

L03: K-Nearest Neighbors & K-Means

Classification and Clustering with Distance

Methods and Algorithms

Spring 2026

Outline

1 Problem

2 Method

3 Solution

4 Practice

5 Decision Framework

6 Summary

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

Two Distinct Problems

1. Classification (Supervised)

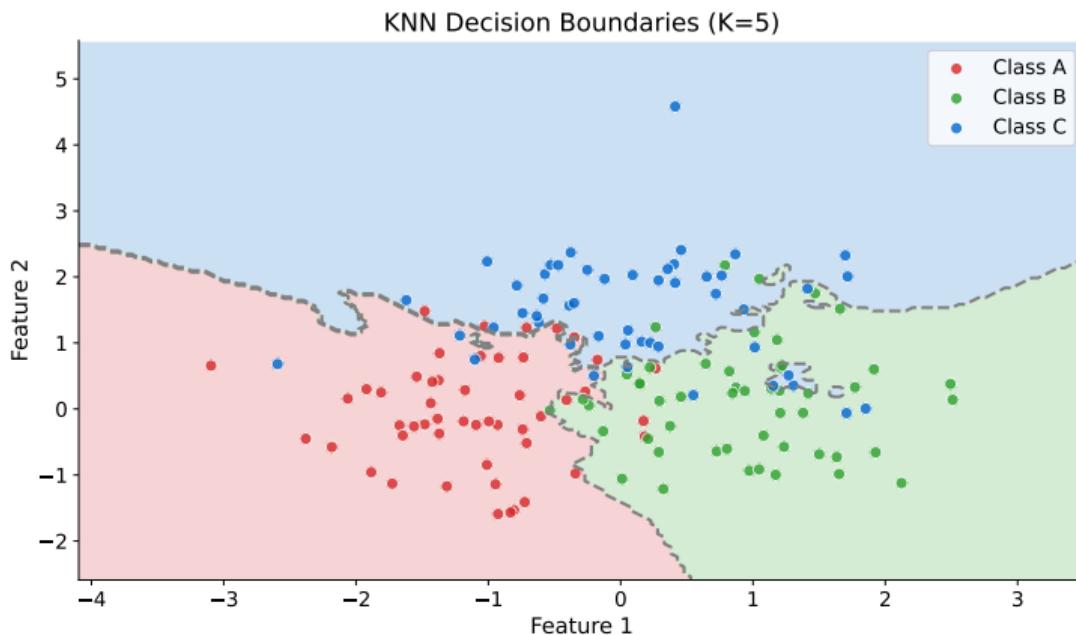
- Given labeled examples: is this transaction fraudulent?
- “Show me similar past transactions and their outcomes”

2. Clustering (Unsupervised)

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

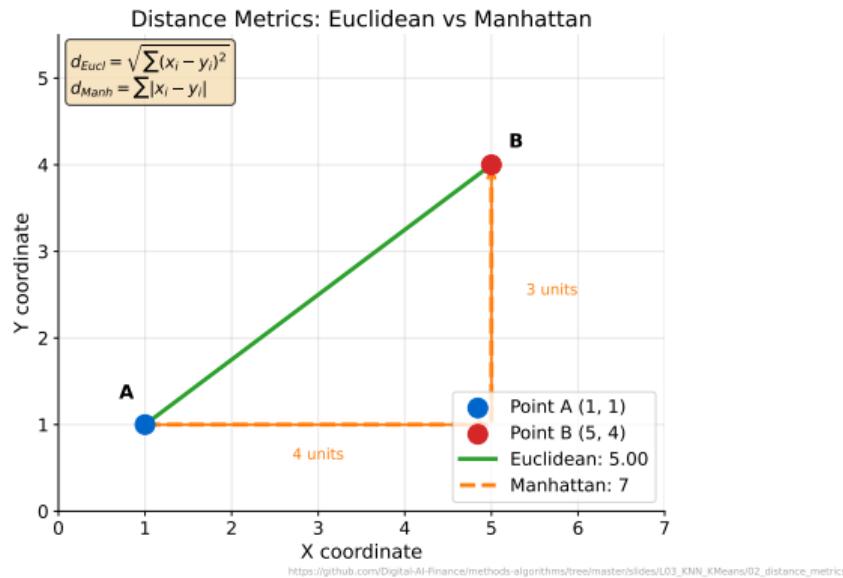
KNN = classification with labels, K-Means = clustering without labels

