Classification 2

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

Match 6th, 2025

Overview

K-Nearest Neighbors (KNN)

A Non-Parametric Classification (Linear) Method

- KNN classifies a test observation based on the majority class among its K closest neighbors.
- No explicit parametric assumptions about the data. Data does not have to be linearly related.
- Can be used for both classification and regression tasks. Both continuous (numeric) or categorical outcomes are permitted. We will focus on classification.

KNN Algorithm

Steps in KNN:

1. Choose the number of neighbors K.
2. Compute the distance between the test observation and all training observations.
3. Identify the K nearest neighbors.
4. Assign the most common class label among the neighbors (majority vote).

Example: KNN Classification

Illustrative Example from ISL

- Consider a dataset with two classes: blue and orange.
- A new test observation (black cross) is classified using KNN with $K = 3$.
- The three closest points (circled) are identified.
- Since two out of three neighbors are blue, the test observation is classified as blue.

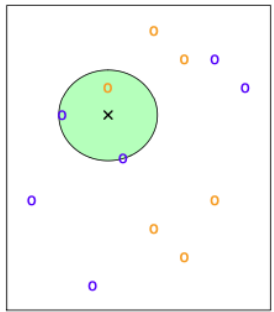


Figure: KNN classification example from ISL.

KNN Decision Boundary

- The decision boundary separates the feature space into regions assigned to each class.
- When K is small, the boundary is highly flexible and follows training data closely.
- When K is large, the boundary becomes smoother and generalizes better.

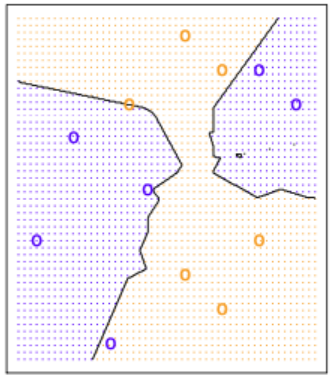


Figure: KNN decision boundary for $K = 3$.

Choosing K in KNN

- **Small K:** High variance, low bias (overfits to noise in the data).
- **Large K:** Low variance, high bias (oversmooths the decision boundary).
- The optimal K is typically chosen using cross-validation.

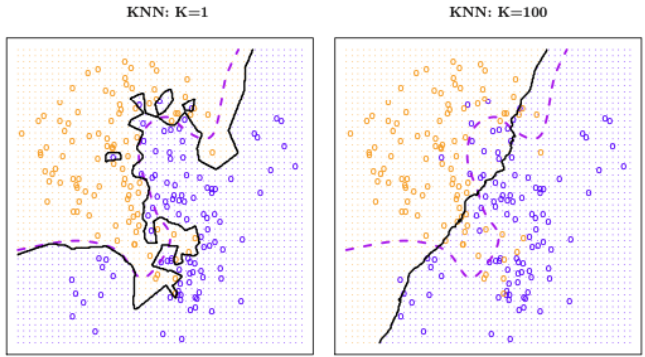


Figure: Effect of different values of K on decision boundary.

KNN Error Rate Analysis

Understanding the Impact of K on Error Rate

- The classification error rate measures how often KNN misclassifies observations.
- Error rate varies with K:
  - Small K: High variance, low bias, more sensitive to noise.
  - Large K: Low variance, high bias, smoother boundaries.
- The optimal K minimizes the test error rate.

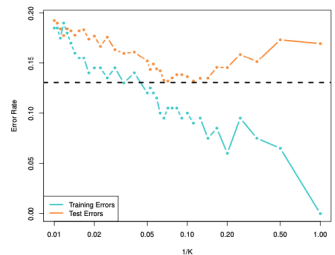


Figure: KNN error rate for different values of K.

Strengths and Weaknesses of KNN

Strengths:

- Simple and intuitive.
- Works well with small datasets.

Weaknesses:

- Computationally expensive for large datasets.
- Performance depends on the choice of distance metric.
- Sensitive to irrelevant or redundant features.

K-Means Clustering

A Partitioning Clustering Method

- An unsupervised machine learning method.
- K-means clustering partitions observations into K distinct clusters.
- Each cluster is defined by its **centroid** (mean of points in that cluster).
- The goal is to minimize within-cluster variation - i.e. to make the observations within each cluster be as similar as possible.

K-Means Algorithm

Steps in K-Means Clustering:

1. Choose the number of clusters K.
2. Randomly assign each observation to one of the K clusters.
3. Iterate until convergence:
   - Compute the centroid of each cluster.
   - Assign each observation to the nearest centroid.

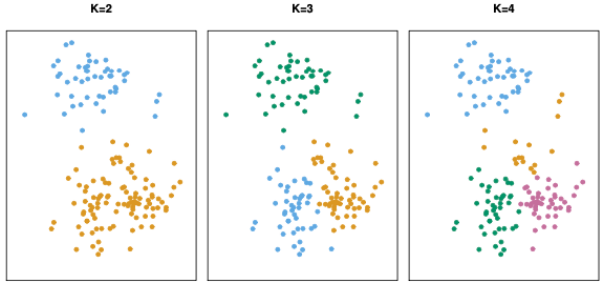
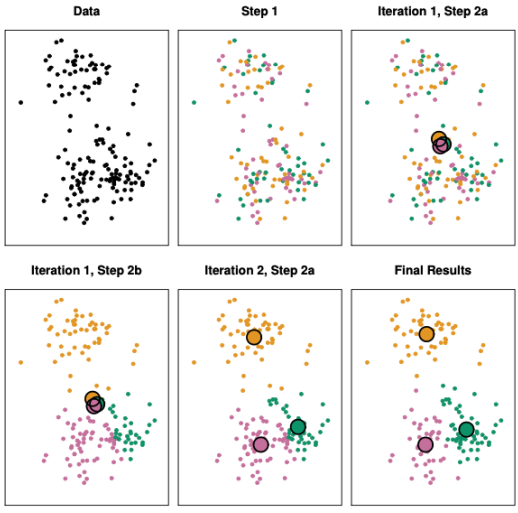


Figure: Illustration of K-Means clustering process.

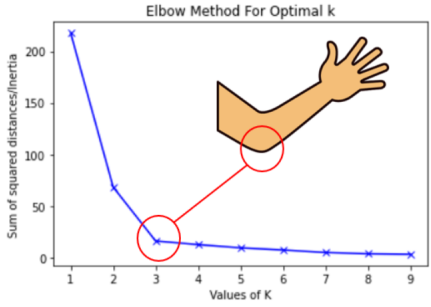
K-Means Algorithm



Evaluating K-Means Clustering

Metrics for Clustering Performance:

- Within-cluster variation: Measures how tight clusters are.
- Total within-cluster sum of squares (WCSS): Sum of squared distances from each point to its cluster centroid.
- Between-cluster variation: Measures how well-separated clusters are.
- Elbow Method: A common technique to select the optimal K.



Line plot between K and inertia

Applications of K-Means Clustering

Real-World Uses of K-Means Clustering:

- Healthcare:
  - Patient segmentation for personalized treatments.
  - Disease subtype identification based on genetic markers.
- Marketing & Customer Segmentation:
  - Identifying customer groups for targeted advertising.
  - Analyzing purchasing behavior.
- Social Network Analysis:
  - Community detection in large networks.
  - Recommender systems for user-group clustering.