CS 456 Data Mining

Central Washington University

Today …

- K-means

Outline

- What is clustering?
- Why would we want to cluster?
- How would you determine clusters?
- How can you do this efficiently?

K-means Clustering

What is Clustering?

Clustering is a technique in unsupervised machine learning that groups similar items together based on their features. This method is used to discover patterns and structures within datasets, where items in the same cluster are more alike than those in different clusters.

Why would we want to cluster?

We use clustering to uncover hidden patterns and structures in data by grouping similar items together. This approach helps in understanding and summarizing large datasets, identifying relationships among data points, and making informed decisions based on these groupings.

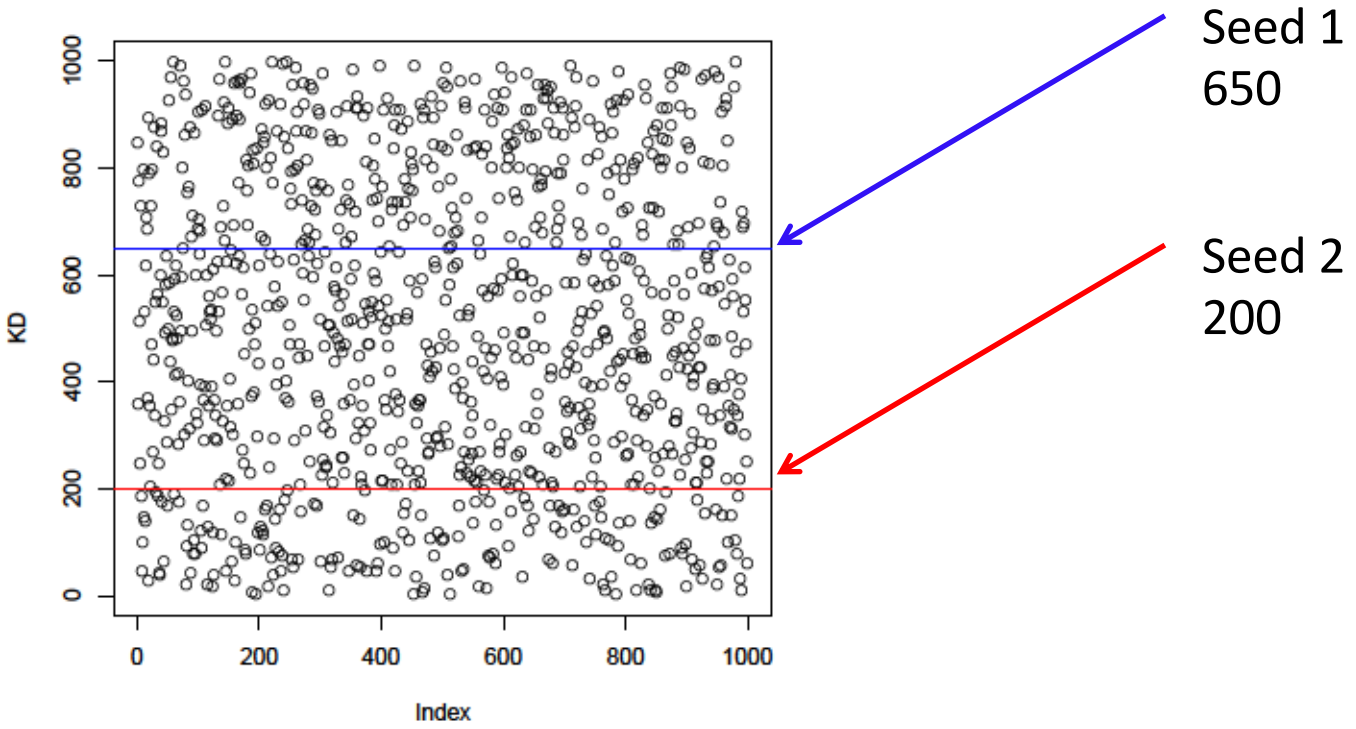
K-means Clustering

- Strengths
  - Simple iterative method
  - User provides “K”, which represents the number of clusters into which the data is to be grouped.
- Weaknesses
  - Often too simple → bad results
  - Difficult to guess the correct “K”

K-means Clustering Steps

Basic Algorithm:

