L03: K-Nearest Neighbors & K-Means

Classification and Clustering with Distance

Methods and Algorithms – MSc Data Science

Learning Objectives

By the end of this lecture, you will be able to:

1. Apply KNN for classification with appropriate K selection
2. Implement K-Means clustering and evaluate cluster quality
3. Compare distance metrics and their effects on results
4. Distinguish between supervised (KNN) and unsupervised (K-Means)

Finance Applications: Customer segmentation, fraud detection

From parametric models (regression) to instance-based methods

The Business Problem

Two Distinct Problems

1. Classification (Supervised)

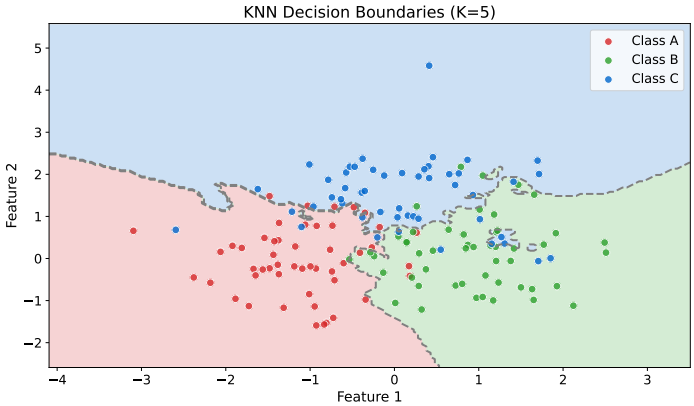
- Given labeled examples: is this transaction fraudulent?
- "Show me similar past transactions and their outcomes"

2. Clustering (Unsupervised)

- No labels: what natural customer segments exist?
- "Group customers by behavior for targeted marketing"

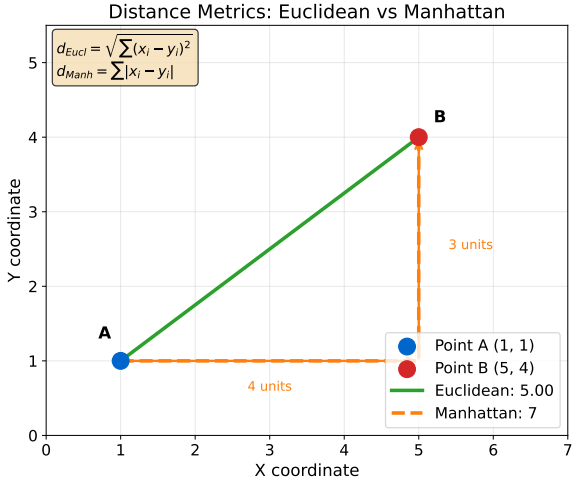
KNN = classification with labels, K-Means = clustering without labels

KNN: Decision Boundaries



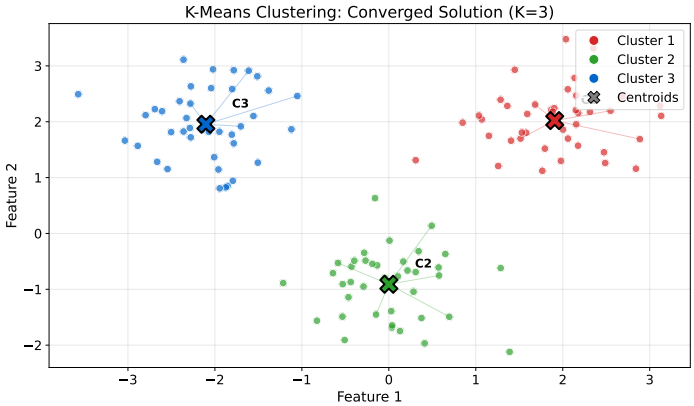
KNN creates non-linear, flexible decision boundaries based on local data