KNN: Decision Boundaries



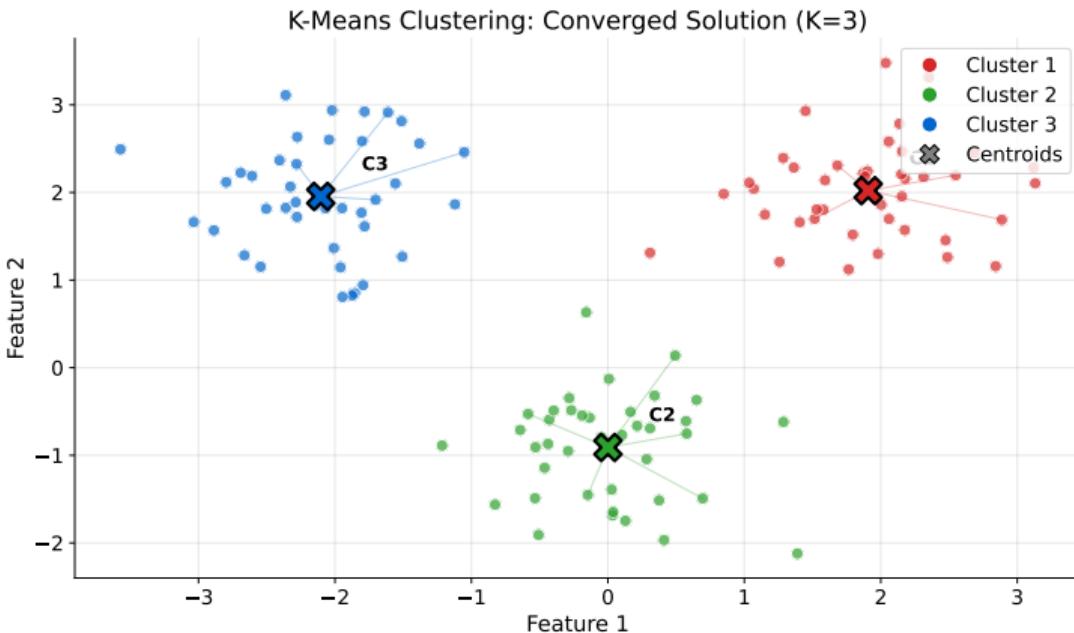
KNN creates non-linear, flexible decision boundaries based on local data

