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

Learning Objectives

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

1. Analyze the bias-variance tradeoff in KNN as a function of K and derive the consistency bound
2. Prove K-Means convergence and evaluate initialization strategies (K-Means++)
3. Evaluate cluster validity using silhouette analysis, Hopkins statistic, and Gap statistic
4. Compare distance metrics and analyze their impact on algorithm performance in high dimensions

Finance Applications: Customer segmentation, fraud detection

Bloom's Level 4–5: Analyze, Evaluate, Prove, Compare

The Business Problem

Two Distinct Problems

1. Classification (Supervised)

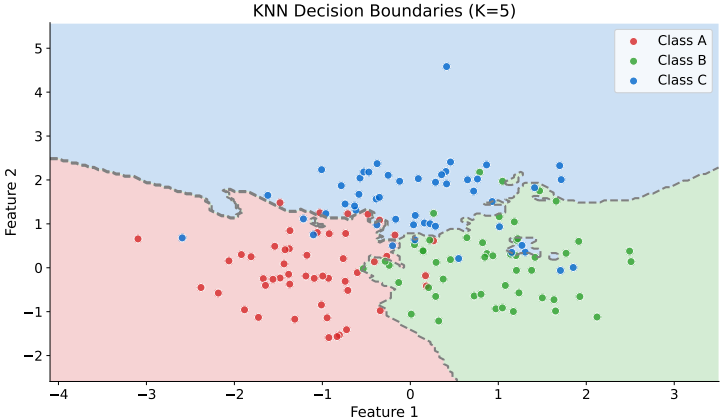
- Given labeled examples: is this transaction fraudulent?
- "Show me similar past transactions and their outcomes"

2. Clustering (Unsupervised)

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

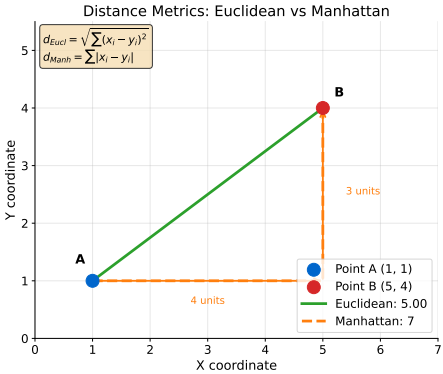
KNN = classification with labels, K-Means = clustering without labels

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"

Key Equations

Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2} \tag{1}$$

KNN classification (majority vote among k nearest neighbors):

$$\hat{y} = \text{majority vote among } k \text{ nearest neighbors} \tag{2}$$

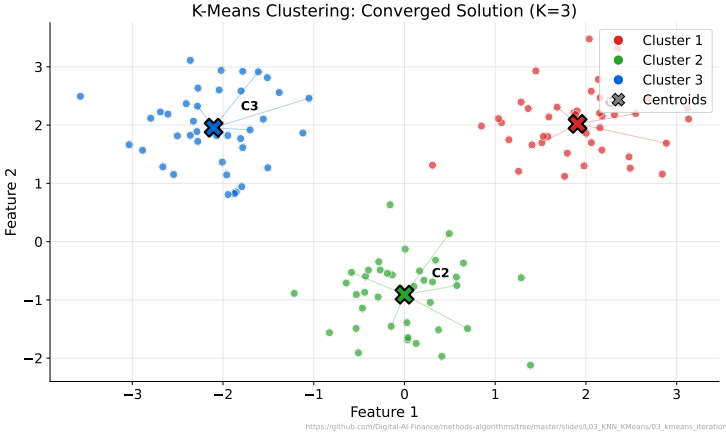
K-Means objective:

$$\min_{\mu_1, \ldots, \mu_K} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \tag{3}$$