Distance Metrics



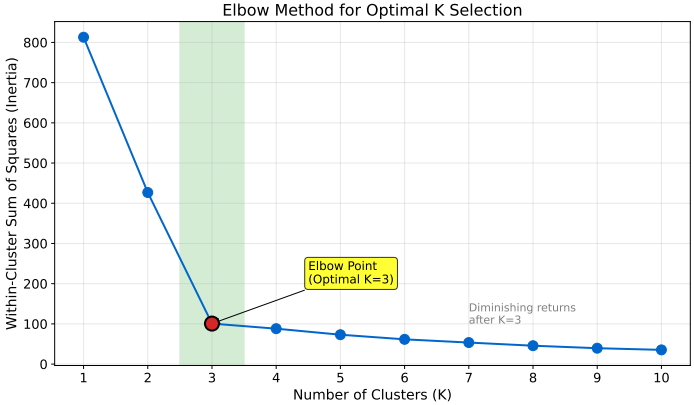
Choice of metric affects which points are considered "nearest"

K-Means: The Algorithm



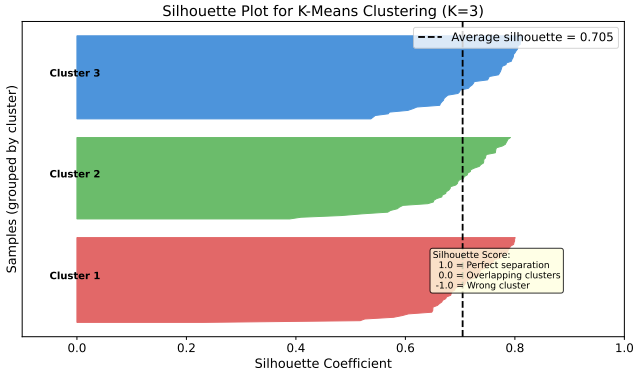
Iteratively assign points and update centroids until convergence

Choosing K: Elbow Method



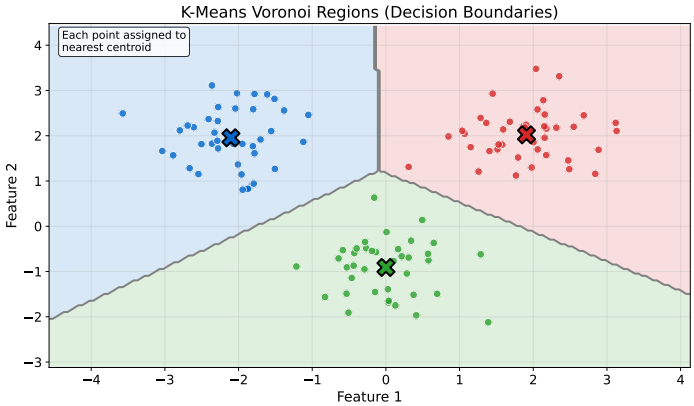
Look for the "elbow" where adding clusters gives diminishing returns

Cluster Quality: Silhouette Analysis



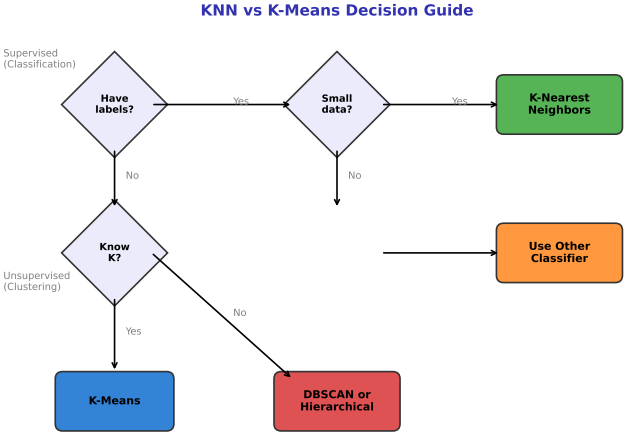
Silhouette score measures how similar points are to their own cluster

K-Means Decision Regions



Each region contains all points closest to one centroid

Decision Framework



KNN for labeled data classification, K-Means for unlabeled clustering