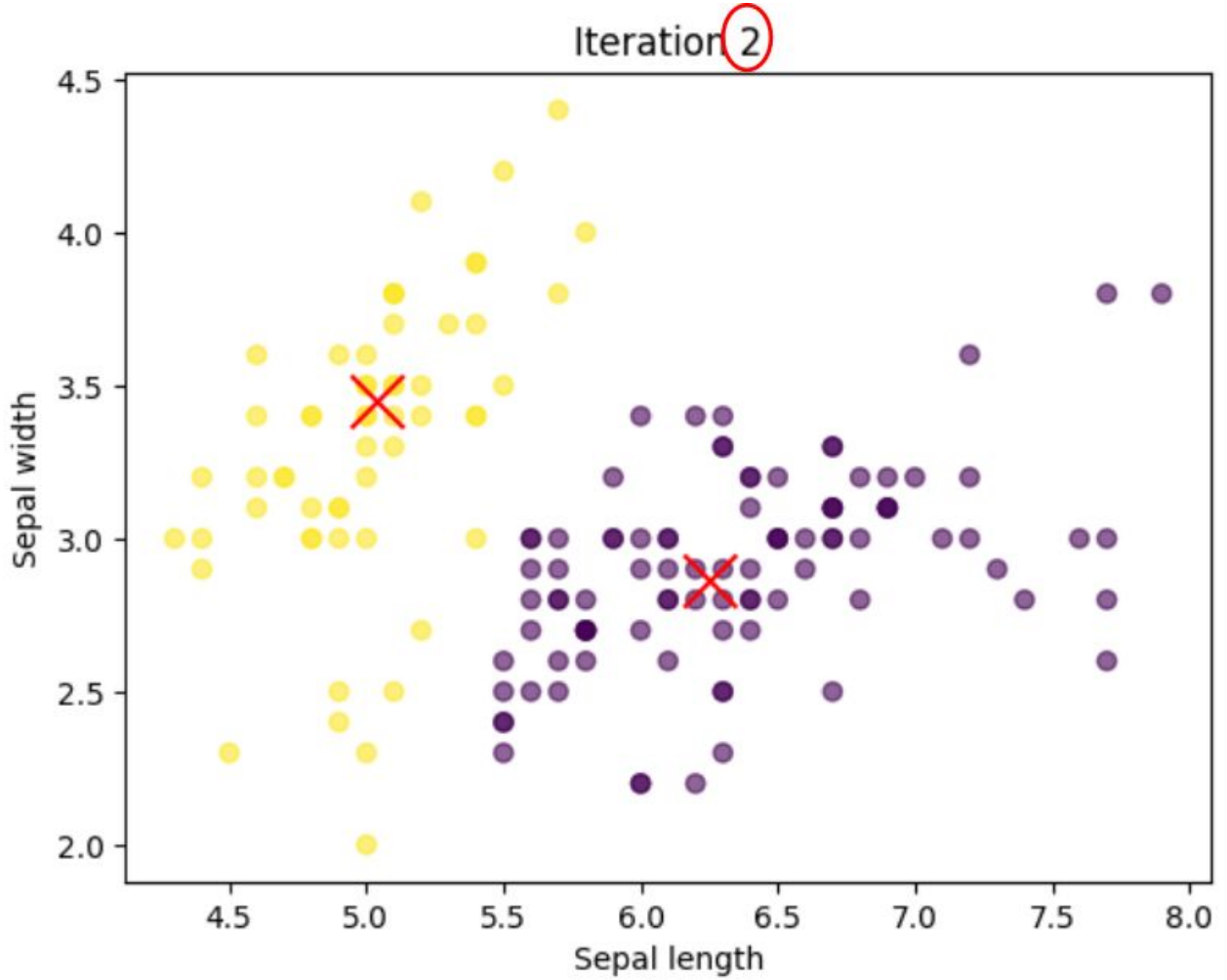
2– K-means clustering

Example - Iris dataset case study

Repeat Step 2: Allocation

- With each data point xi, calculate the Euclidean Distance of it with 2 centroids. If xi near centroid 1 than centroid 2, assign xi to the centroid 1. And vice versa.

$$d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - c_{jl})^2}$$



2– K-means clustering

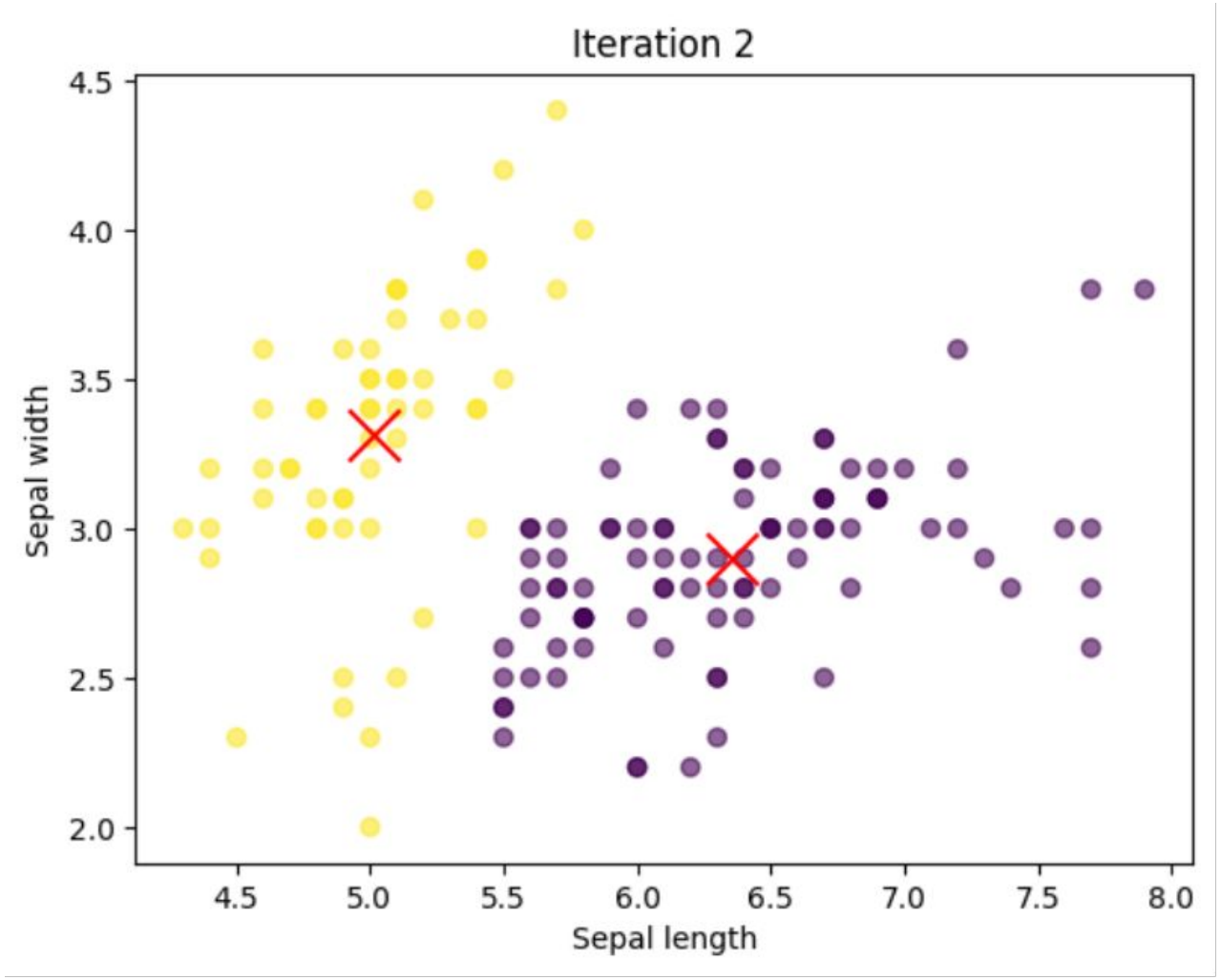
Example - Iris dataset case study

Repeat Step 3: Move centroid

- With each cluster, we calculate the mean position of all data points.