Silhouette: $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]$

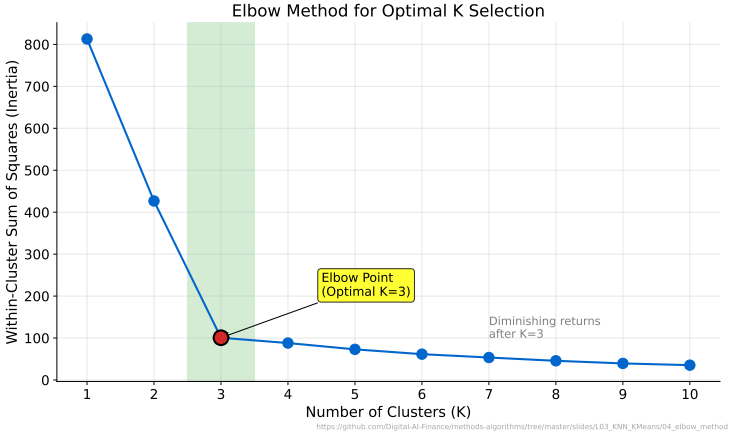
$a(i)$ = **mean intra-cluster distance;** $b(i)$ = **mean nearest-cluster distance**

K-Means: The Algorithm



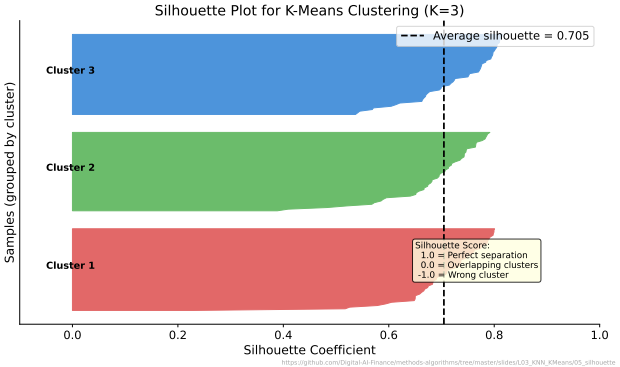
Iteratively assign points and update centroids until convergence

Choosing K: Elbow Method



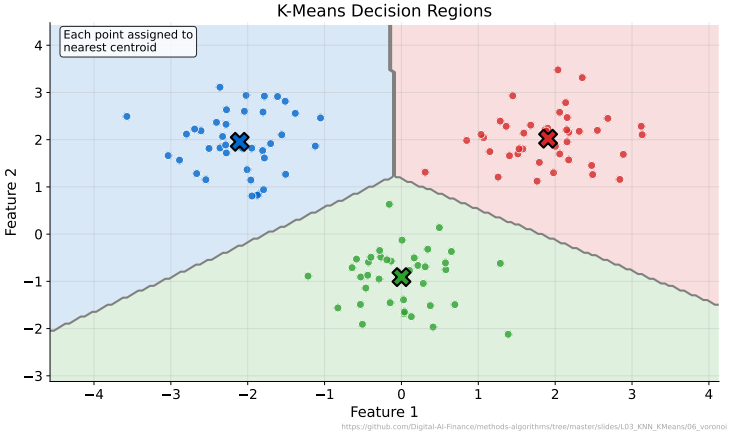
Look for the "elbow" where adding clusters gives diminishing returns

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

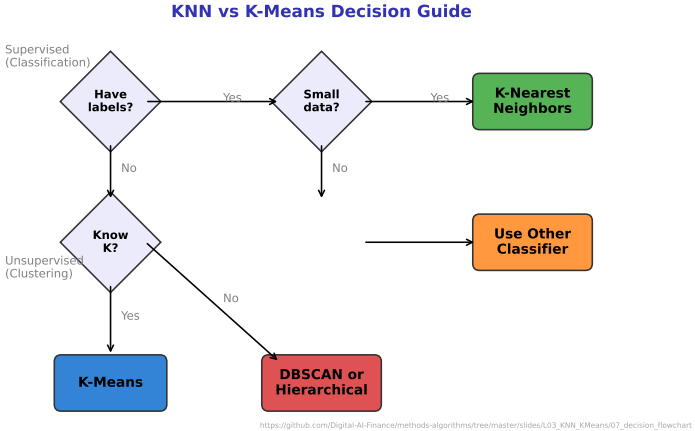
Hands-on Exercise

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: See course materials for Colab notebook

Decision Framework



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

Key Takeaways

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