Week 11-1: K-Nearest Neighbors (k-NN)

Ekarat Rattagan

April 9, 2025

Supervised Learning: Overview

- **Regression**:
  - Linear Regression (LR), Model representation: $\theta^T x$
  - Cost Function: Mean Squared Error (MSE)
  - Evaluation Metric: MSE
- **Classification**:
  - Logistic Regression, Model representation: $\sigma(\theta^T x)$
  - Cost: Negative Log-Likelihood Loss (NLLL)
  - Evaluation Metric: Confusion Matrix

k-NN

Assumption: Similar inputs should have similar outputs.

- Model representation: ?
- Cost function: ?
- Evaluation Metric: Confusion Matrix (for classification), MSE (for regression)

Properties:

- Non-parametric models
- Lazy learning / instance-based method
- Works for both classification and regression tasks

k-NN Algorithm

1. Find the Euclidean distance between the testing data and each training data point.
2. Sort the distances from minimum to maximum
3. Pick k nearest neighbors
4. Use majority vote for classification

Euclidean Distance (L_2):

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{d} (x_{i,j} - x_{k,j})^2}$$

k-NN for Regression and Classification

Sample Data:

Sample	x_1	x_2	y_1	y_2
S1	1	5	1000	yes
S2	2	6	1200	yes
S3	3	1	1100	no
S4	2	4	2000	yes
S5	**2**	**5**	?	?

Predict: To estimate the value for S5 using the k-nearest neighbors (k-NN) algorithm, compute the average of the y_1 values for regression tasks and apply a majority vote to the y_2 values for classification tasks.

k-NN with Categorical Data

Example Dataset

Sample	Food	Chat	Fast	Price	Bar	Tip
S1	great	yes	yes	normal	no	yes
S2	great	no	yes	normal	no	yes
S3	mediocre	yes	no	high	no	no
S4	great	yes	yes	normal	yes	yes
S5	great	no	no	normal	no	?

Distance: Hamming Distance Match $= 0$, Mismatch $= 1$

Define $k = 2$, then compute:

- $H(S5, S1)$
- $H(S5, S2)$
- $H(S5, S3)$
- $H(S5, S4)$
- Choose two nearest neighbors with lowest Hamming distances

Definition: Hamming Distance

Definition

The Hamming distance between two equal-length strings of symbols is the number of positions at which the corresponding symbols are different.

Examples

The symbols may be letters, bits, or decimal digits, among other possibilities. For example, the Hamming distance between:

- "**karolin**" and "**kathrin**" is 3.
- "**karolin**" and "**kerstin**" is 3.
- "**kathrin**" and "**kerstin**" is 4.
- **0000** and **1111** is 4.
- **2173896** and **2233796** is 3.

