# K-means Clustering

Lesson 7 – Section 5

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

## K-means Clustering

Partitional clustering approach

- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

# K-means Clustering Algorithm

1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

# Performance Metrics of K-Means Clustering: SSE

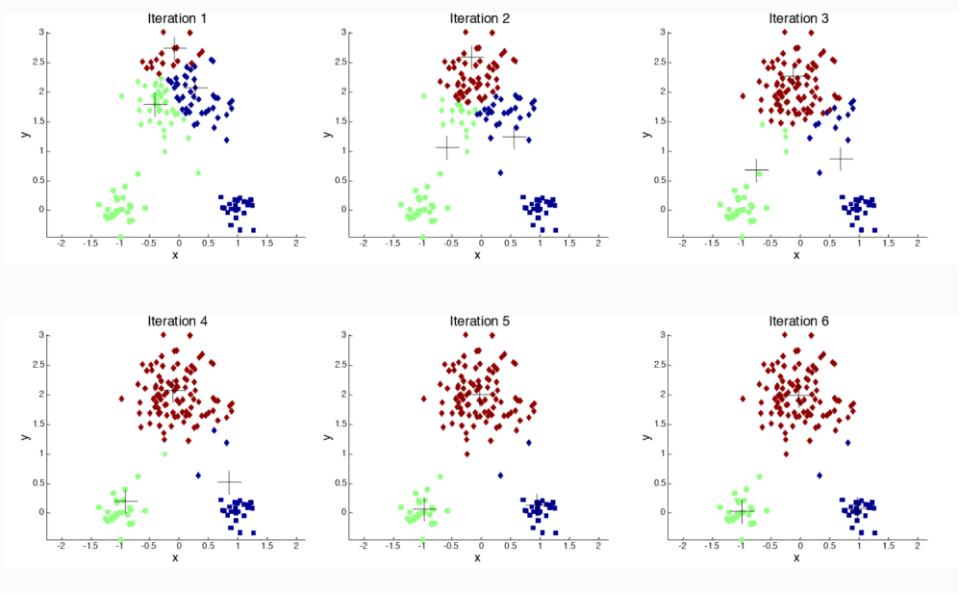
Suppose the centroid of cluster $C_j$ is $m_j$

1. For each object x in $C_m$, compute the squared error between x and the centroid $m_j$
2. Sum up the error of all the objects

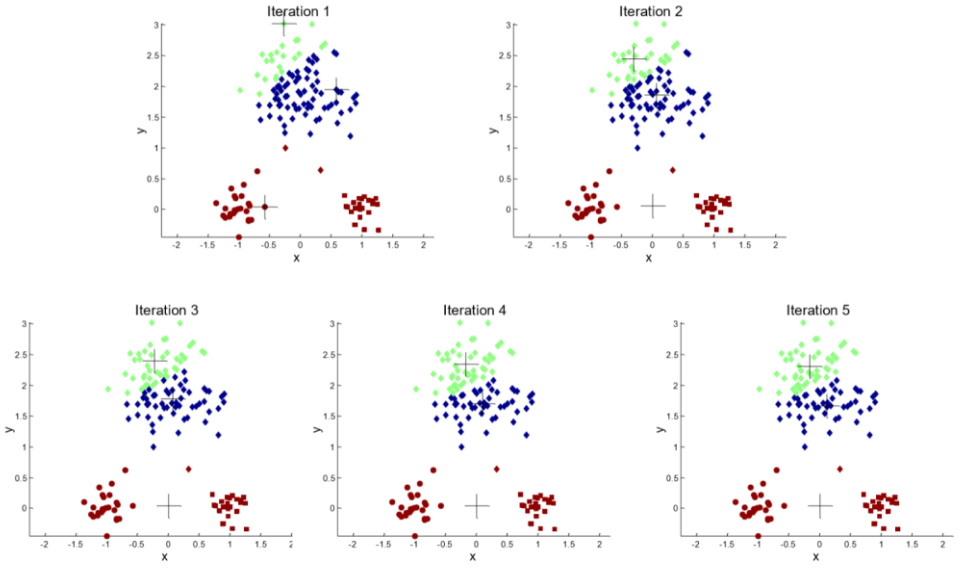
$$SSE = \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$


$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

# Importance of Choosing Initial Centroids




# Importance of Choosing Initial centroids



## Solutions to Initial Centroids Problem

In order to solve the centroid initialization problem, usually we do k-means for multiple times with fixed k

- Each time calculate SSE
- Choose the run with the minimal SSE as the final clustering result



## Data Preprocessing in K-means Clustering

- K-means clustering requires all variables are numerical
- Numerical variables need to be scaled to remove the impact of scales of different variables
- Non-numerical variables?
  - Ordinal non-numerical variable, reasonable to represent the values as 1, 2, 3, …. For instance, education middle school, high school, college, masters, Ph.D. can be replaced by 1, 2, 3, 4, and 5
  - Other non-numerical variable, one-hot encoding.

# K-Means Clustering Discussion

## Non-numeric data

Feature values are not always numbers

- Example
  - Boolean Values: Yes or no, presence or absence of an attribute
  - Categories: Colors, educational attainment, gender

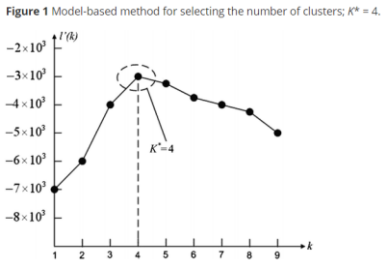
How do these values factor into the computation of distance?



# K-Means Clustering: How to Determine K?

- Is minimizing SSE a good way to choose K?
  - –If you make each observation as a single cluster, SSE=0

$$SSE = \sum_{j} \sum_{x \in C_j} (x - m_j)^2$$




- Consider regularization:
  - –We can choose to minimize

$$\sum_{j=1}^{k} \sum_{x \in C_j} (x - m_j)^2 + \lambda \times N_k$$

for k = 1, 2, ..., K, where K is a reasonably maximal possible number of clusters,

$N_k$ : number of independent parameters to be estimated in the k models, assuming that each cluster is generated by an underlying multivariate normal distributed model

## Summary

>K-means in details

>Practiced K-means in Python

# Performance Metrics and Clustering Analysis

Lesson 7

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON