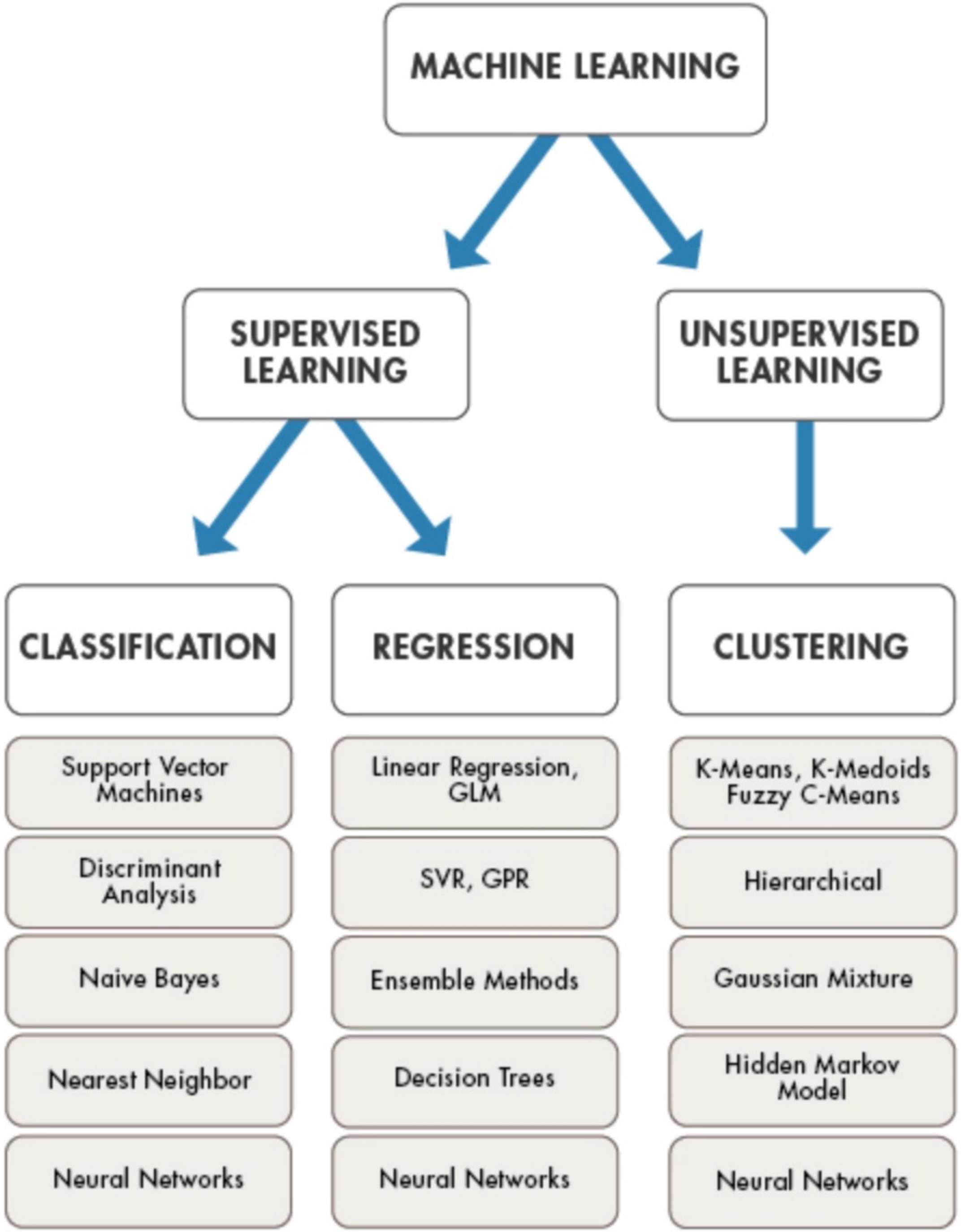
Clustering

- **Outline:**

1. Introduction

2. K-means clustering algorithm

3. Gaussian mixture model clustering algorithm


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

Clustering

- **Outline:**

  1. Introduction

  2. K-means clustering algorithm

  3. Gaussian mixture model clustering algorithm

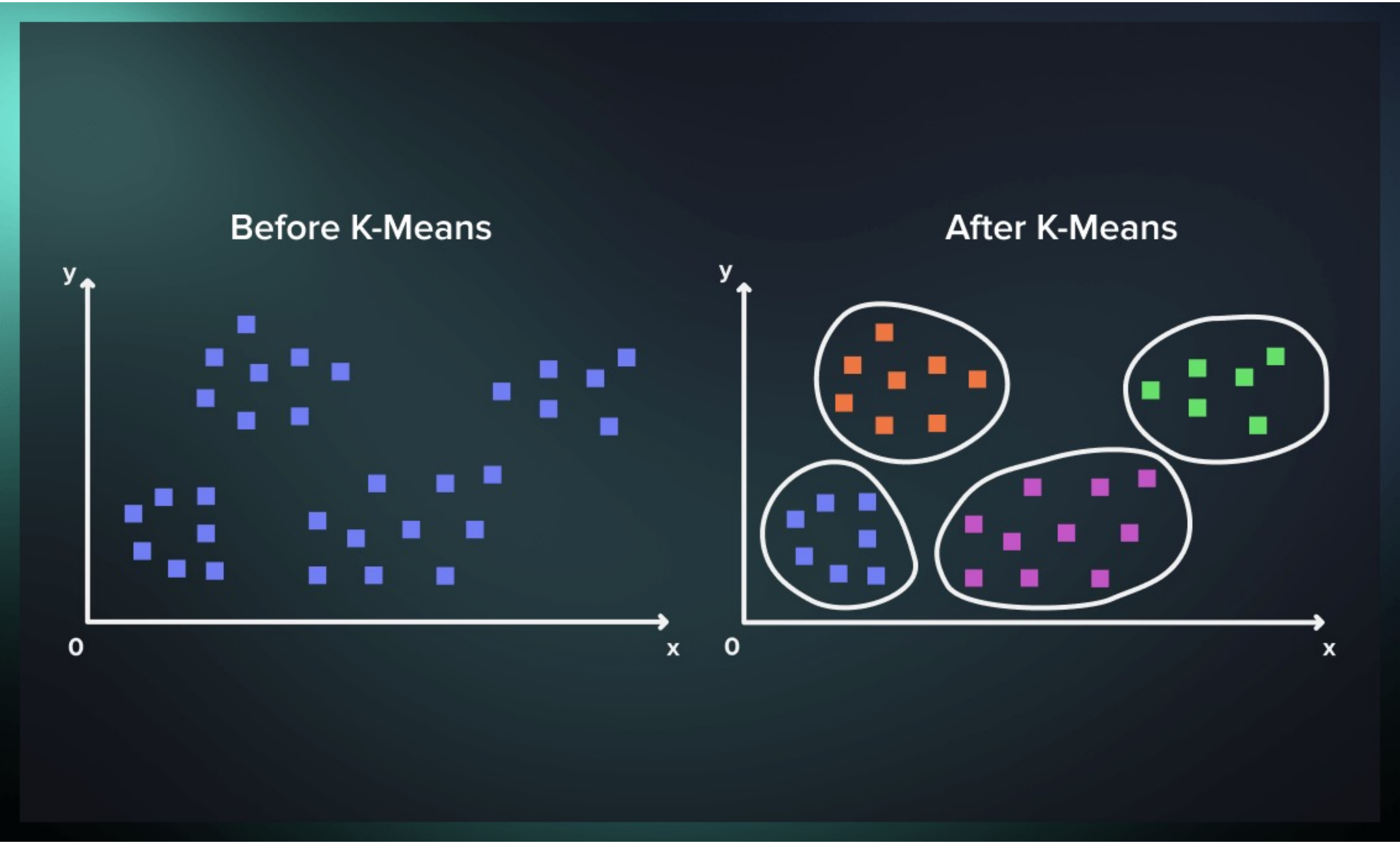
What is clustering?