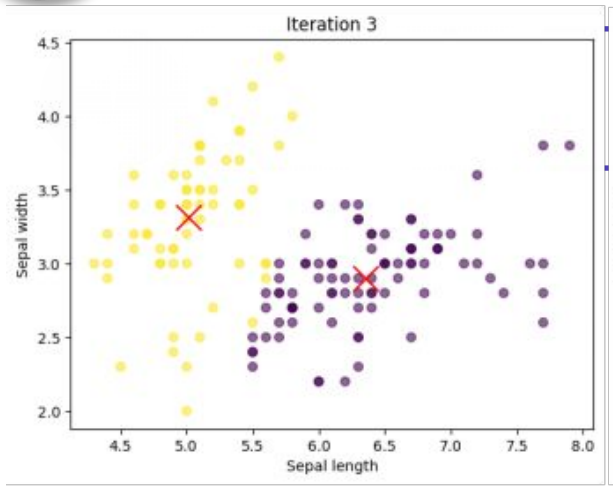
$$\mathbf{c}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i$$

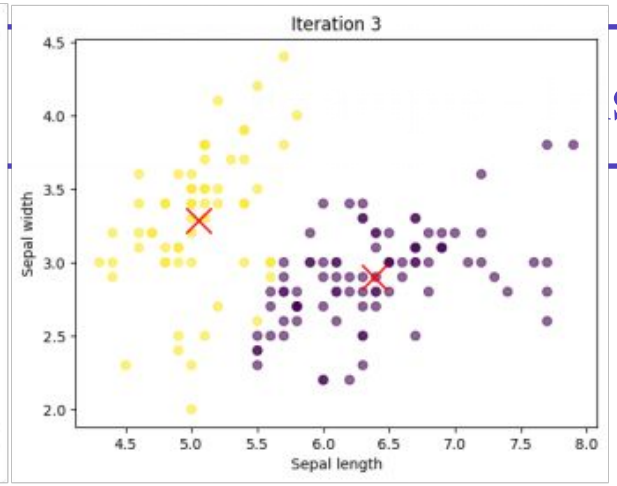
- Then this position will become the new centroid of this cluster.