Distance Metrics



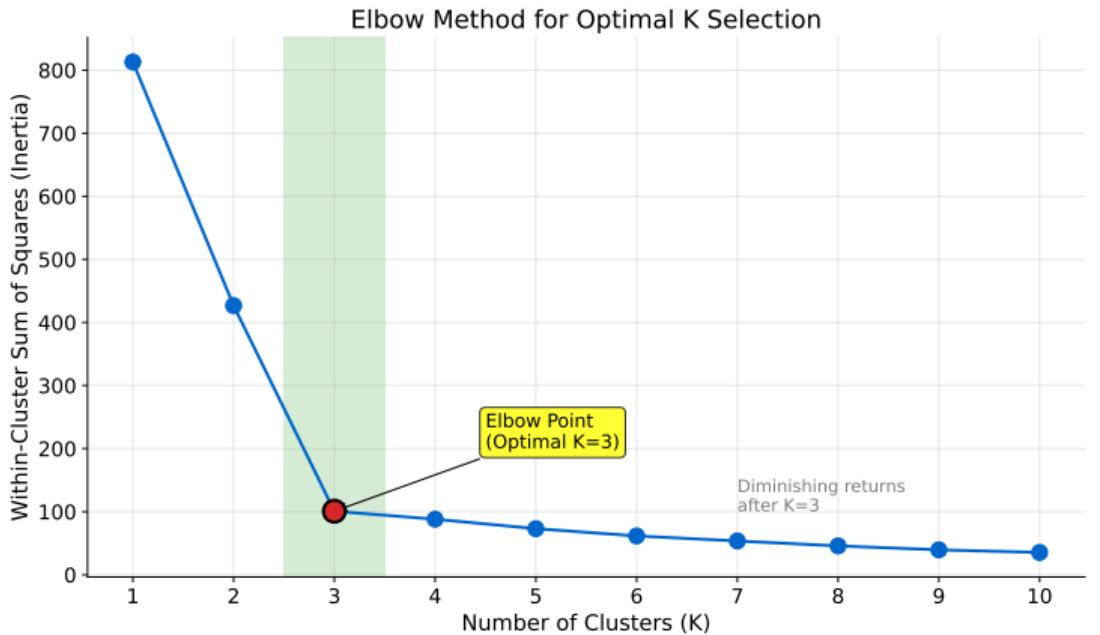
Choice of metric affects which points are considered “nearest”

K-Means: The Algorithm



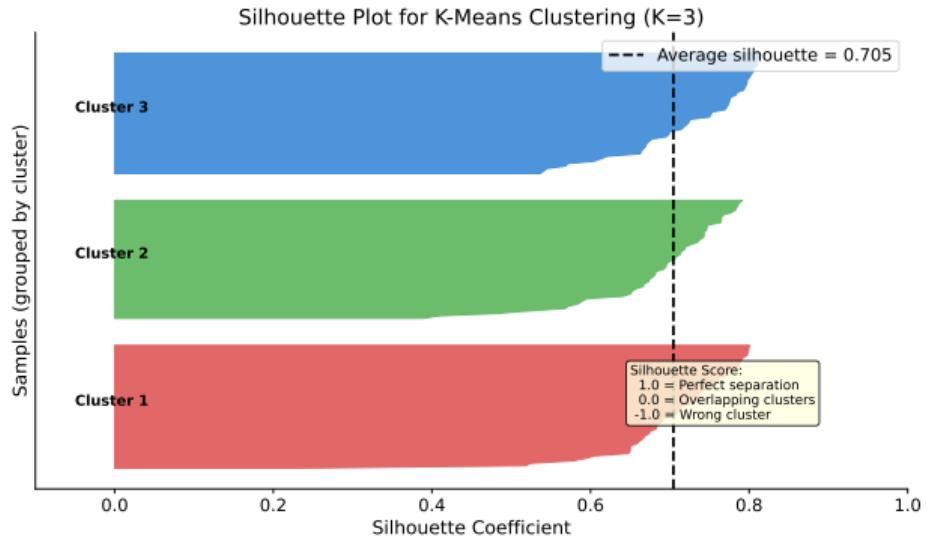
Iteratively assign points and update centroids until convergence