- Unsupervised learning
- Input: an unlabeled dataset
- Output: **groups** (**clusters**)
- Principle: dividing the examples into a number of **groups** (**clusters**) such that examples in the same group are more similar to other examples in the same group than those in other groups.
- **Goal:** to find distinct groups or "clusters" within a data set.

Clustering

- **Duration:** 2 hrs
- **Outline:**

1. Introduction
2. **K-means clustering algorithm**
3. Gaussian mixture model clustering algorithm


Before K-Means
y
0
x
After K-Means
y
0
x

General

- K-means clustering is the most commonly used clustering algorithm.
- K-means clustering is a distance-based algorithm.
- K-means tries to to group the closest points to form a cluster (K-means tries to minimize the variance of data points within a cluster).
- K-means is best used on small data sets because it iterates over ***all*** of the data points → it'll take more time to classify data points in the large data set.

K-means clustering implementation

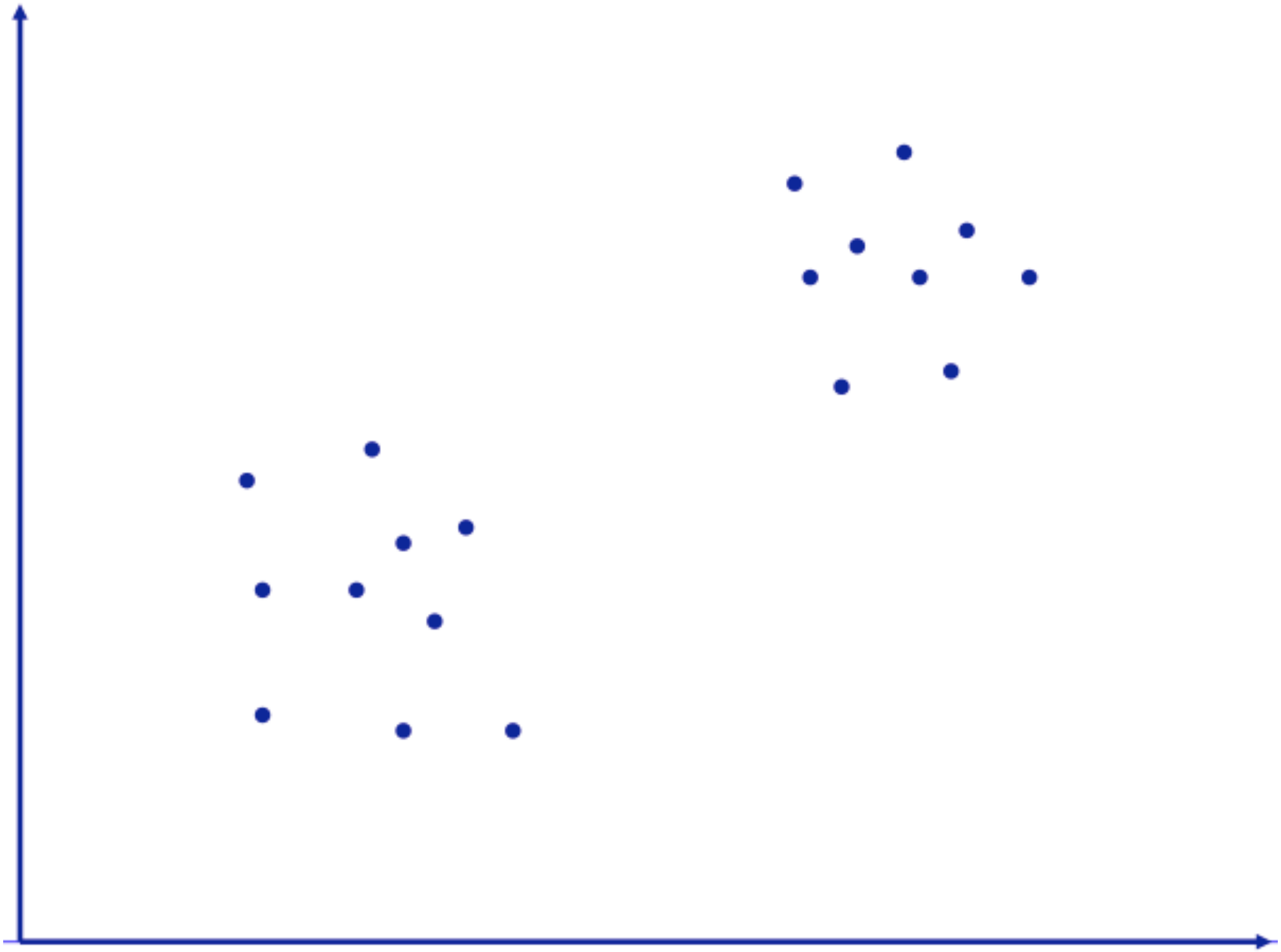
- Step 1: initialization
  - Partition the data points into K clusters randomly. Find the centroids of each cluster
- Step 2: data clustering

  For each data point:
  - Calculate the distance from the data point to each cluster
  - Assign the data point to the closest cluster