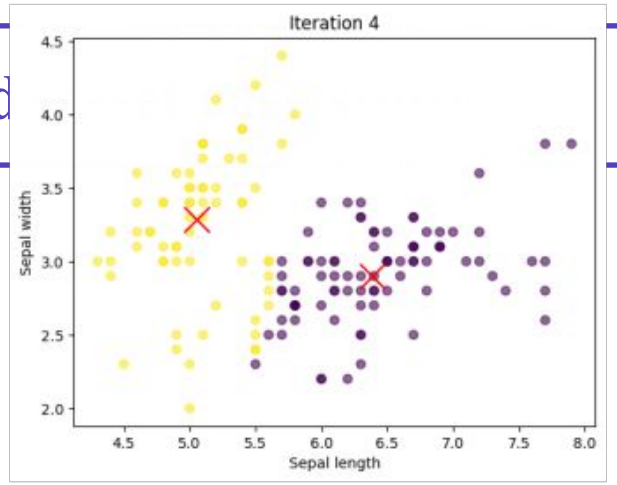


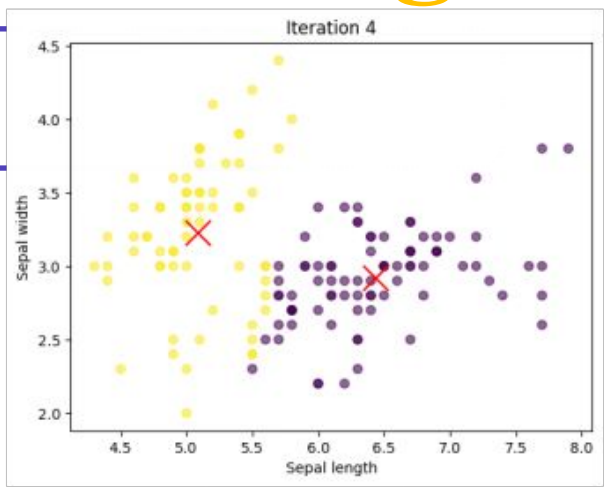
AI VIET NAM
@aivietnam.edu.vn

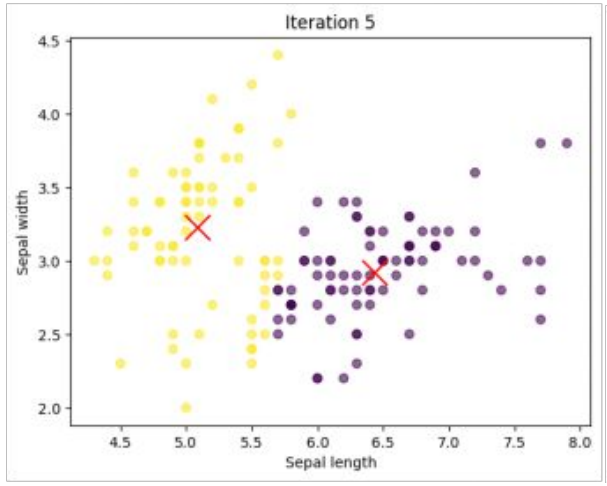
2– K-means clustering

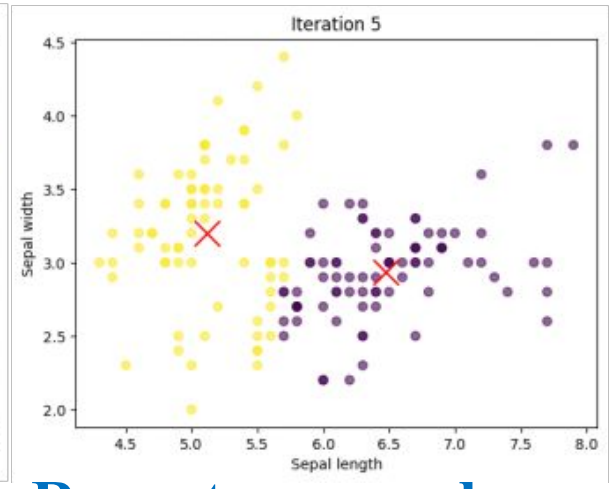


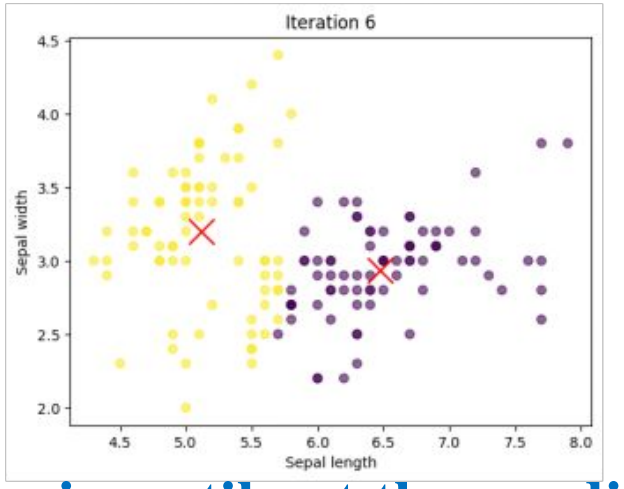


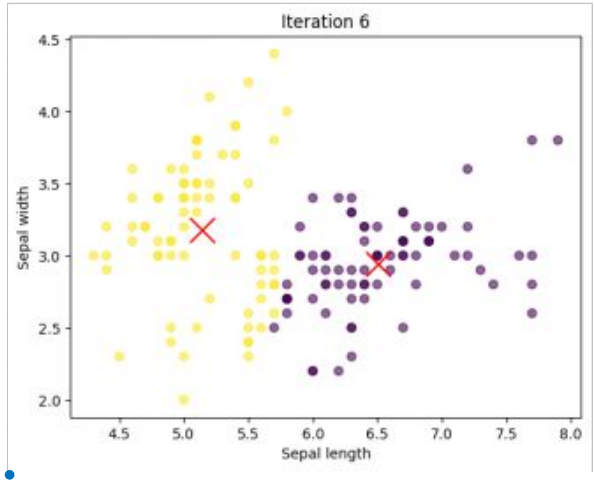












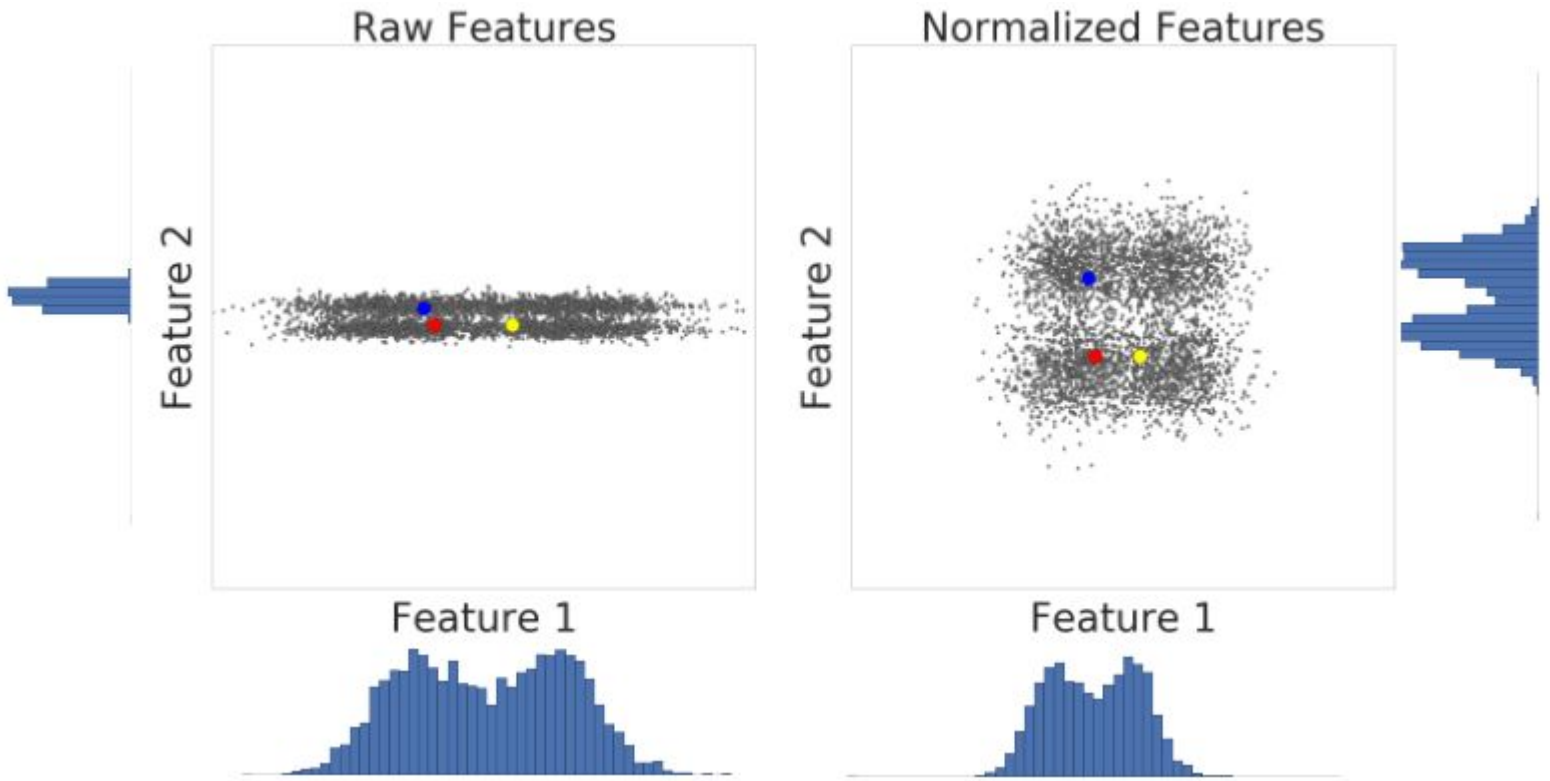
Repeat over and over again until got the condition

- The iteration is over max_iter.
- The centroid not change.

2– K-means clustering

Normalizing data

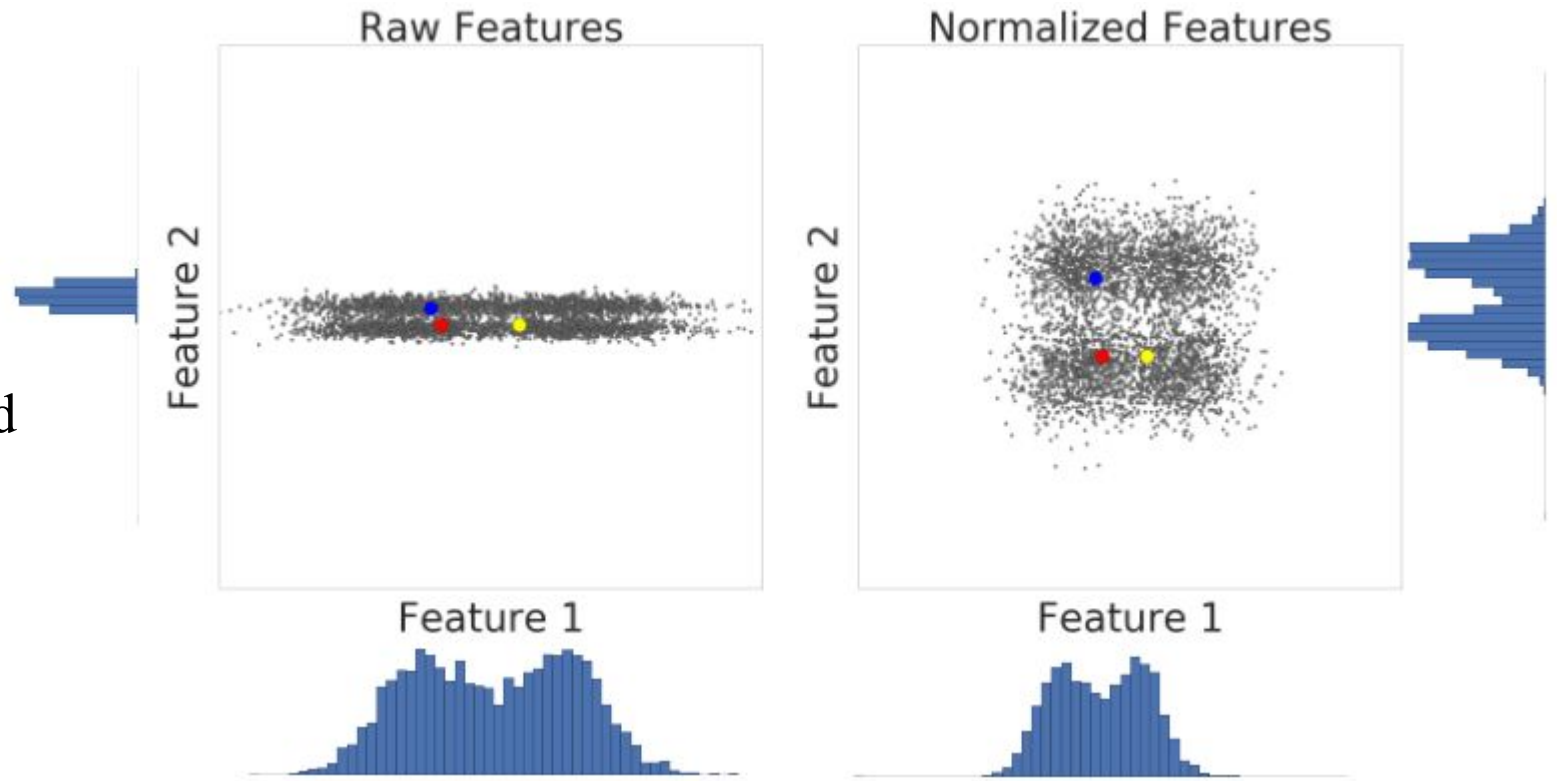
- Transform data for multiple features to the same scale by normalizing the data.
  - **Z-scores:** Whenever you see a dataset roughly shaped like a **Gaussian distribution**, you should calculate z-scores for the data. Z-scores are the number of standard deviations a value is from the mean.