Choosing K: Elbow Method



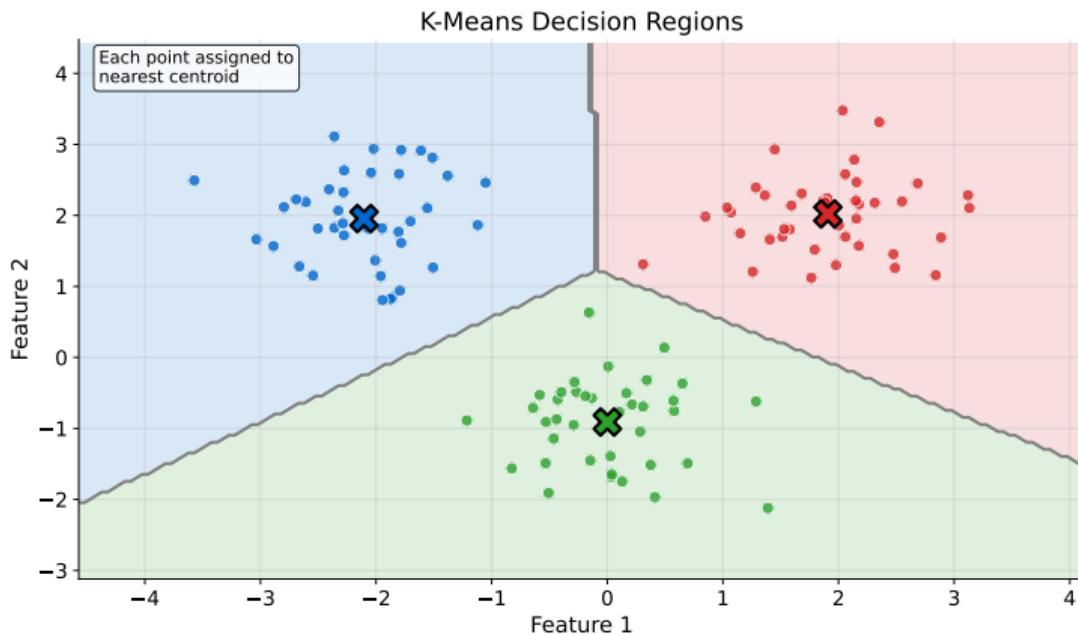
Look for the “elbow” where adding clusters gives diminishing returns

Cluster Quality: Silhouette Analysis



Silhouette score measures how similar points are to their own cluster

K-Means Decision Regions



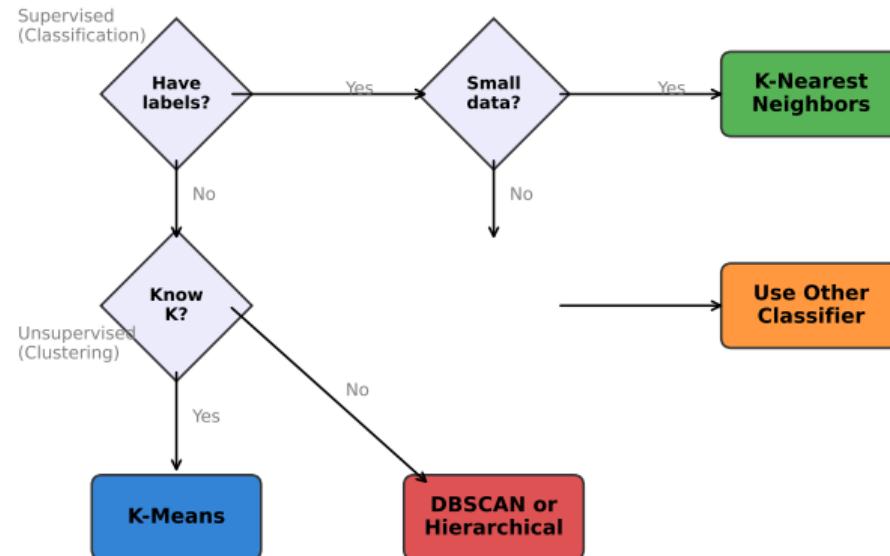
Each region contains all points closest to one centroid

Open the Colab Notebook

- Exercise 1: Implement KNN classifier from scratch
- Exercise 2: Apply K-Means to customer segmentation data
- Exercise 3: Compare distance metrics and k values

Link: <https://colab.research.google.com/> [TBD]

KNN vs K-Means Decision Guide



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/l03_KNN_KMeans/07_decision_flowchart

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

Remember

- KNN: supervised classification using nearest neighbors
- K-Means: unsupervised clustering with iterative centroids
- Distance metrics and K selection are critical choices
- Finance use cases: fraud detection, customer segmentation

Next lecture: L04 Random Forests