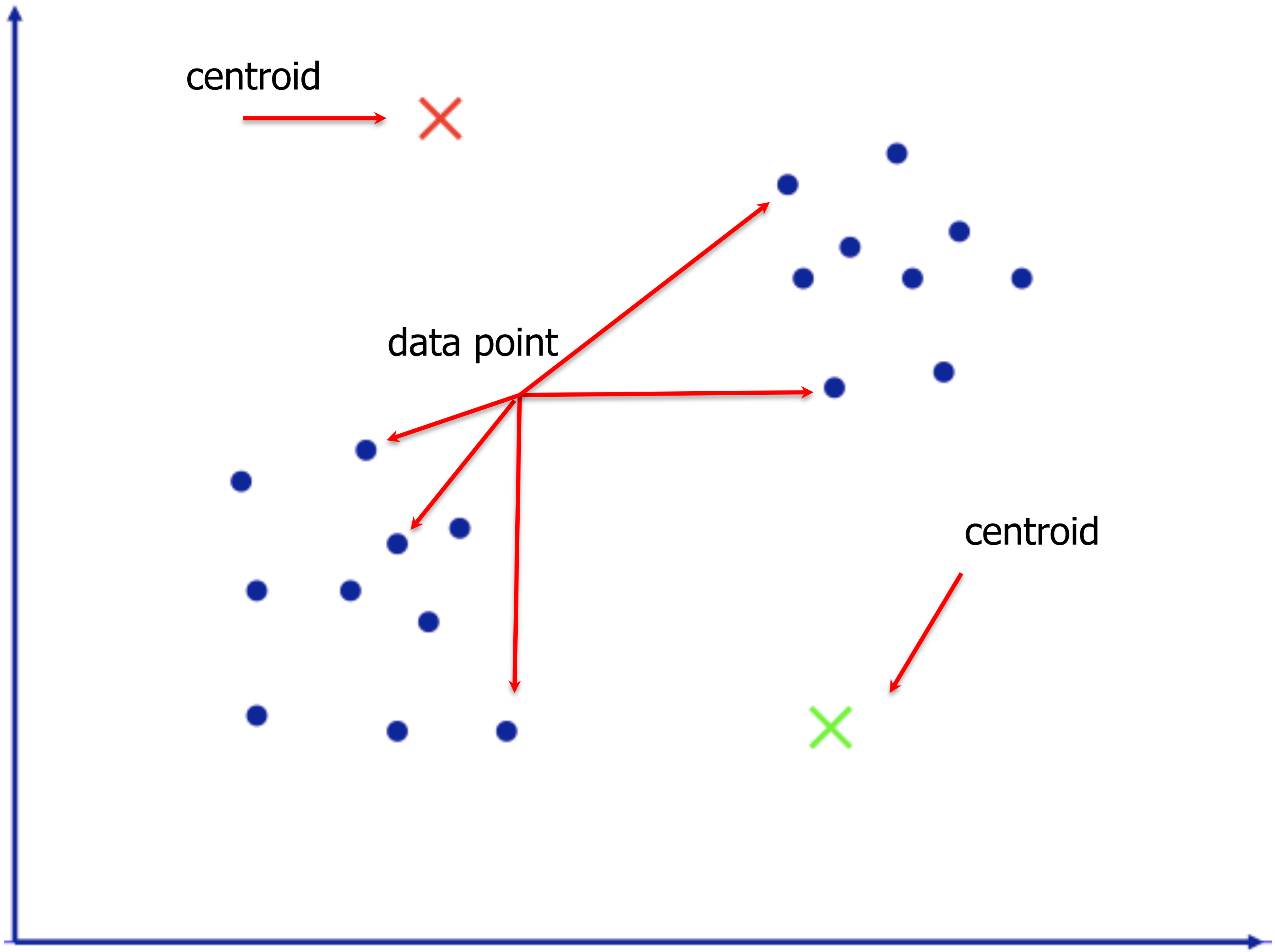
K-means clustering implementation

- Step 3: centroid determination
  - Re-compute the centroid of each cluster
- Step 4: iteration
  - Repeat step 2 and step 3 until terminated