2– K-means clustering

Normalizing data

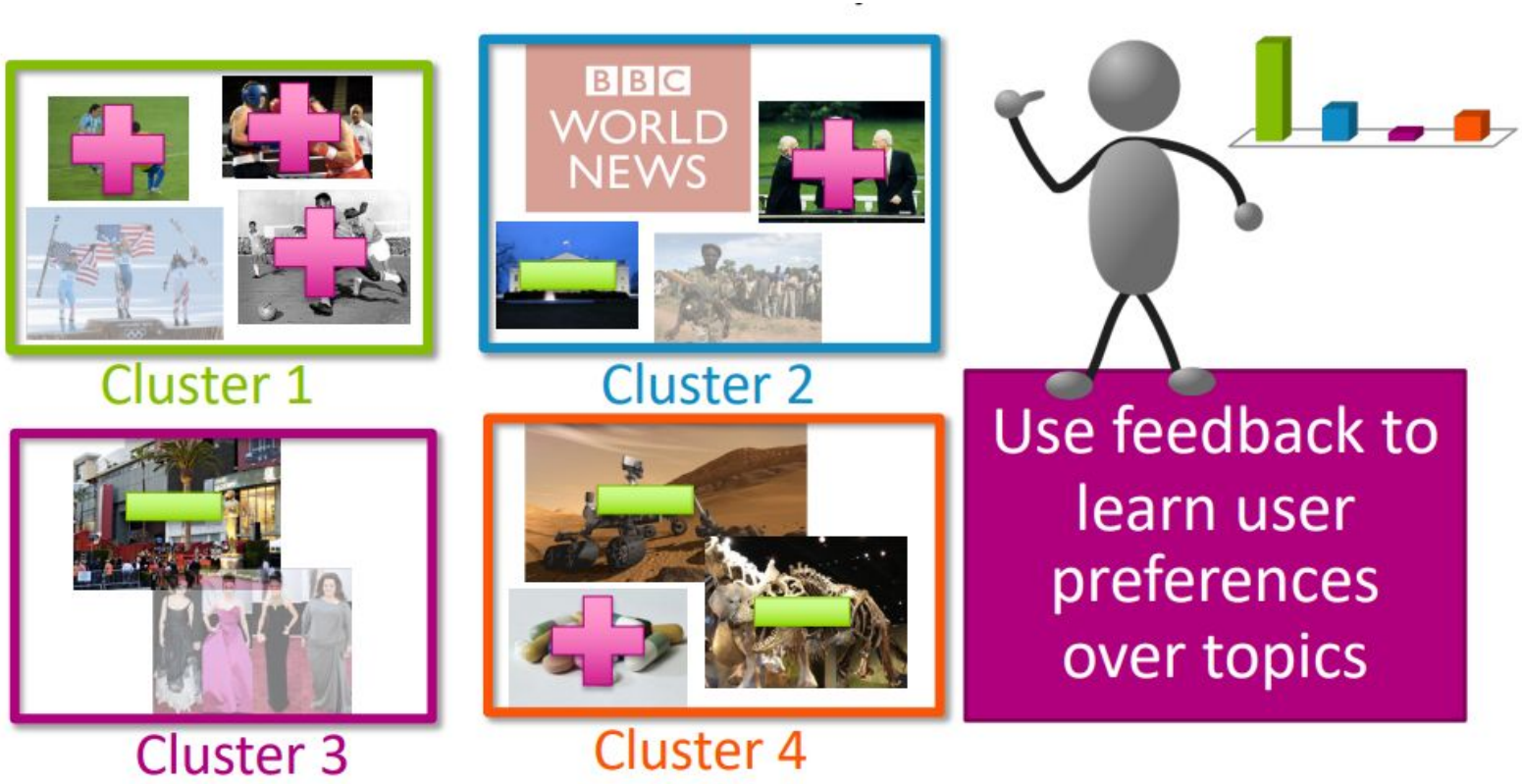
In the **unnormalized dataset** on the left, Feature 1 and Feature 2, respectively graphed on the x and y axes, **don't have the same scale**. On the left, the **red** example **appears closer**, **or more similar**, **to blue than to yellow**. On the right, **after z-score scaling,** Feature 1 and Feature 2 have the **same scale**, and the **red example appears closer to the yellow example**. The normalized dataset gives a more accurate measure of similarity between points.



2– K-means clustering

Applications

K-means clustering can be applied to user feedback data to **discern patterns in user preferences across different topics**. By grouping similar feedback responses, K-means helps in identifying common interests or disinterests among users, enabling personalized content delivery or recommendations tailored to user clusters. This approach enhances user engagement by aligning content more closely with each user's specific interests.



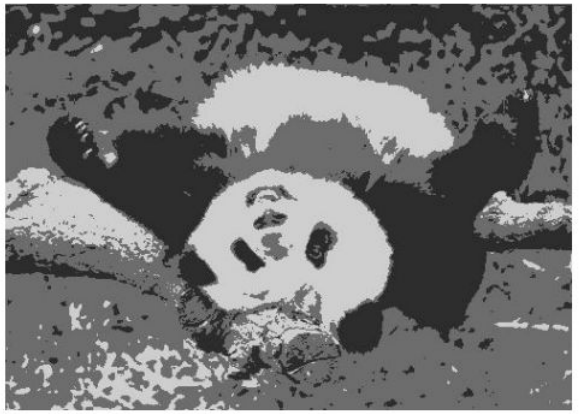
2– K-means clustering

Applications

K-means clustering can also be applied to **image compression by reducing the number of colors in an image**. This is done by **treating each color as a point in a three-dimensional space** (representing the RGB values) and **clustering these points into K groups**. Each cluster centroid becomes the **new color for all points** (pixels) in that cluster, significantly **reducing the image's color palette**, which leads to a reduction in image size without severely impacting visual quality.

Original image

2 clusters



3 clusters

2– K-means clustering

Applications

K-means clustering can also be applied to **image compression by reducing the number of colors in an image**. This is done by **treating each color as a point in a three-dimensional space** (representing the RGB values) and **clustering these points into K groups**. Each cluster centroid becomes the **new color for all points** (pixels) in that cluster, significantly **reducing the image's color palette**, which leads to a reduction in image size without severely impacting visual quality.