- Step 0: select K
- Step 1: randomly select initial cluster seeds



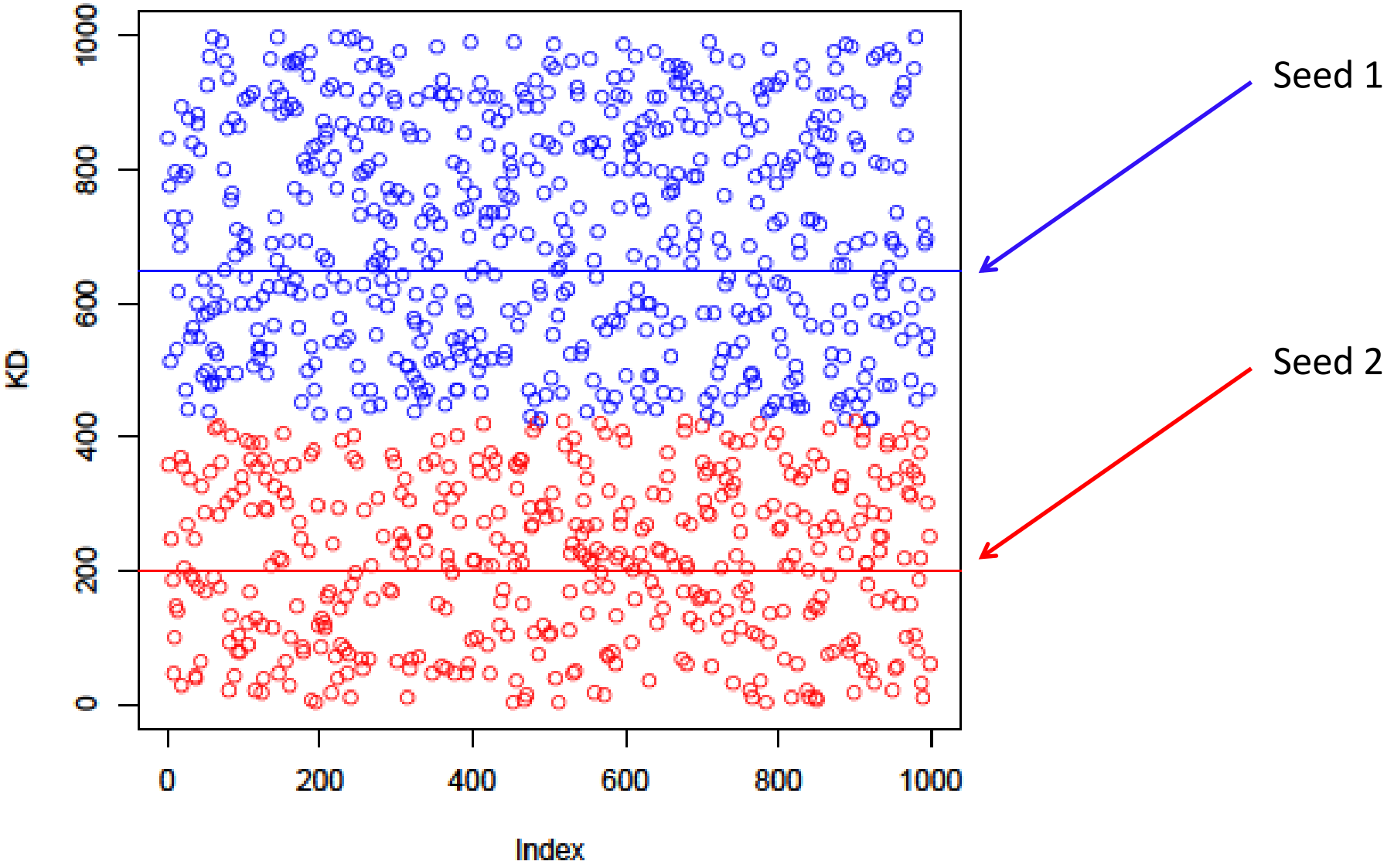
K-means Clustering Steps

- An initial cluster seed represents the “mean value” of its cluster.
- In the preceding figure:
  - Cluster seed 1 = 650
  - Cluster seed 2 = 200