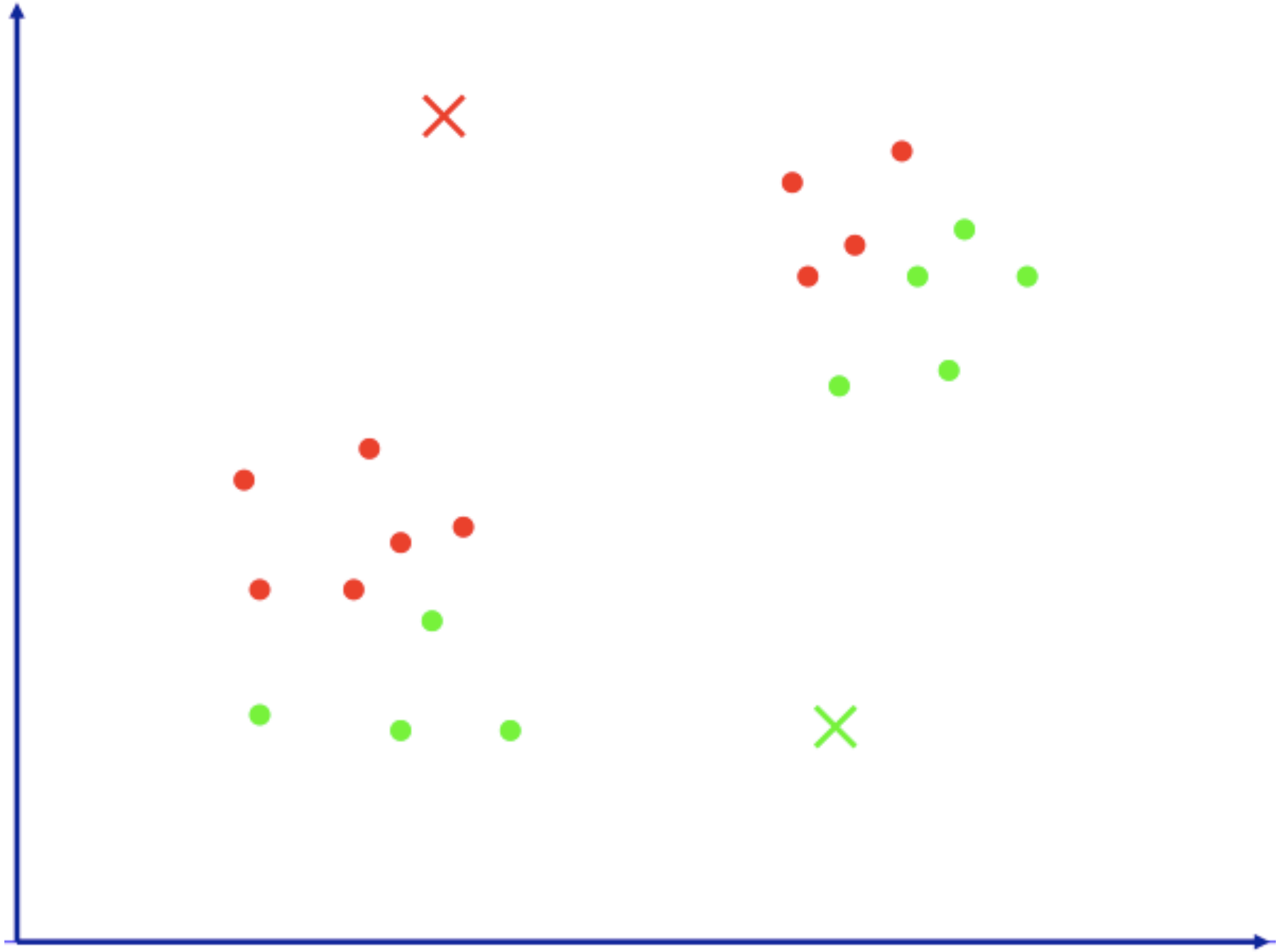
K-means clustering - illustration



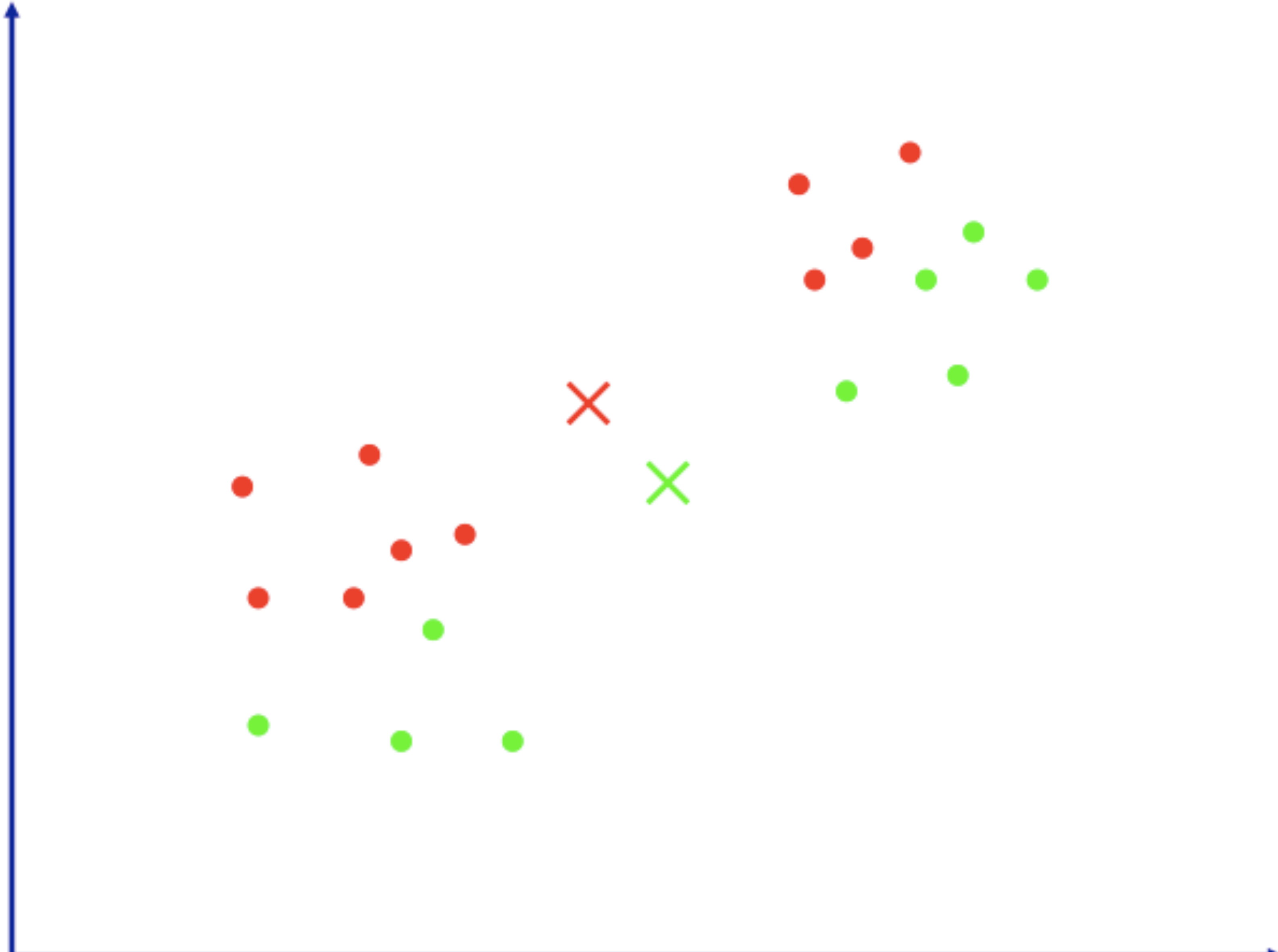
K-means clustering - illustration



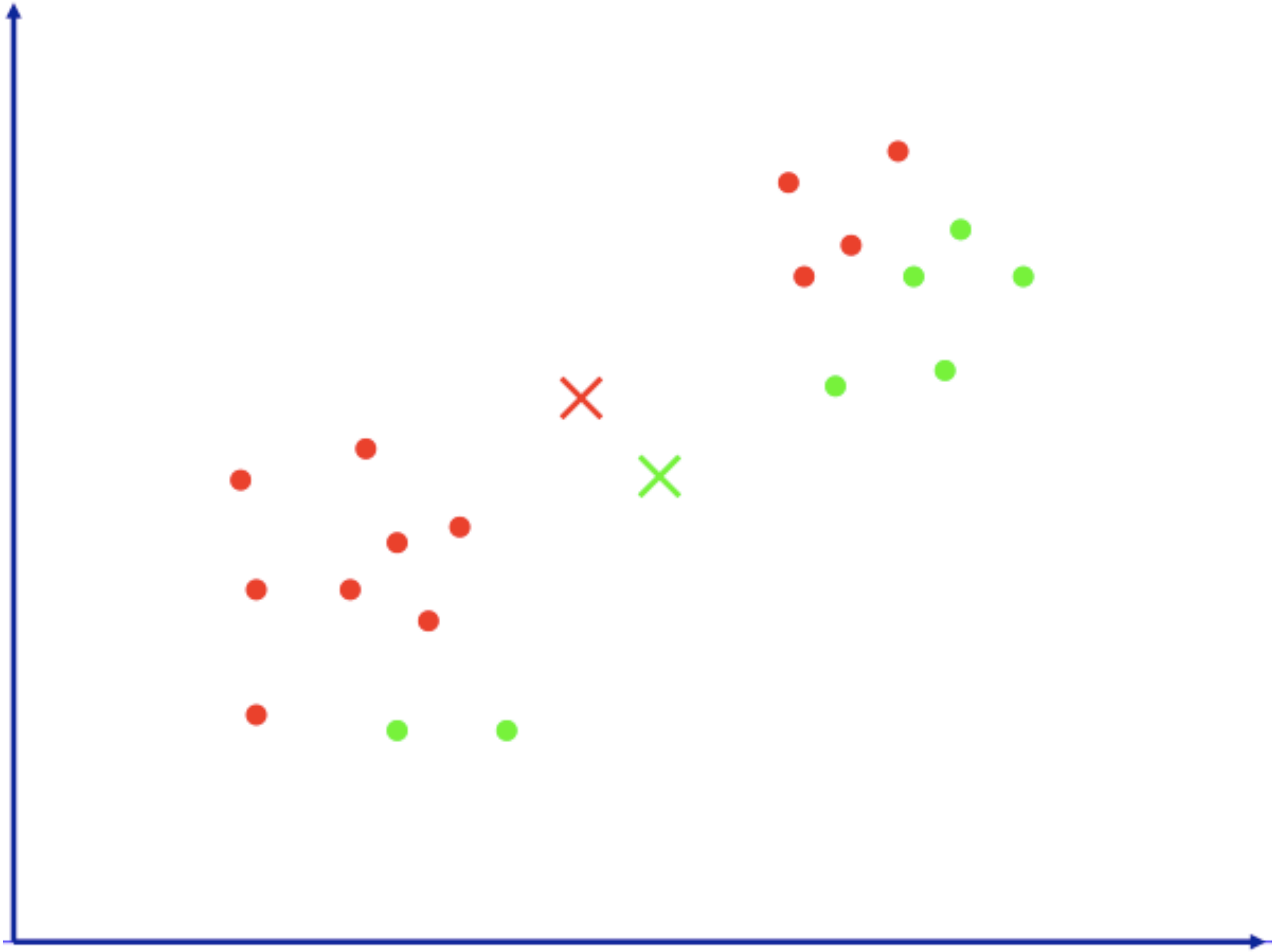