Ref: https://en.wikipedia.org/wiki/Hamming_distance

Tuning the Hyperparameter k

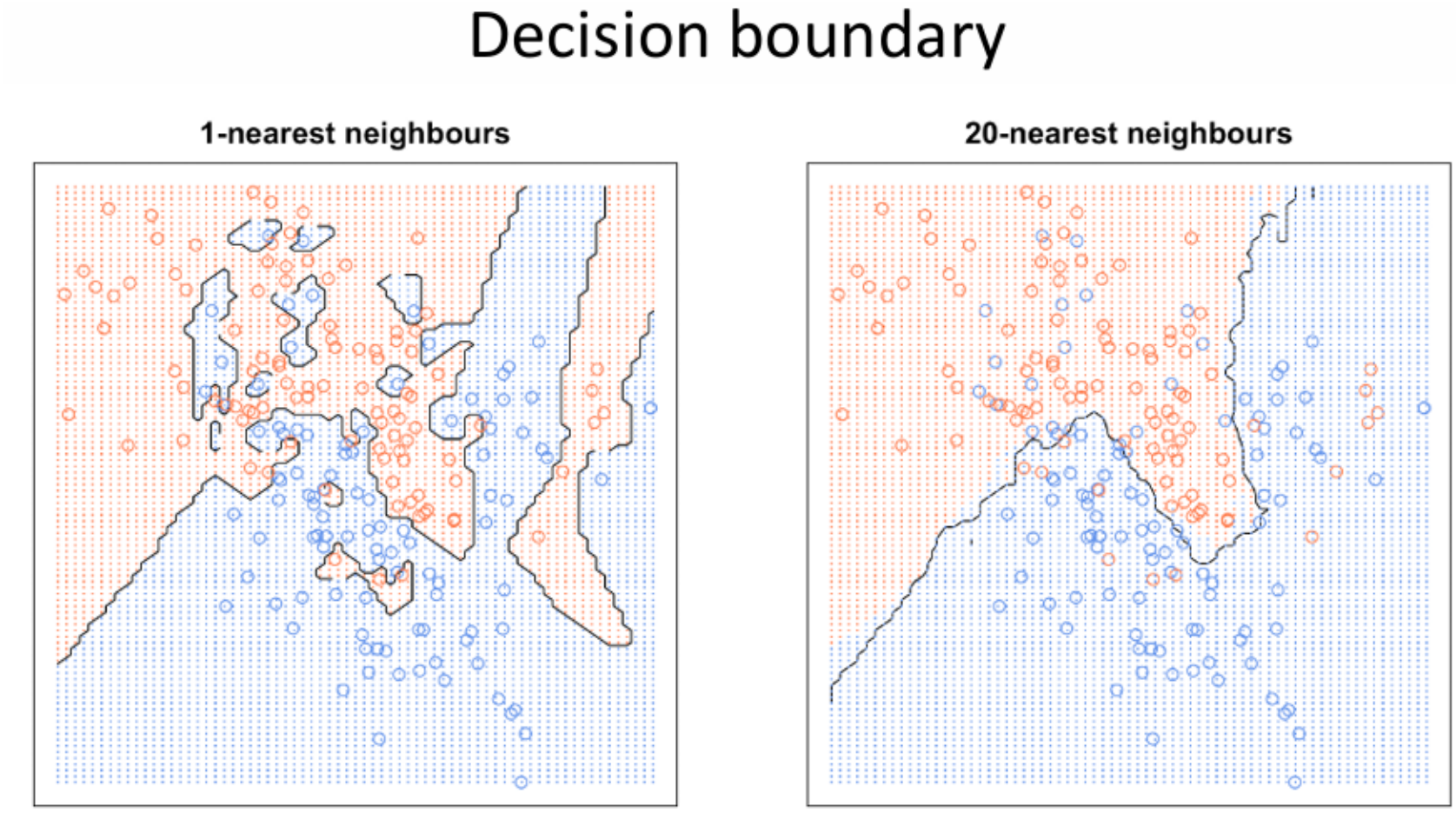
How to choose k?

- Use odd k values to avoid ties: $k = 1, 3, 5, ..., \sqrt{N}$
- Perform cross-validation using `GridSearchCV`

Effect of k:

- k too small: overfitting
- k too large: underfitting

Decision Boundaries



Ref: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/

Issue 1: Feature Scaling

Problem:

- Features with large magnitude dominate distance calculations.
- Example: Population $= 120{,}000{,}000$ VS Age $= 60$

Solution: Feature Scaling

- **z-score normalization**: Maps to range $(-3, 3)$
- **Min-max scaling**: Maps to range $[0, 1]$

Other Issues in k-NN

Issue 2: Noisy Data

- Solution: Remove outlier data
- Methods: box plots, histograms, Z-score or IQR

Issue 3: Slow Testing Time for Big Data

- Solution: Use KD-Trees

Issue 4: Storage Requirements

- Solution: Compression (e.g., ZIP)

Issue 5: Curse of Dimensionality

- Solution: ?

Application

- Im2gps: https://graphics.cs.cmu.edu/projects/im2gps/im2gps.pdf
- Data imputation: https://www.sciencedirect.com/science/article/pii/S0164121212001586?casa_token=EQPFMncizwsAAAAA:wF2cst5NhtlUAUHmVEqkg2c8JskyGTFywEUtfUaqSRp6Qi7vux04wG3
- Recommendation system: https://www.sciencedirect.com/science/article/pii/S221083271400026X