K-means Clustering Steps

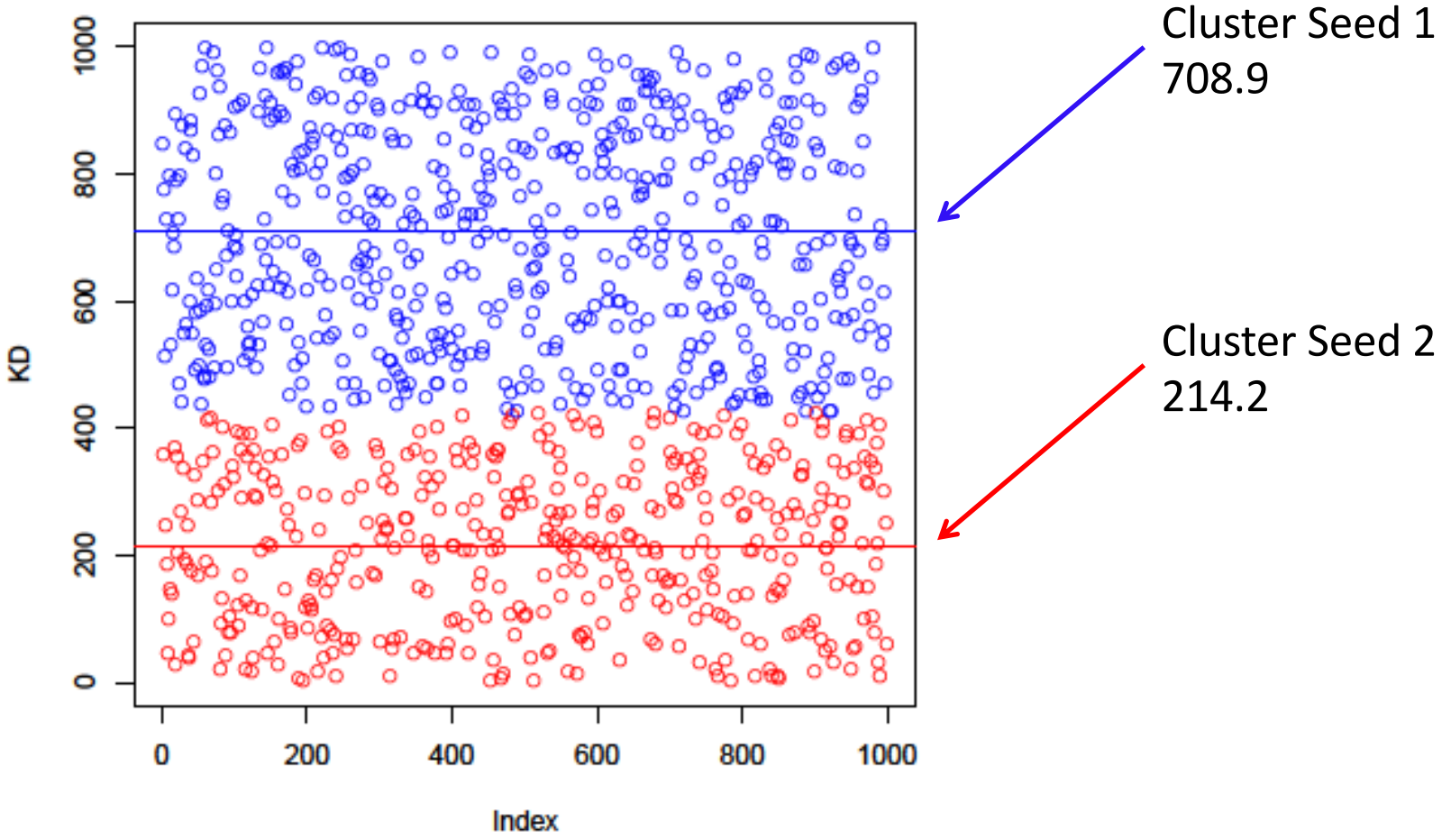
- Step 2: calculate distance from each object to each cluster seed.
  - What type of distance should we use?
    - Squared Euclidean distance
- Step 3: Assign each object to the closest cluster

K-means Clustering example



K-means Clustering Steps

- Step 4: Compute the new centroid for each cluster



K-means Calculation

- **Iterate**:
  - Calculate distance from objects to cluster centroids.
  - Assign objects to closest cluster
  - Recalculate new centroids
- **Stop based on convergence criteria**
  - No change in clusters
  - Max iterations

K-means Issues

- Distance measure is squared Euclidean
  - Scale should be similar in all dimensions
    - Rescale data?
  - Not good for nominal data. Why?
- Approach tries to minimize the within-cluster sum of squares error (WCSS)
  - Implicit assumption that SSE is similar for each group

WCSS

- The over all WCSS is given by: $\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$
- The goal is to find the smallest WCSS
- Does this depend on the initial seed values?
  - Possibly.