K-means clustering - illustration



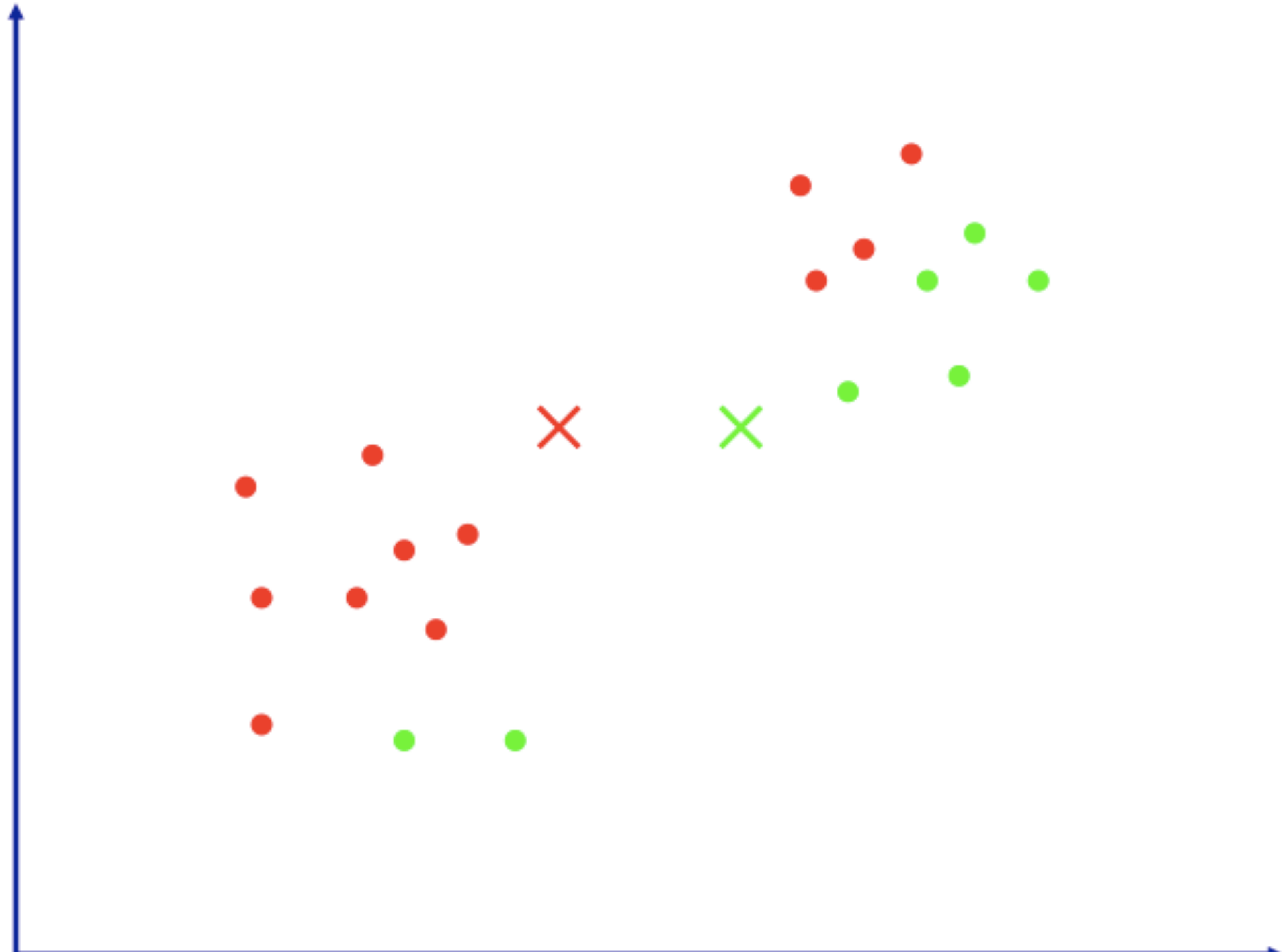
K-means clustering - illustration



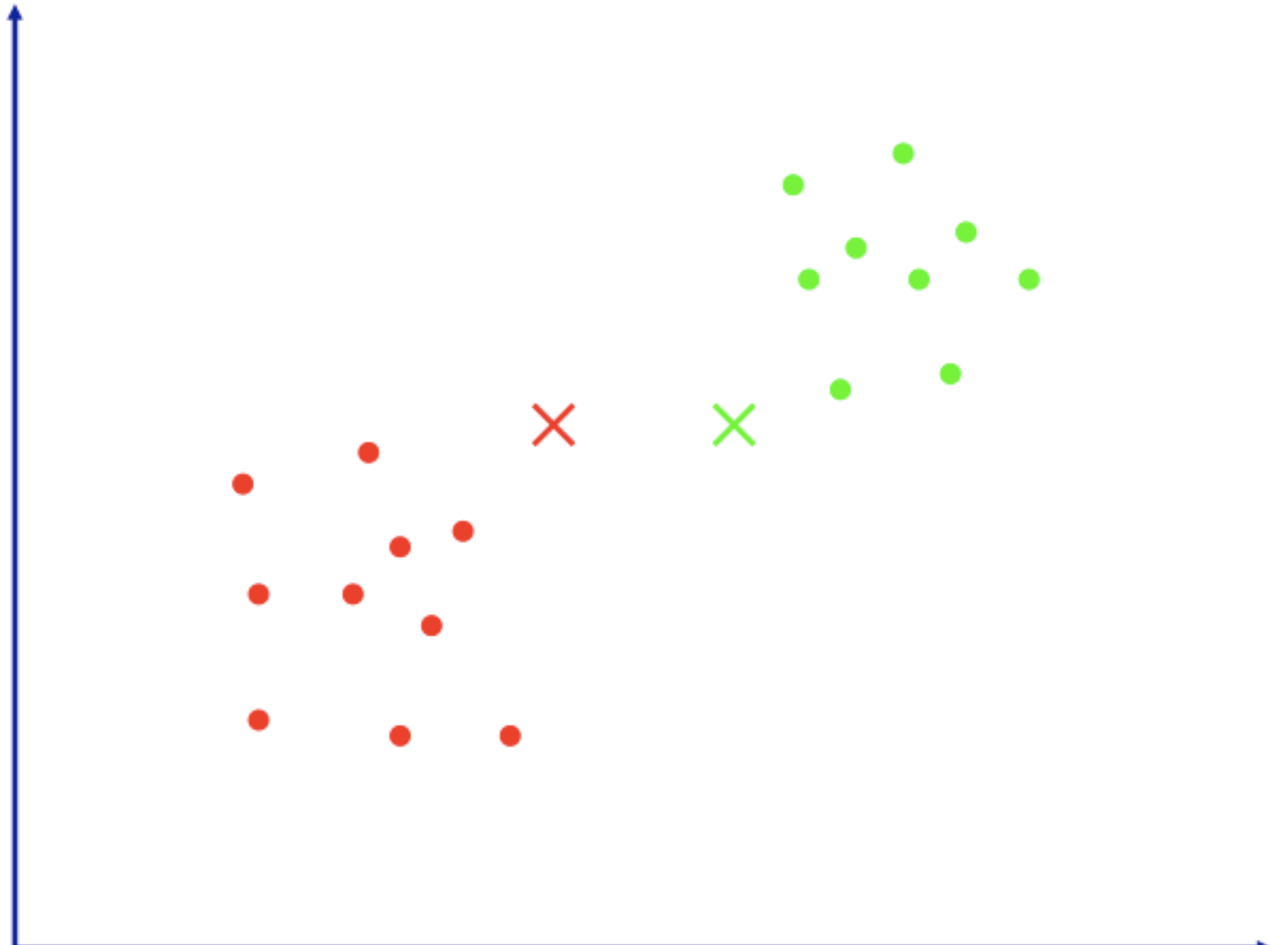
K-means clustering - illustration



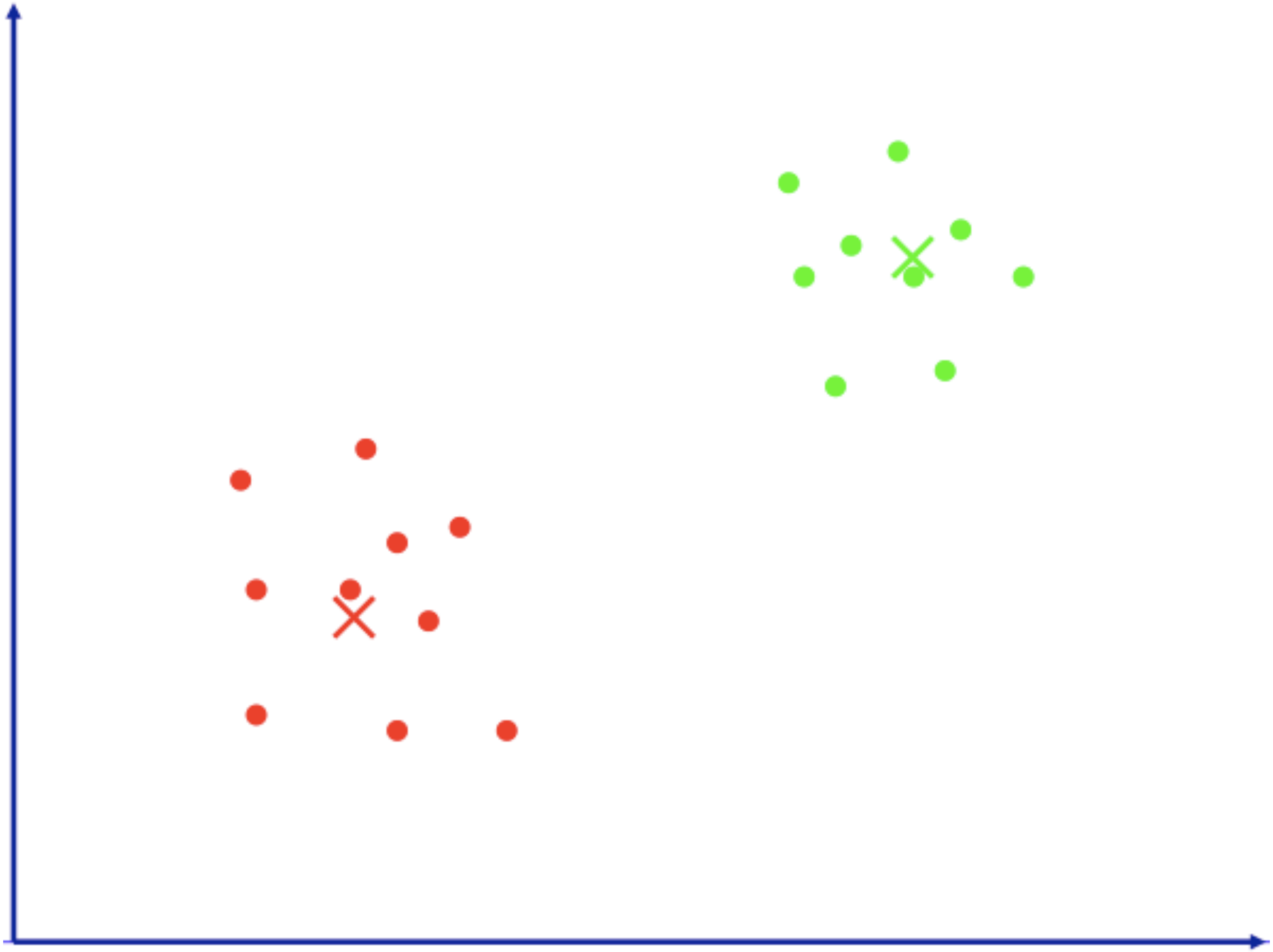