2– K-means clustering

Applications

Example: K-Means for Segmentation

K=2

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original

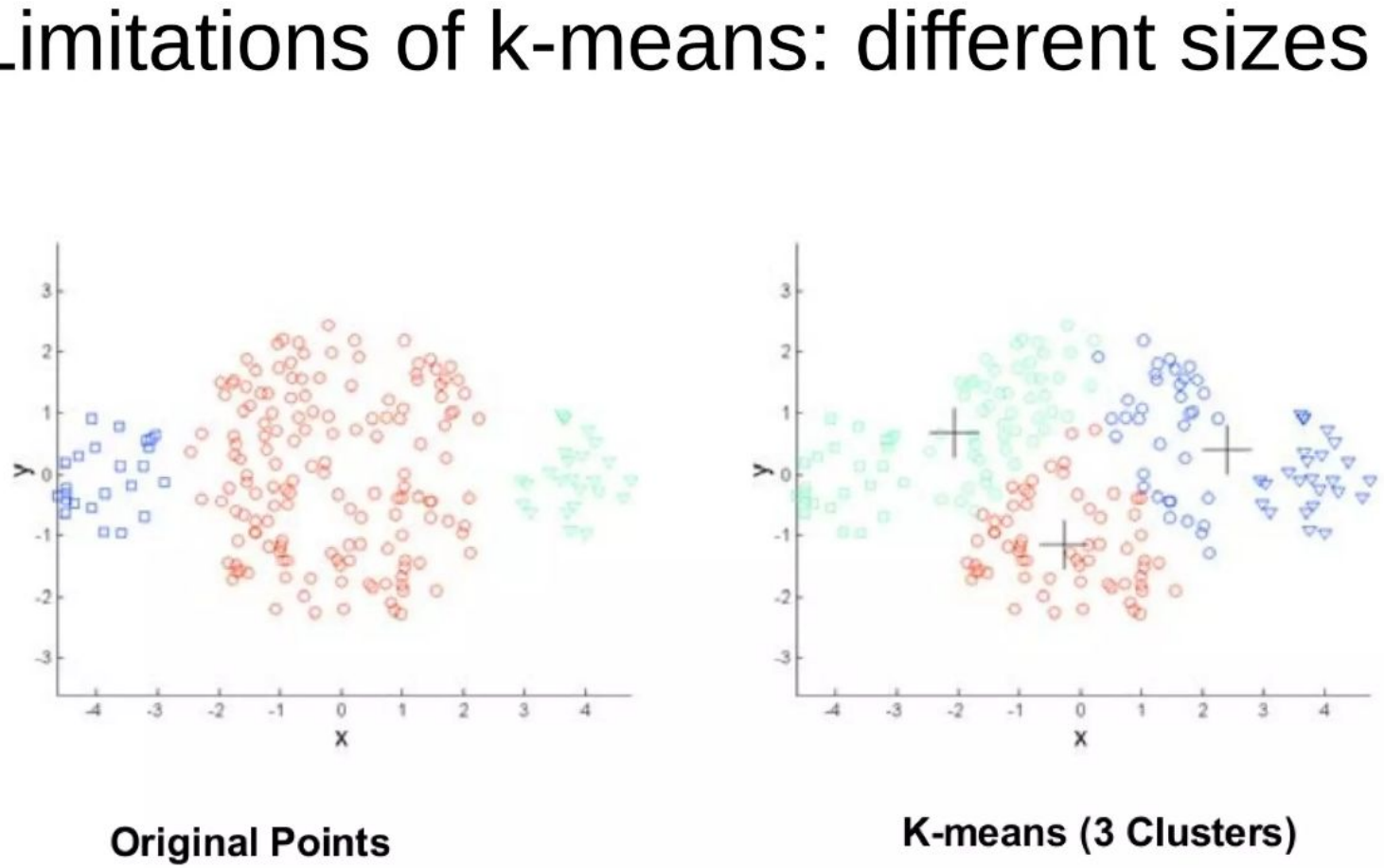
2– K-means clustering

Applications

2– K-means clustering

Advantages and Disadvantages

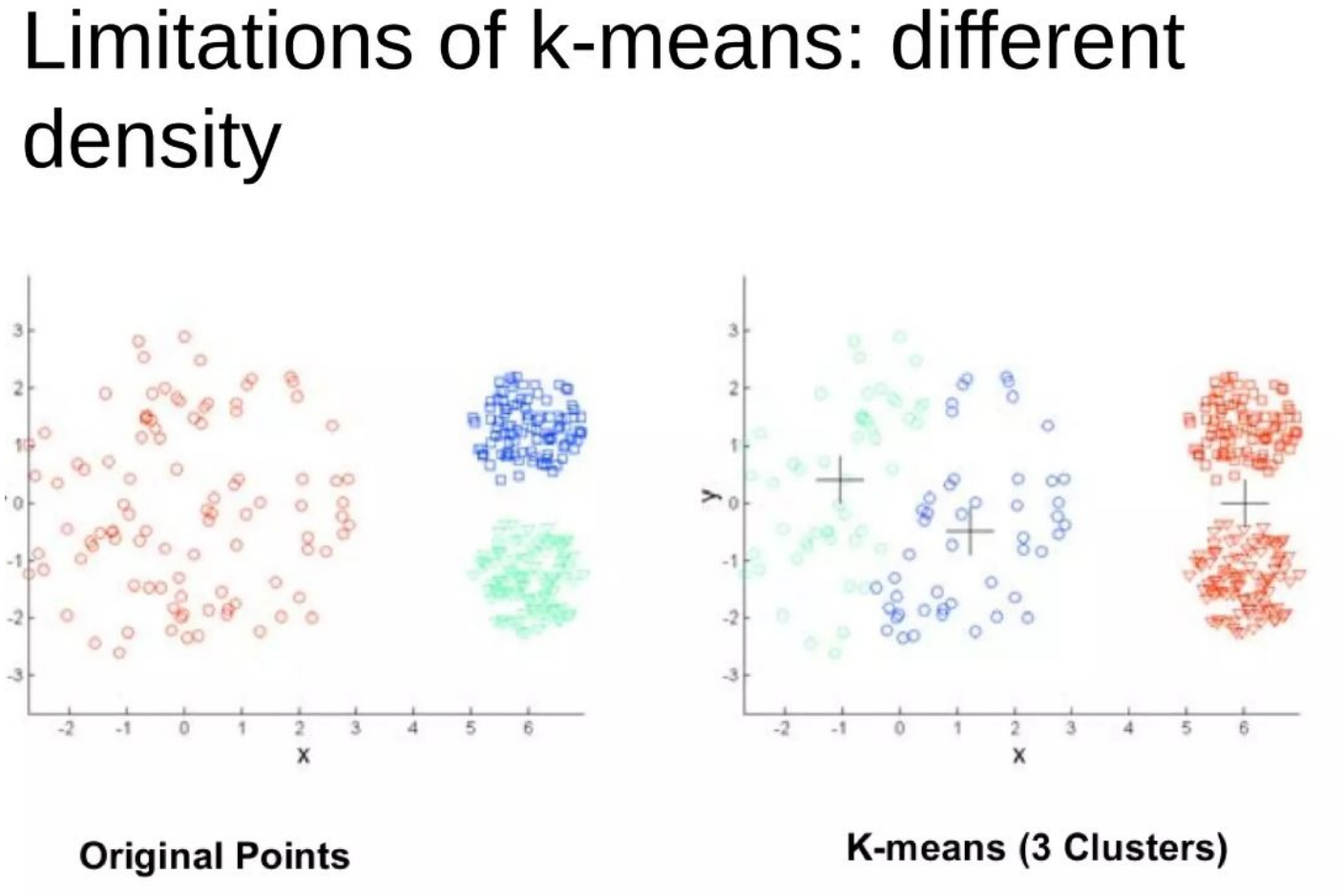
One limitation of K-means clustering is its **difficulty in handling clusters of varying sizes and densities**. Since K-means uses the mean of cluster members to define the centroid, it inherently **assumes that clusters are spherical and similar in size**. This can lead to **suboptimal clustering** when the actual data clusters vary significantly in size or density, as the algorithm tends to favor creating clusters of roughly equal spatial extent.



2– K-means clustering

Advantages and Disadvantages

A limitation of K-means clustering is its **ineffectiveness with clusters of varying densities**. The algorithm calculates centroids based on the average of points within a cluster, leading to an assumption of **uniform density across clusters**. This assumption can cause K-means to **perform poorly when dealing with data where clusters have different densities**, often **misclassifying dense regions as separate clusters and sparse regions as a single cluster.**

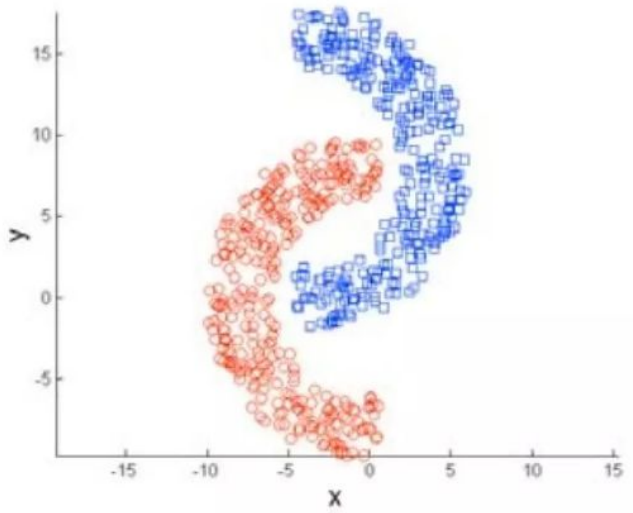


2– K-means clustering

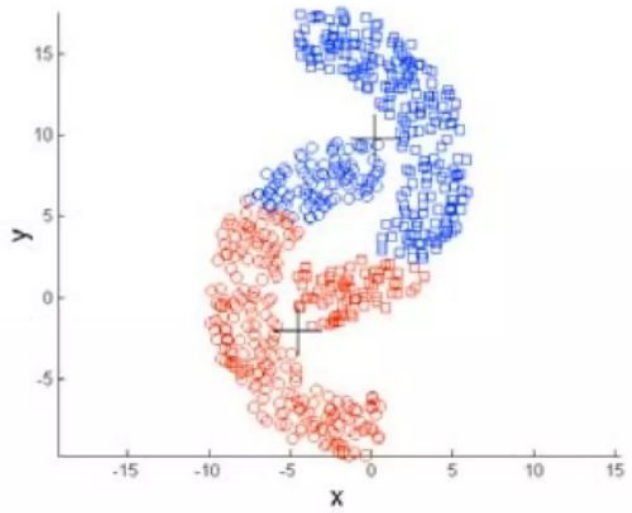
Advantages and Disadvantages

K-means clustering also struggles with **non-spherical cluster shapes** due to its reliance on **Euclidean distance to assign points to the nearest centroid**. This metric assumes clusters are spherical; thus, K-means can **inaccurately partition data that naturally forms elongated or irregular shapes**, often splitting these into multiple clusters or merging them incorrectly with other groups. This limitation significantly affects the versatility of K-means in practical applications where data may not conform to spherical distributions.

Limitations of k-means: non-spherical shapes



Original Points



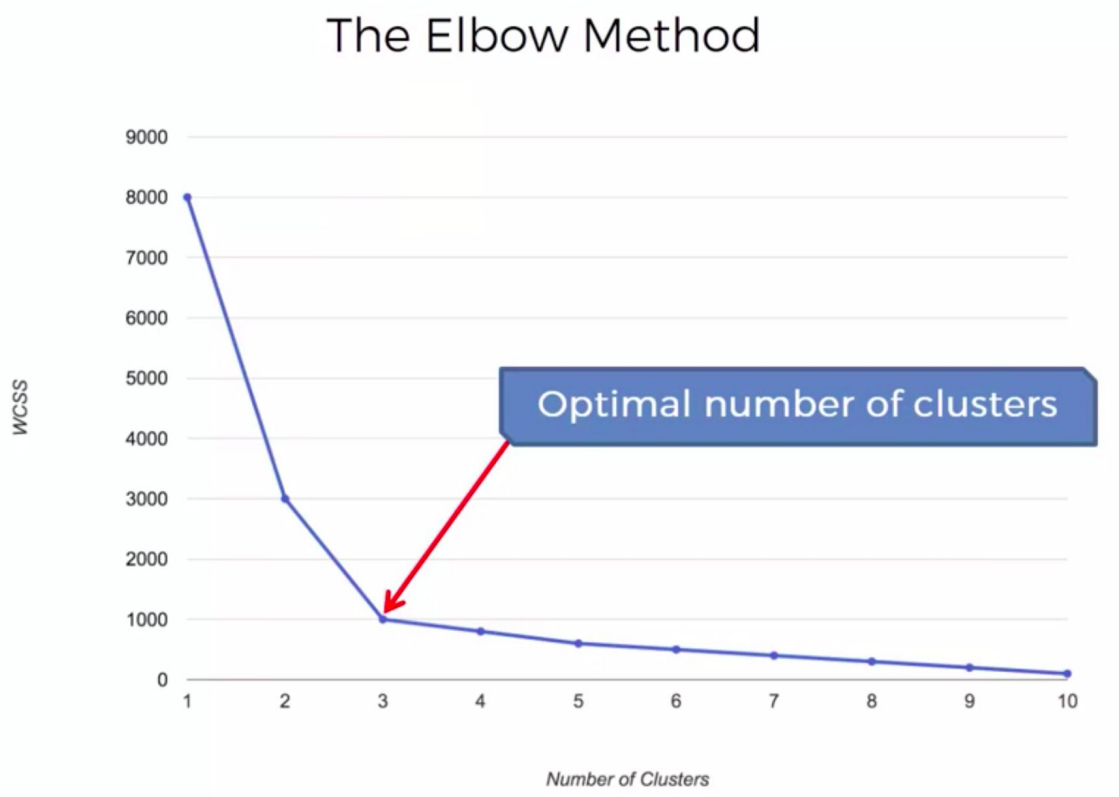
K-means (2 Clusters)

