MAT8034: Machine Learning

Clustering and the K-means Algorithm

Fang Kong

https://fangkongx.github.io/Teaching/MAT8034/Spring2025/index.html

Part of slide credit: Stanford CS229

Outline

- K-means algorithm
- Convergence analysis

Unsupervised learning

- In previous lectures, we consider the supervised learning with training set

$$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$$

- Now, consider the unsupervised learning with training set

$$\{x^{(1)}, \dots, x^{(n)}\}$$

- Hope to group the data into a few cohesive "clusters"

The k-means clustering algorithm

- 1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ randomly.

  2. Repeat until convergence: {

     For every i, set

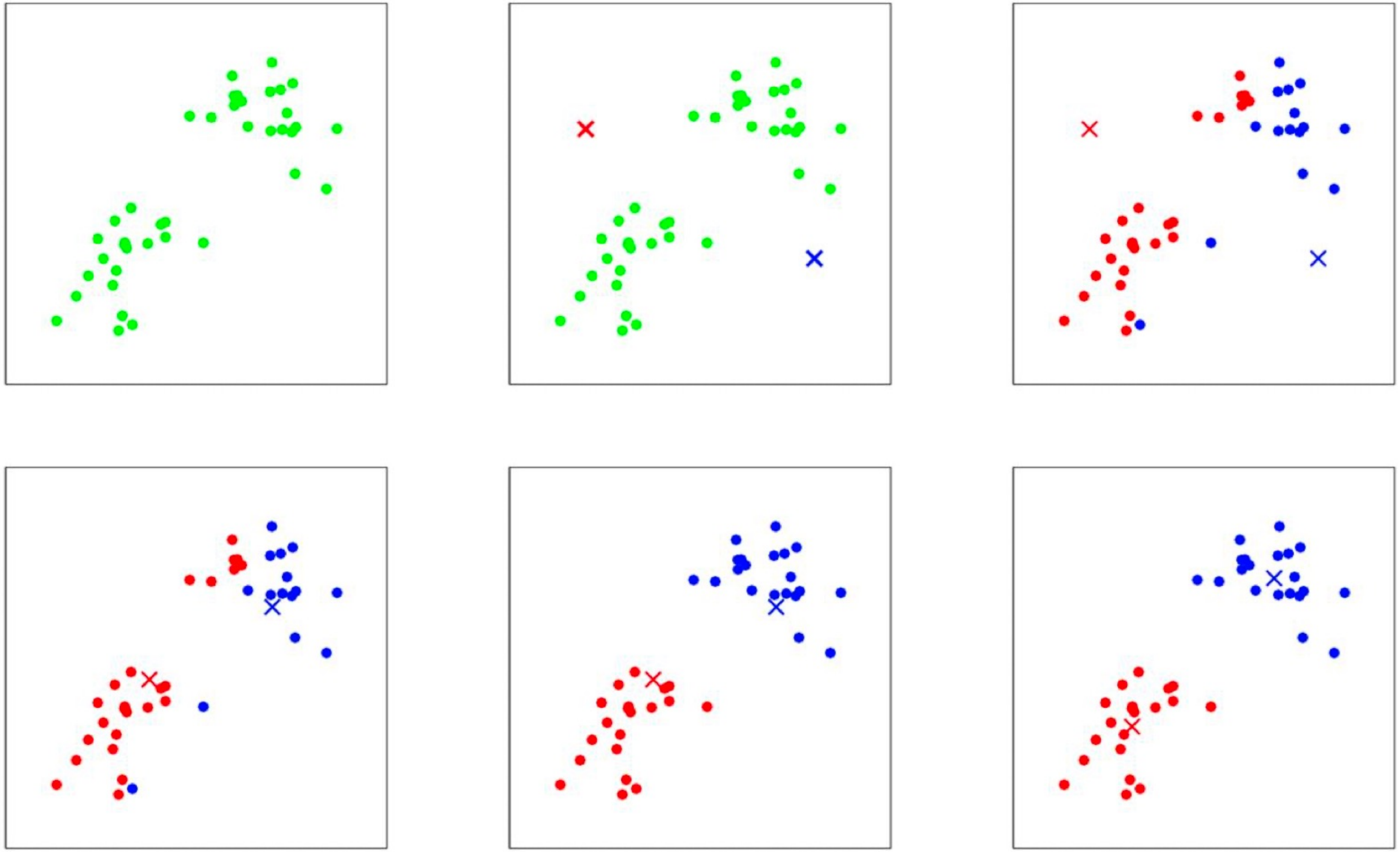
$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

     For each j, set

$$\mu_j := \frac{\sum_{i=1}^n 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^n 1\{c^{(i)} = j\}}.$$

     }

The k-means clustering algorithm: Illustration

- 

Convergence analysis

- Is the k-means algorithm guaranteed to converge?

- The procedure of K-means (in each loop):
  - Fix cluster centroids μ, minimize the distance between $x^{(i)}$ and $\mu^{c(i)}$ by optimizing $c(i)$
  - Fix $c(i)$, minimize the distance between $x^{(i)}$ and $\mu^{c(i)}$ by optimizing μ

Convergence analysis (cont'd)

- Define the distortion function

$$J(c, \mu) = \sum_{i=1}^{n} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

- K-means is exactly coordinate descent on J
- J must monotonically decrease, and the value of J must converge

Convergence analysis (cont'd)

- Define the distortion function

$$J(c,\mu) = \sum_{i=1}^{n} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

  - J is a non-convex function, and so coordinate descent on J is not guaranteed to converge to the global minimum
  - K-means can be susceptible to local optima
  - It typically performs well
  - Tricks: run many with different initial values, pick the one with the lowest distortion J

Summary

- K-means algorithm
  - Fix cluster centroids μ, minimize the distance between $x^{(i)}$ and $\mu^{c(i)}$ by optimizing $c(i)$
  - Fix $c(i)$, minimize the distance between $x^{(i)}$ and $\mu^{c(i)}$ by optimizing μ
- Convergence analysis
  - Each loop is an exactly coordinate descent on the distortion function