K-means clustering - illustration



K-means clustering - illustration

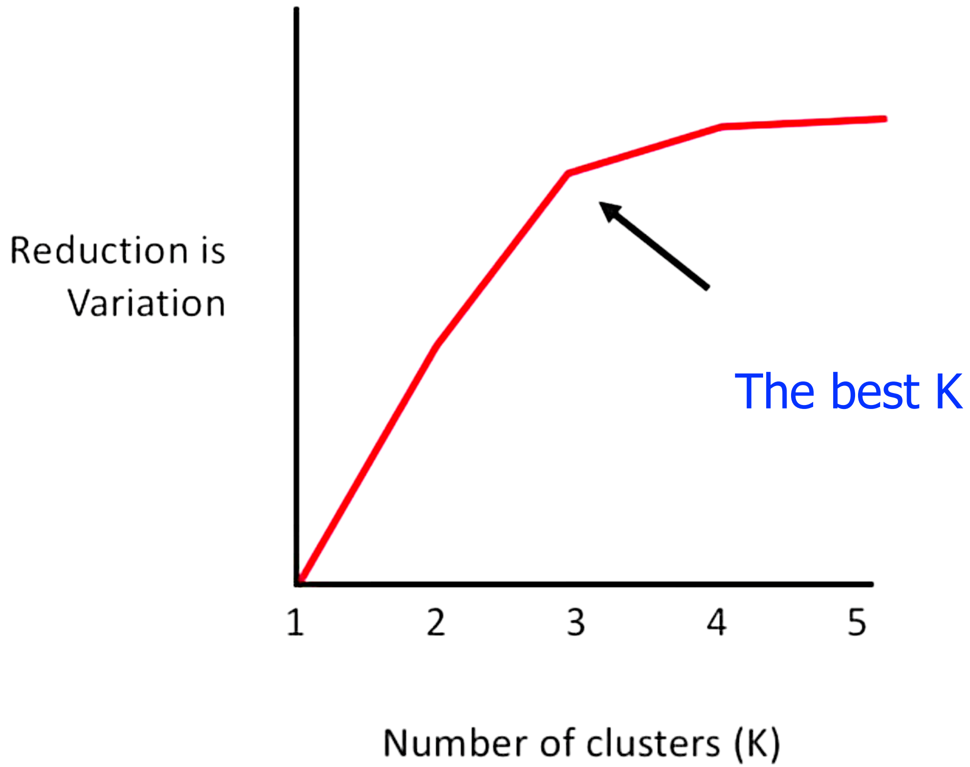


K-means clustering - illustration



How to figure out the best value of K?