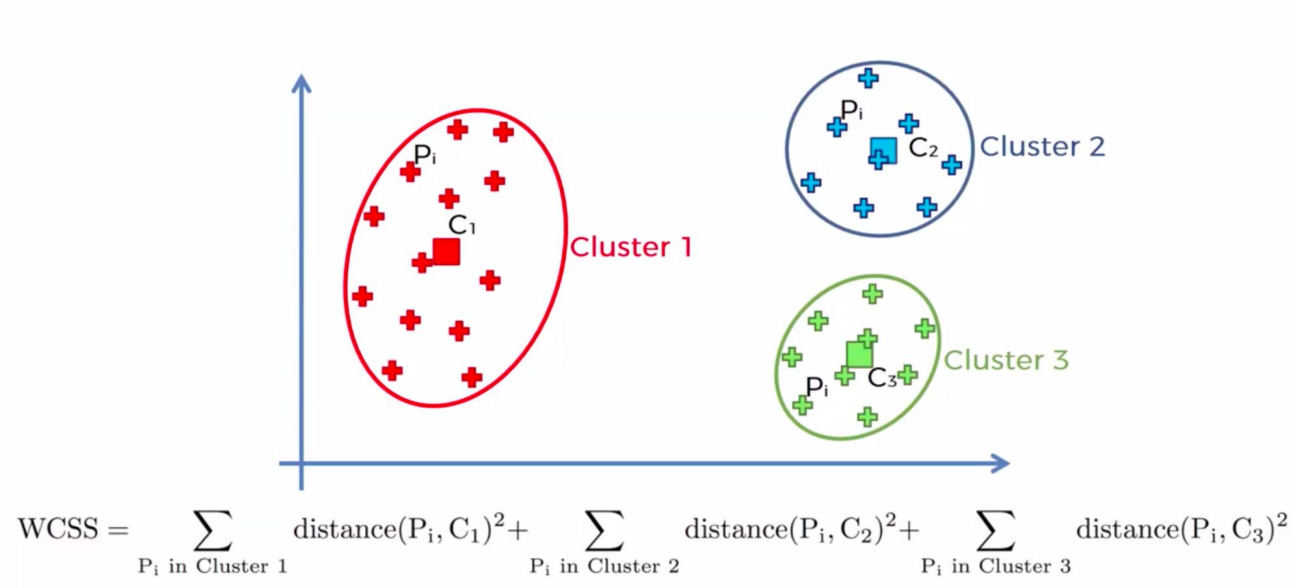

QUIZ
QUIZ
QUIZ

AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

How to Determine the Optimal K for K-Means?





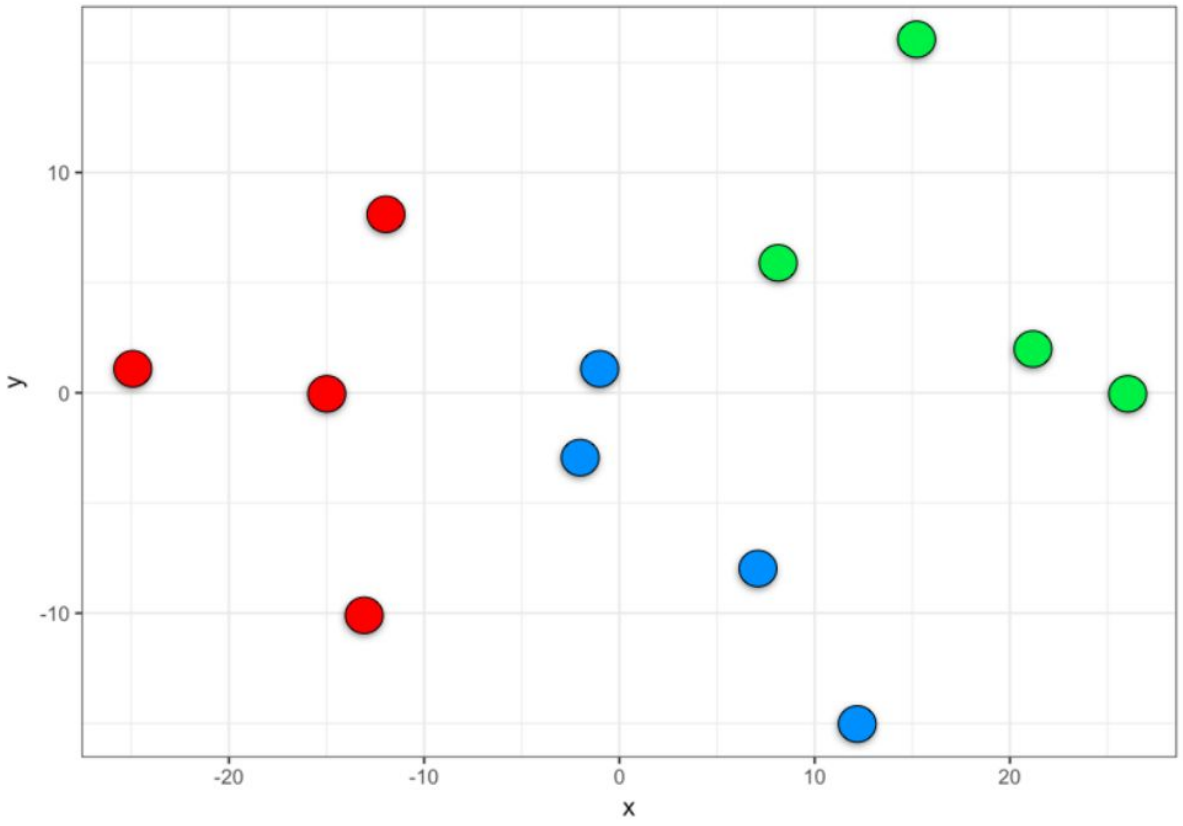
https://github.com/serodriguez68/machine-learning-cheatsheet/blob/master/cheatsheets/Part%204%20-%20Clustering/k_means_clustering.md

