Module 8: Nearest Neighbour Methods

Learning outcomes

1. Analyse the real-life applications and limitations of KNN methods.
2. Calculate common distance functions in ML algorithms.
3. Apply preprocessing methods, including normalisation and encoding, with KNN to predict plant types in Python.
4. Identify the optimal value of k in relation to decision boundaries, validation results and bias–variance trade-offs.
5. Apply KNN methods to classification and regression problems.

k-nearest neighbours (KNN) overview

- KNN is a non-parametric, distance-based ML method that assumes similar inputs lead to similar outputs.
- It is used for both classification and regression tasks.
- It requires selecting a parameter k, representing the number of nearest neighbours considered.

Data preparation for KNN

1. Distance functions
   - Proximity between data points is defined using the following common distance metrics:
     - Euclidean distance
       - It measures the straight-line distance between two points in n-dimensional space.
       - It is best used for continuous numerical variables.
       - Formula: $\mathrm{d(x,y)} = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i} - \mathrm{y_i})^2}$
     - Manhattan distance
       - It measures distance as the sum of the absolute differences of their coordinates.
       - It is used in certain spatial contexts.
       - Formula: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$
     - Minkowski distance
       - It is a generalisation of Euclidean and Manhattan distances with a parameter p.
       - When $p = 1$, it equals Manhattan distance, and when $p = 2$, it equals Euclidean distance. For other values of p, it produces fractional or generalised distances.
       - Formula: $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$

    - Hamming distance
        - It measures the number of positions at which two binary strings of equal length differ.
        - It is used primarily for categorical or binary data, error detection and coding theory.
    - Cosine *similarity* distance
        - It measures the cosine of the angle between two vectors and effectively measures orientation rather than magnitude.
        - It is used especially in text analysis and high-dimensional positive spaces (e.g. term frequency–inverse document frequency vectors).
- The choice of metric affects model performance and depends on the data context.

2. Feature scaling
    - Scaling ensures each feature contributes equally to distance calculations.
    - Two common techniques:
        - *Min–max* normalisation, which scales values to a $[0, 1]$ range
        - *z-score* normalisation, which centres data around 0 with unit variance and is better for handling outliers
3. Categorical and binary predictors
    - Binary predictors (e.g. yes/no) are encoded as 0 and 1.
    - Categorical predictors (e.g. colour) require:
        - One-hot encoding (most common)
        - Ordinal encoding if order exists (e.g. low = 1, medium = 2 and high = 3)
    - Alternatively, subsetting by category can be used in certain modelling contexts.

Model tuning and evaluation

- Small k:
    - Highly flexible
    - Captures local patterns
    - Susceptible to overfitting (high variance)
- Large k:
    - Smoother decision boundary
    - Less sensitive to noise
    - Risk of underfitting (high bias)
- Optimal k is data-dependent and should be selected via a training–validation split or cross-validation.

Bias–variance trade-off in the context of KNN

- Low k has low bias and high variance.
- High k has high bias and low variance.
- Balance is key to generalisation performance.

Curse of dimensionality

- In high-dimensional spaces:
  - Distance measures become less meaningful.
  - Most data lies near the boundaries or corners of the feature space.
  - Models require exponentially more data to maintain performance.
- Feature selection or dimensionality reduction is essential for KNN in such contexts.

Applications of KNN

- Classification
  - The label is determined by a majority vote among the *k*-nearest neighbours.
  - Examples include disease diagnosis, credit risk prediction and personalised recommendations (e.g. products and music).
- Regression
  - Prediction is based on the average value of *k*-nearest neighbours.
  - It uses metrics such as mean squared error for evaluation.