- Just try the different value of K
- Check the total variation within each cluster



Bài tập áp dụng 1

- Cho 2 trọng tâm của 2 cụm (cluster) của dữ liệu 2D như sau:
  - Centroid của cụm 1: (1,5)
  - Centroid của cụm 2: (4,1)
- Giả sử có 3 mẫu dữ liệu A, B, C có các vector đặc trưng lần lượt là: (1.1,1.2), (2.0,3.0) và (6.3,1.5)
- Cho biết các mẫu dữ liệu này thuộc về cụm nào?

Bài tập áp dụng 2

- Cho ảnh sau:

- Bằng phương pháp K-means clustering với K = 3, hãy trích ra bông hoa trên nền đen như ảnh sau:

Bài tập về nhà

- Ứng dụng phương pháp Kmeans clustering phát hiện quả chín trên cây.

Bài tập về nhà (tt)

- Ứng dụng phương pháp Kmeans clustering phát hiện quả chín trên cây.

Clustering

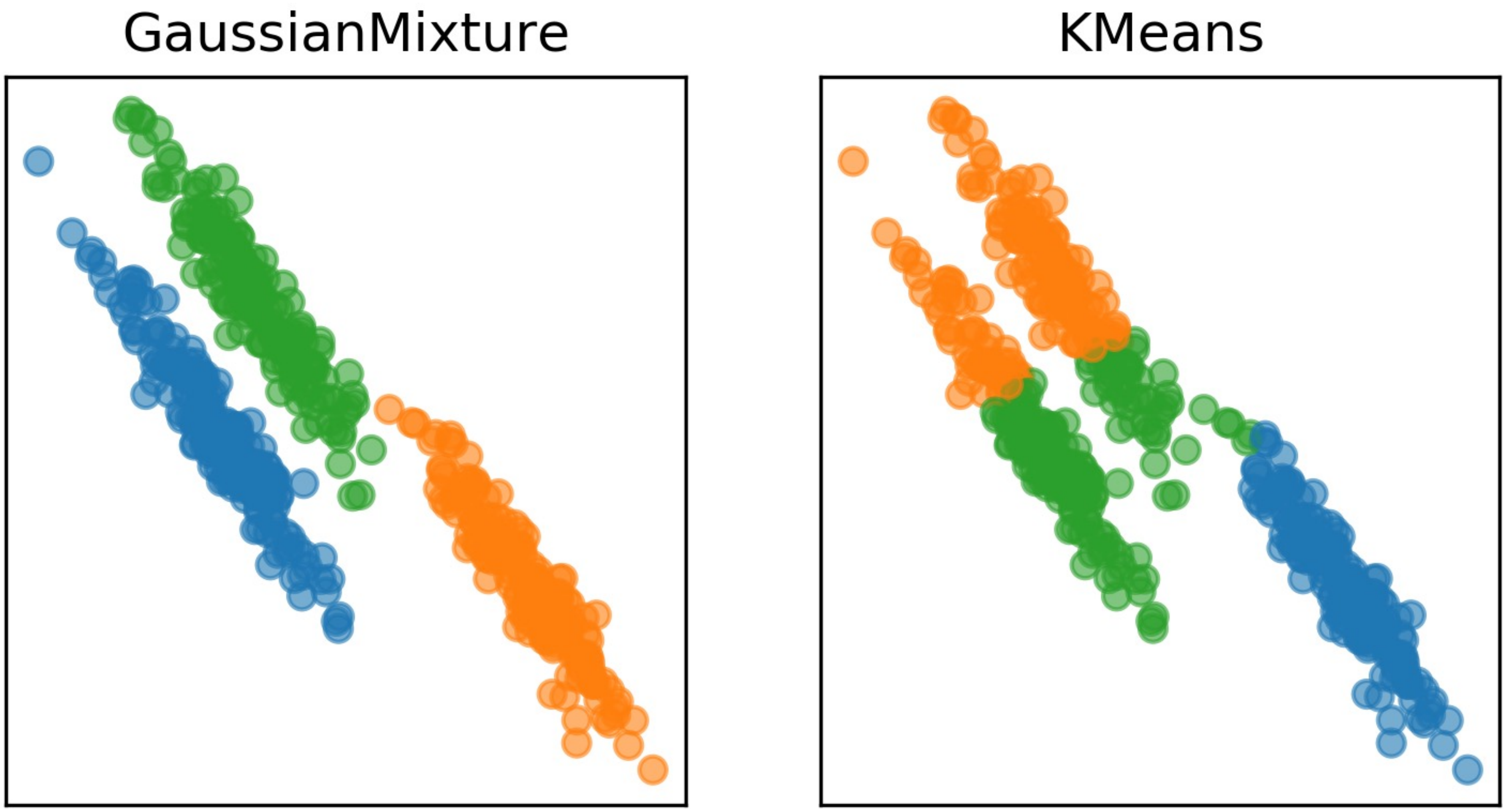
- **Duration:** 2 hrs
- **Outline:**

1. Introduction

2. K-means clustering algorithm

3. **Gaussian mixture model clustering algorithm**

Drawback of K-means



GMM clustering

- GMM clustering is a powerful clustering algorithm.
- GMM clustering is distribution-based.

Gaussian distribution

- Gaussian distribution ≡ Normal distribution
- Gaussian distribution has a bell-shaped curve.
- The data points symmetrically distributed around the mean value.