AI VIET NAM
@aivietnam.edu.vn

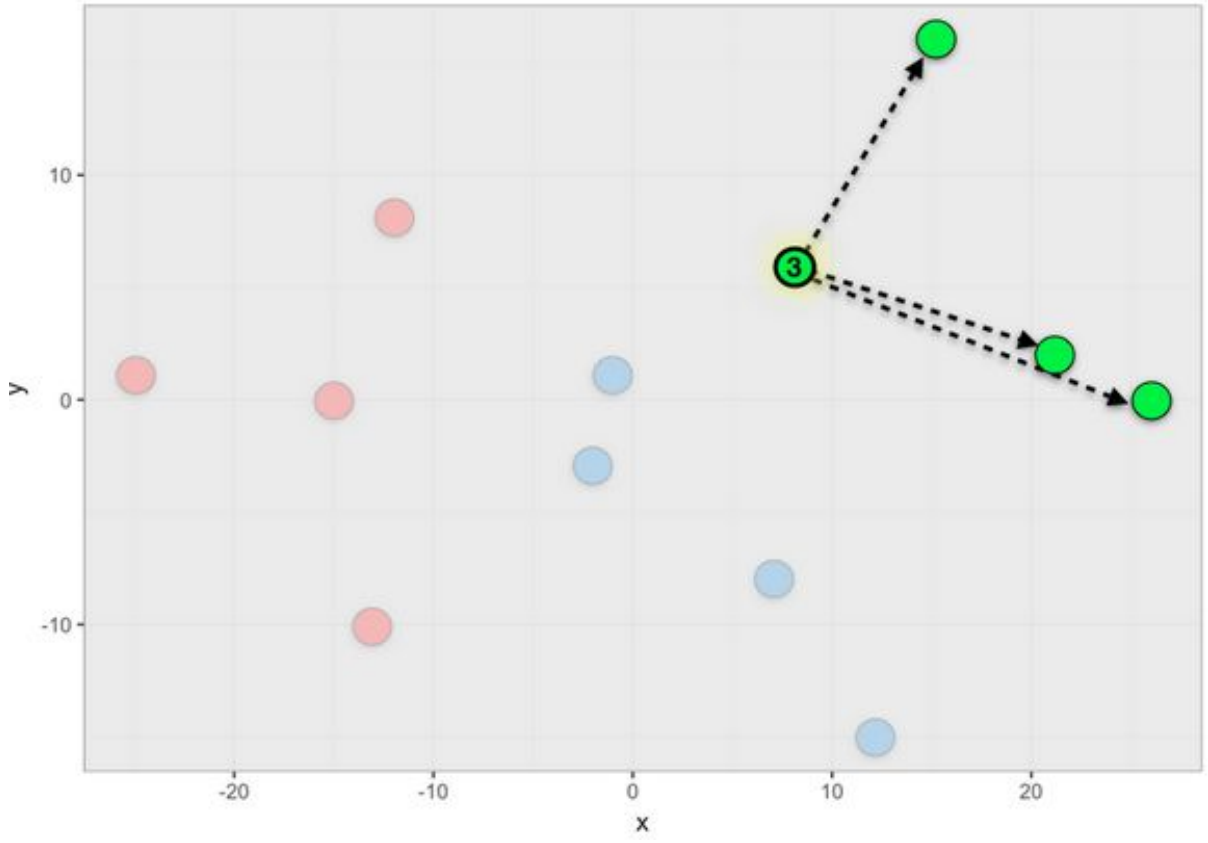
2– K-means clustering

How to Determine the Optimal K for K-Means?

Silhouette



Within Cluster Distance: C(i)

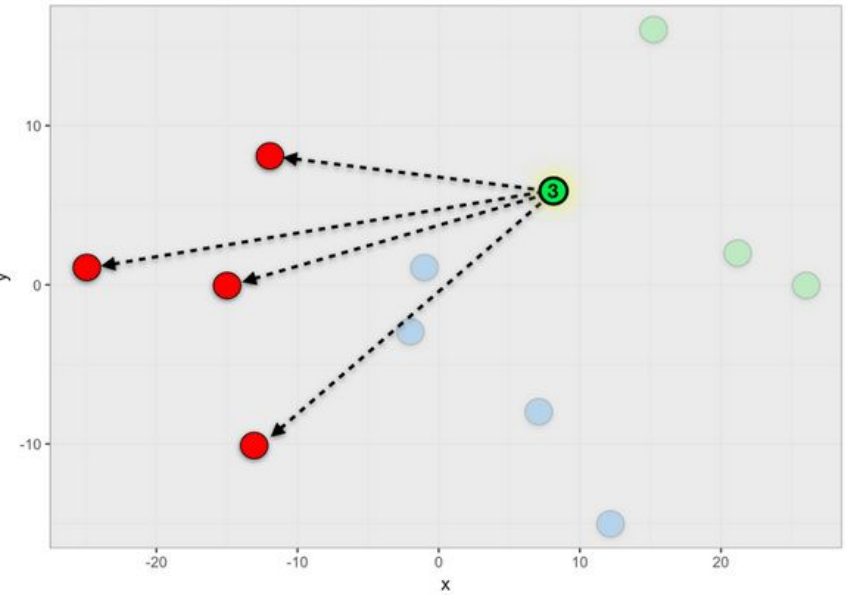


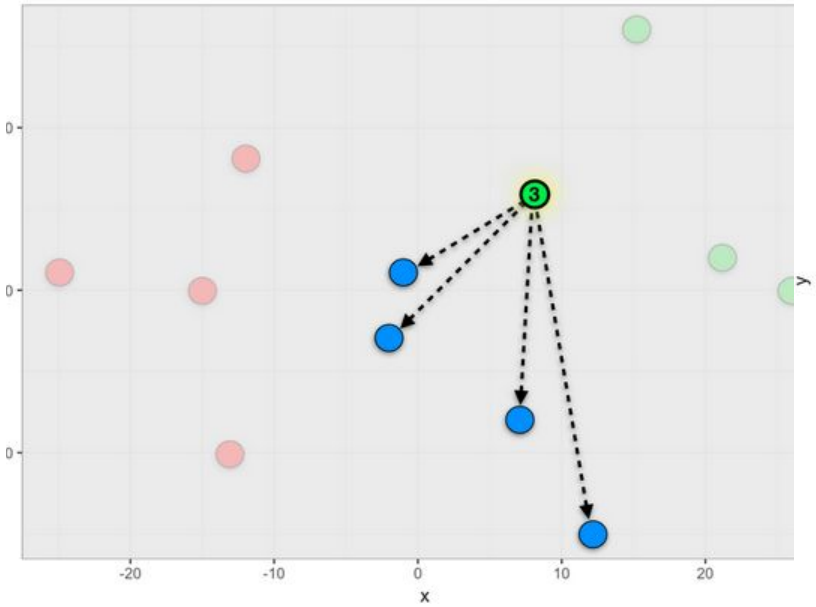
AI VIET NAM
@aivietnam.edu.vn

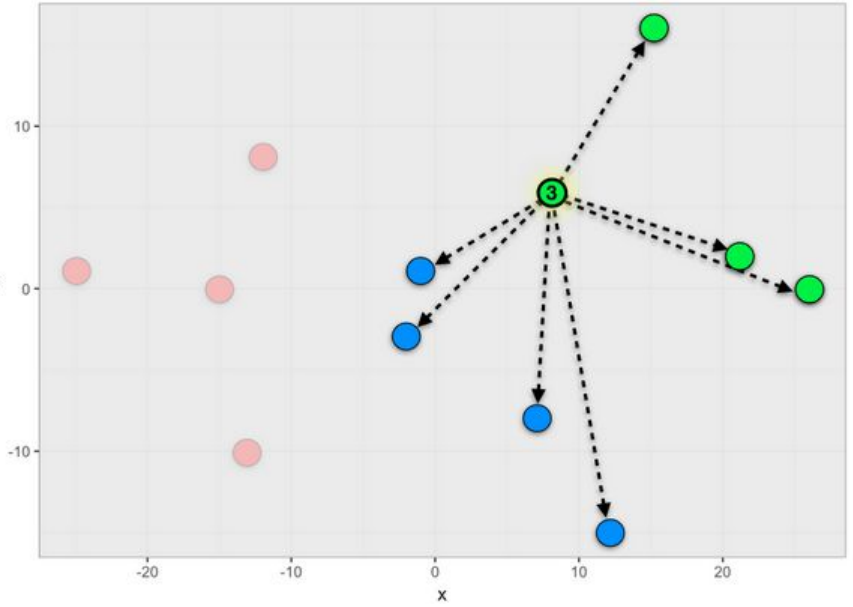
2– K-means clustering

How to Determine the Optimal K for K-Means?

Closest Neighbor Distance: N(i)





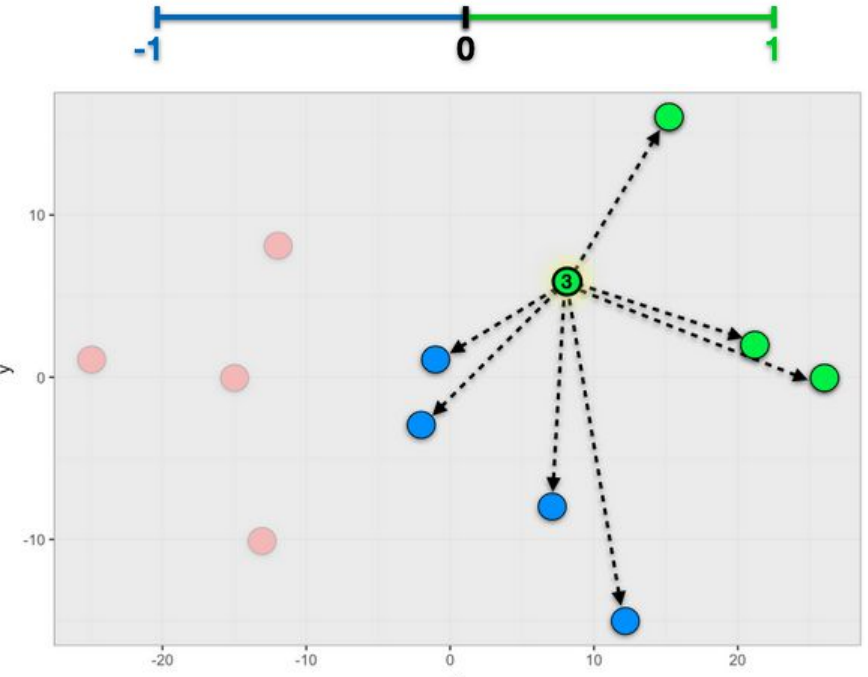


$$s(i) = \begin{cases} 1 - C(i)/N(i), & \text{if } C(i) < N(i) \\ 0, & \text{if } C(i) = N(i) \\ N(i)/C(i) - 1, & \text{if } C(i) > N(i) \end{cases}$$

AI VIET NAM
@aivietnam.edu.vn

2– K-means clustering

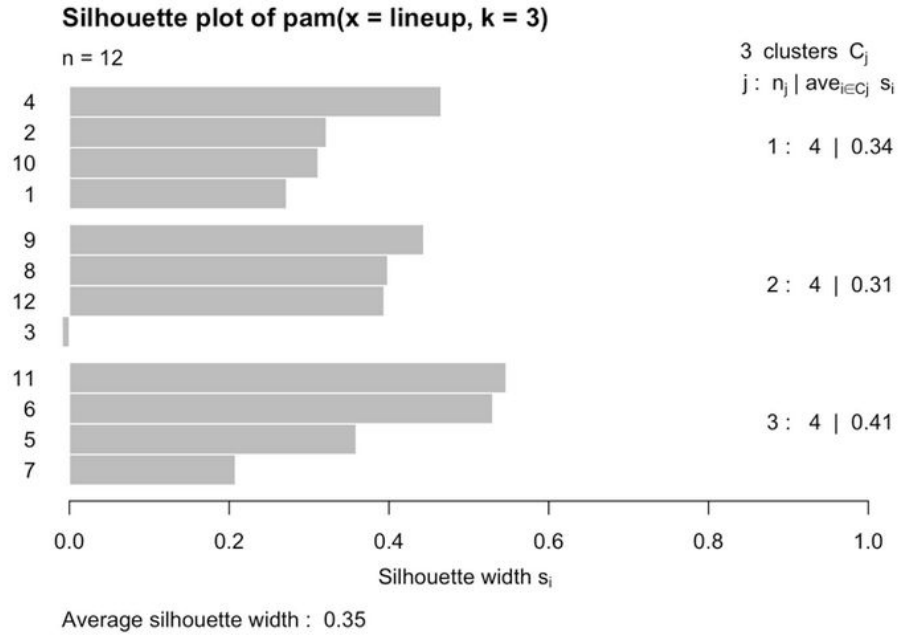
How to Determine the Optimal K for K-Means?



1: Well matched to cluster

0: On border between two clusters

-1: Better fit in neighboring cluster



1: Well matched to each cluster

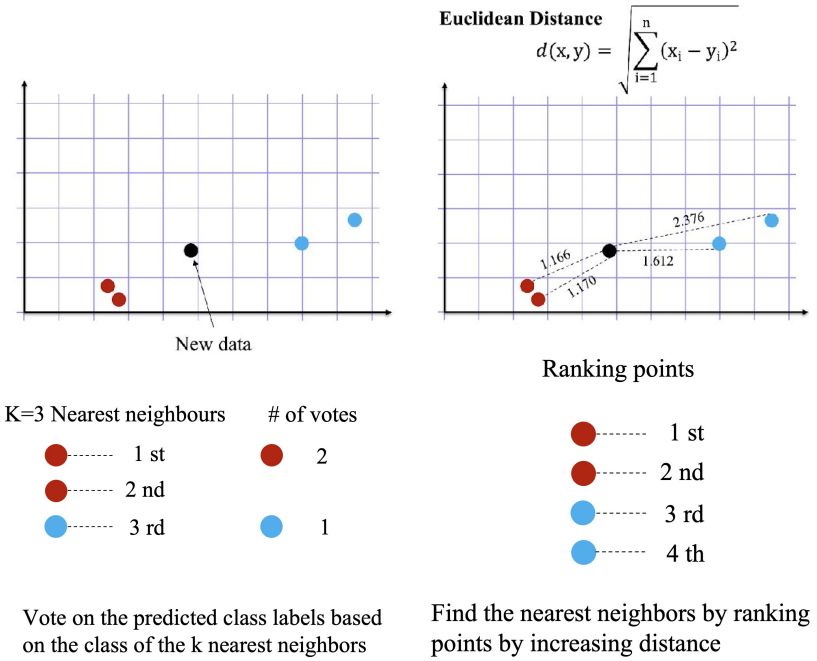
0: On border between clusters

-1: Poorly matched to each cluster

SUMMARY

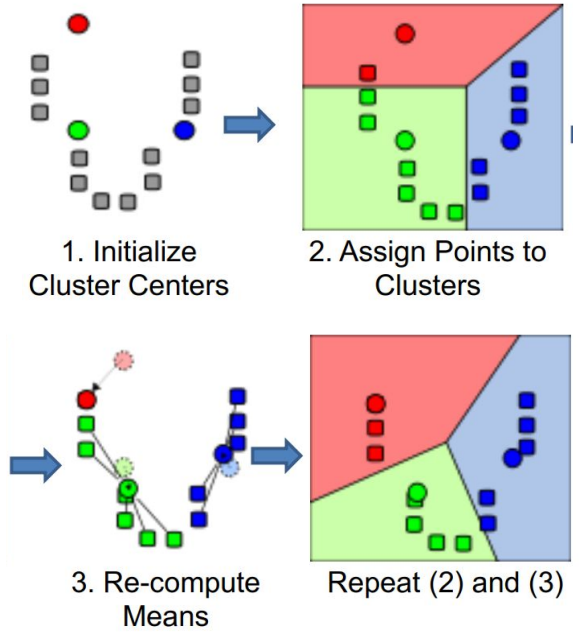
KNN

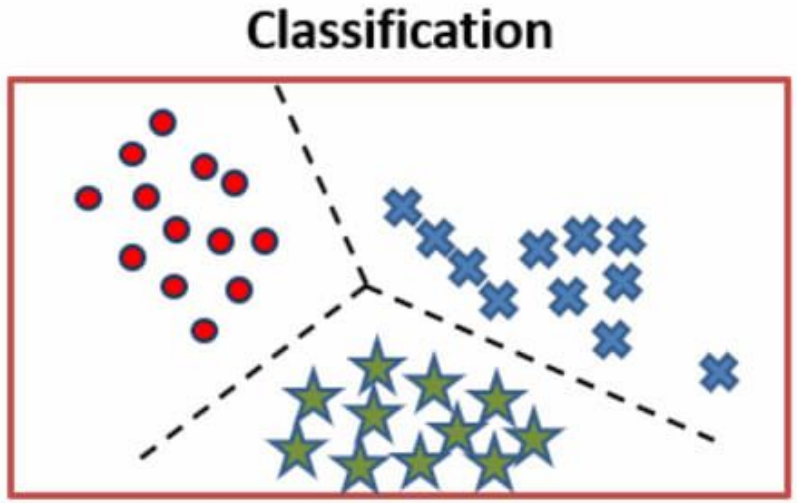
- Predicted based on K-Nearest Neighbors from the training data through Geometry Distance Functions



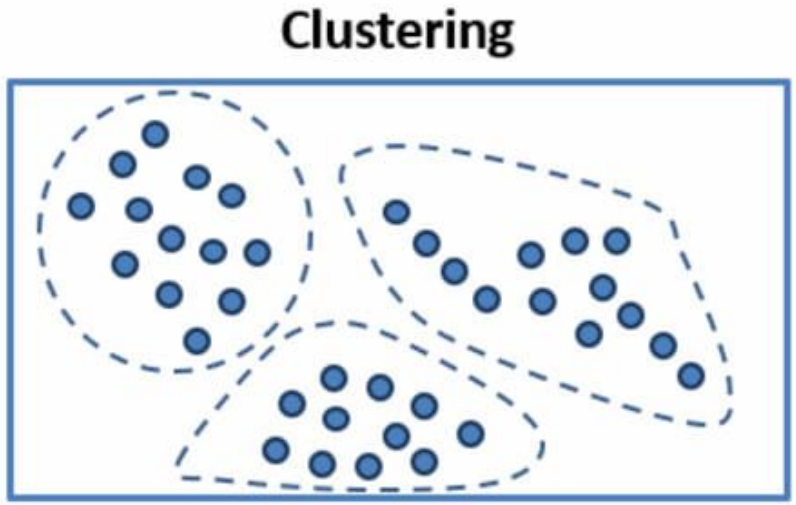
Kmeans

- A method used in data analysis to partition a dataset into K distinct, non-overlapping groups based on feature similarity.





Supervised learning



Unsupervised learning

- ✔ Provide an overview and explanation of the fundamental concepts behind the KNN
- ✔ Show how KNN is implemented in classifying data.
- ✔ Describe the use of KNN in solving regression problem.
- ✔ Showcase examples of KNN
- ✔ Explore the applications and uses of the KNN

- ✔ Offer a definition and detailed description of how K-Means clustering works.
- ✔ Detail the procedural execution of the K-Means algorithm step-by-step.
- ✔ Conduct an analysis on how K-Means is applied to cluster the Iris dataset.
- ✔ Discuss the applications of K-Means