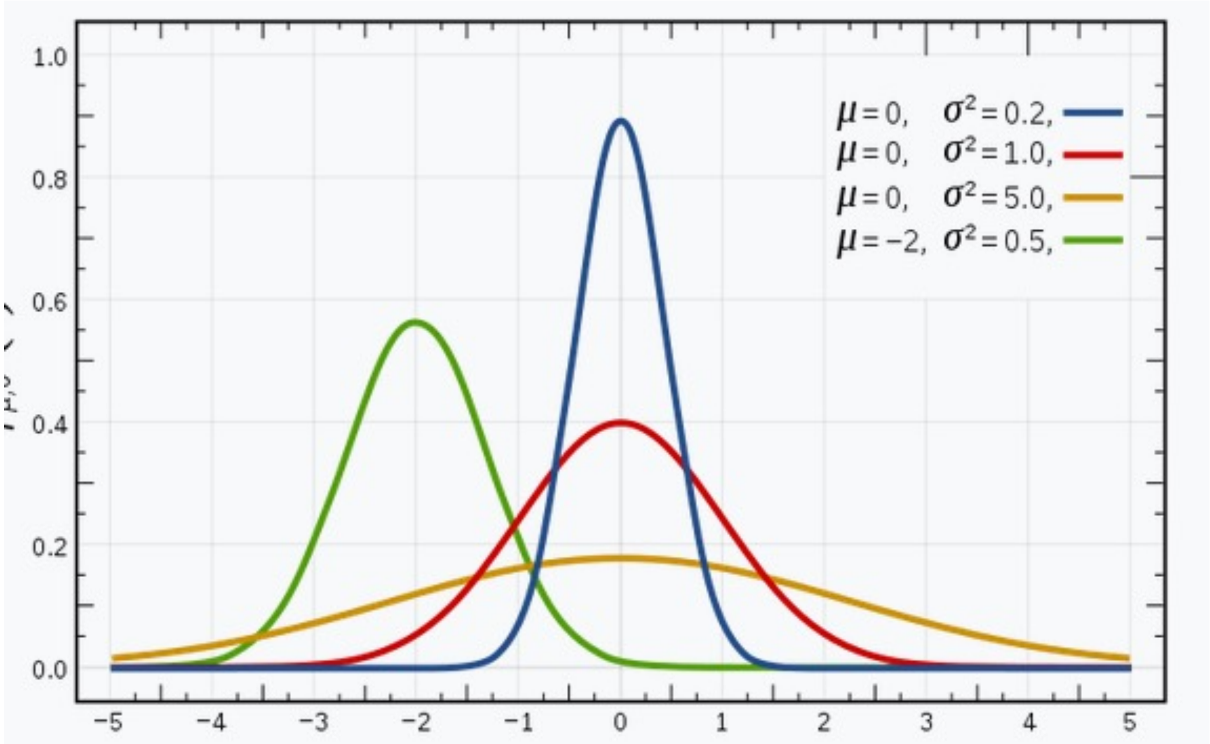
1D Gaussian pdf

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

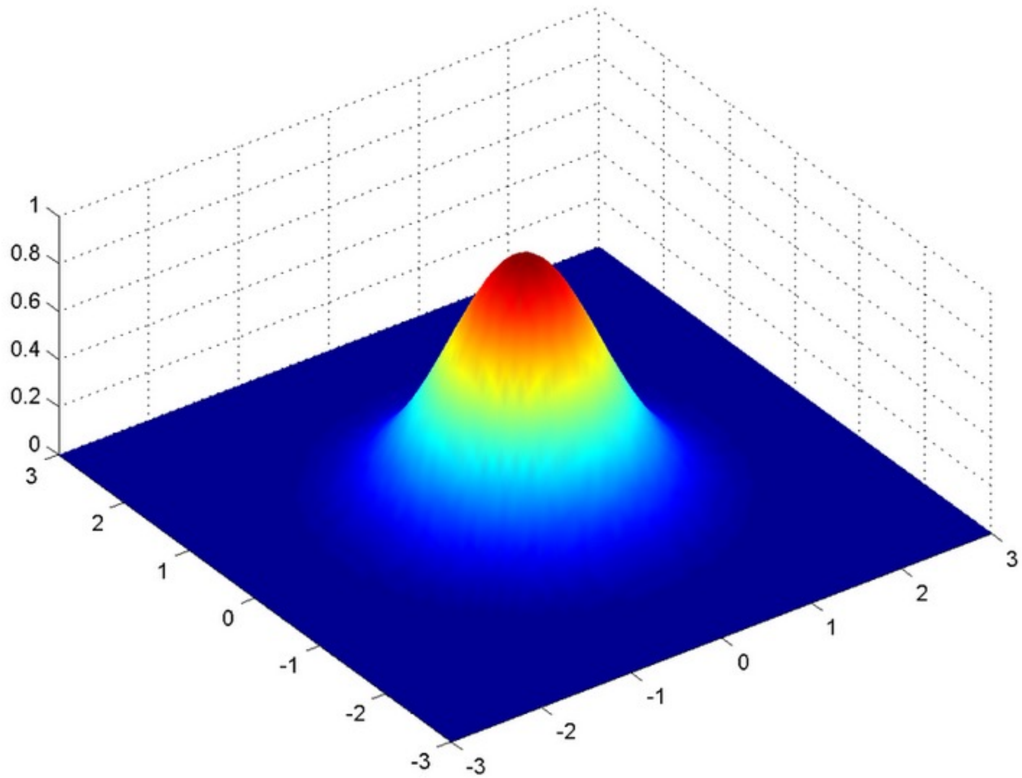
x: input data

μ: mean

σ^2: variance.



2D Gaussian pdf



x: input vector (length = 2)

μ: mean vector (length = 2)

Σ: 2 × 2 covariance matrix

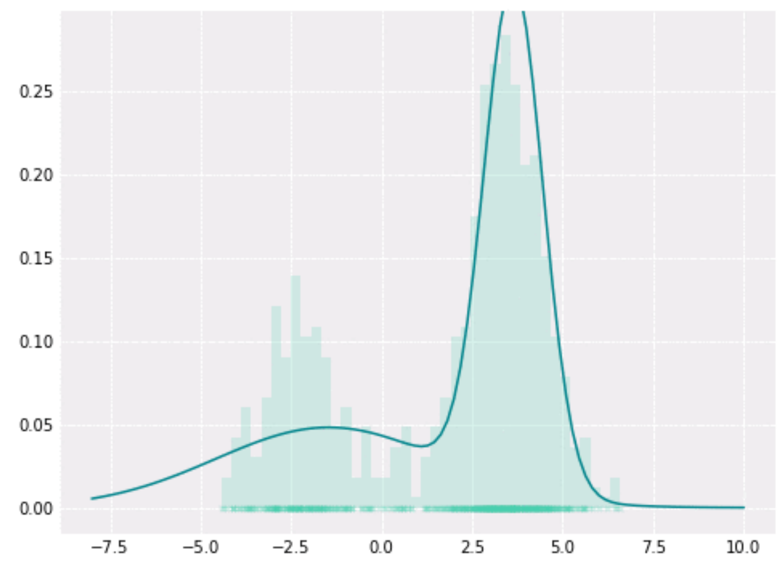
$$f(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Gaussian mixture model

- Linear combination of *M* Gaussian distributions
- pdf of GMM:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \; g(\mathbf{x}|\mu_i, \Sigma_i),$$

Ex: M = 2



- $\mathbf{x}$: D-dimension data
- $\omega_i: mixing\ coeffcients,$
- $1 \leq \omega_i \leq M\ for\ all\ i = 1, \ldots, \mathrm{M}$

and $\sum_{i=1}^{M} \omega_i$ = 1

- $g: Gaussian\ density\ components$

GMM clustering alogrithms

- GMM parameters:
  - number of Gaussian components (M)
  - weights (ω_i)
  - Gaussian components (mean μ, covariance Σ)
- GMM assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions, and each of these distributions represent a cluster → tends to group the data points belonging to a single distribution together.

GMM clustering algorithm

- GMM training input:
  - number of Gaussian components (*M*) ≡ number of clusters
  - training data points (***x***)
- Goal: to model this data using GMM
  - Mixing coefficients $\omega_1, \omega_2, \ldots, \omega_M$
  - Mean $\mu_1, \mu_2, \ldots, \mu_M$
  - Covariance $\Sigma_1, \Sigma_2, \ldots, \Sigma_M$
- Solution: EM algorithm

Expectation-Maximization (EM) algorithm

- EM is a statistical algorithm for finding the right model parameters.
- EM is used when the data has missing values (latent variables).
- EM tries to use the existing data → determine the optimum latent variables → find the model parameters → go back and update the latent variable, and so on.

> E-step: the available data is used to estimate (guess) the values of the missing variables

> M-step: based on the estimated values generated in the E-step, the complete data is used to update the parameters

GMM-based motion detection

https://www.youtube.com/watch?v=0nz8JMyFF14&t=844s