DEPARTMENT OF BIOMEDICAL ENGINEERING

24AI671-FUNDAMENTALS OF DATA SCIENCE

UNIT IV- DATA ANALYTICS AND MACHINE LEARNING FUNDAMENTALS

4. 4. CLASSIFICATION AND REGRESSION

Classification and regression are the two main problems of supervised learning. The difference lies in the type of output that is predicted: **classification predicts a category (discrete label)** and **regression predicts a continuous number**.

Classification

In classification, the model learns from labelled examples and then assigns a new input to one of the **predefined classes**. When there are only two classes (yes/no, spam/not spam) it is called **binary classification**; when there are more than two (cat/dog/horse) it is **multi-class classification**.

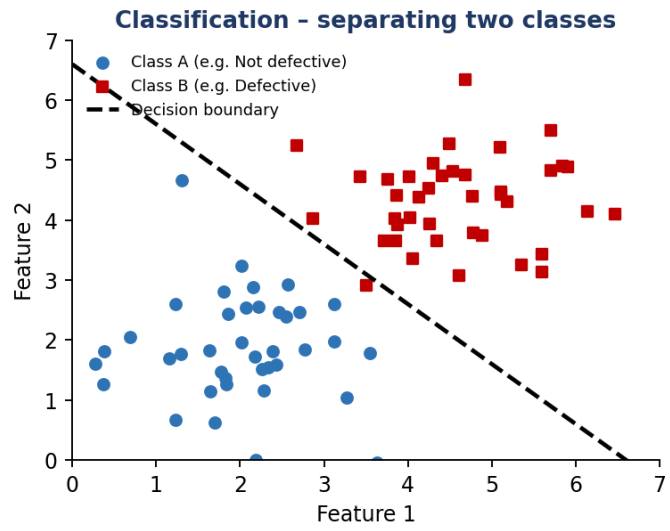


Fig 4.1 Classification – a decision boundary separates the classes

Common Classification Algorithms

- **Logistic Regression** – despite its name it is a classification method. It uses the **sigmoid function** to convert a value into a probability between 0 and 1; the class is decided by a threshold, usually 0.5.

$$\sigma(z) = 1 / (1 + e^{-z}), \quad \text{where } z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- **K-Nearest Neighbours (KNN)** – finds the K training points closest to the new point (using Euclidean distance) and assigns the majority class. It is simple but slow for big data.

$$\text{Euclidean distance } d = \sqrt{[\,(x_2 - x_1)^2 + (y_2 - y_1)^2\,]}$$

- **Decision Tree** – a tree-like model in which each internal node tests a feature, each branch is the result of the test and each leaf is a class. The best split is chosen using **Entropy / Information Gain** or the **Gini index**.

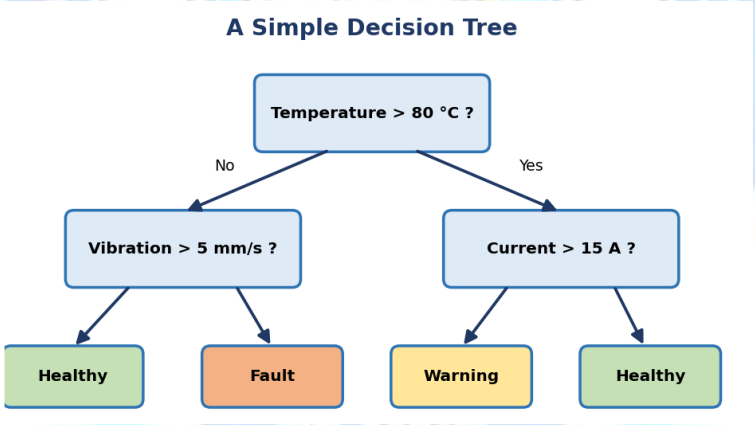


Fig 4.2 Example of a decision tree for machine health

$$\text{Entropy} = -\Sigma\, p_i \log_2 p_i \qquad \text{Gini} = 1 - \Sigma\, p_i^2$$

- **Random Forest** – an ensemble of many decision trees trained on random samples of data and features; the final class is by majority voting. It is more accurate and reduces overfitting.

- **Support Vector Machine (SVM)** – finds the **hyperplane with the maximum margin** between two classes. Kernel functions (linear, polynomial, RBF) allow it to separate data that is not linearly separable.

- **Naive Bayes** – based on Bayes' theorem; assumes that features are independent. Fast and very good for text classification.

$$P(A \mid B) = [\, P(B \mid A) \times P(A)\,] / P(B)$$

Regression

In regression, the output is a **continuous value**. The model learns the relationship between the input features (independent variables) and the output (dependent variable) and uses it to predict numeric values.

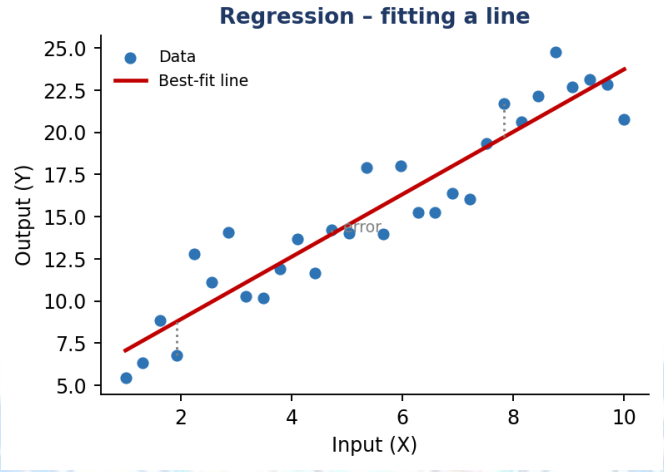


Fig 4.3 Regression – fitting the best line through the data

Common Regression Algorithms

- **Simple Linear Regression** – one input variable; the best-fit line is found by minimising the sum of squared errors (least squares method).

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- **Multiple Linear Regression** – many input variables: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$.
- **Polynomial Regression** – fits a curved line by adding powers of X (X^2, X^3); used when the relationship is not a straight line.
- **Ridge and Lasso Regression** – linear regression with a penalty term (L2 and L1 regularisation) to reduce overfitting. Lasso can also remove unimportant features.
- **Decision Tree / Random Forest / SVR / Neural network regression** – non-linear methods for complex data.

Classification vs Regression

Point	Classification	Regression

Point	Classification	Regression
Output	Discrete class label	Continuous numeric value
Example	Is the tumour benign or malignant?	What will be the house price?
Algorithms	Logistic regression, KNN, SVM, Naive Bayes	Linear, polynomial, Ridge, Lasso regression
Evaluation	Accuracy, precision, recall, F1-score, ROC-AUC	MAE, MSE, RMSE, R^2
Graph	Decision boundary separating classes	Best-fit line or curve through the data

5. CLUSTERING

Clustering is an unsupervised learning technique that **divides unlabelled data into groups (clusters) such that the data points in the same cluster are very similar to each other and different from those in other clusters.** Similarity is usually measured by the distance between points – the smaller the distance, the greater the similarity.

Types of Clustering

- **Partition-based** – the data is divided into K groups. Example: K-Means, K-Medoids.
- **Hierarchical** – builds a tree of clusters (dendrogram). Example: agglomerative and divisive clustering.
- **Density-based** – clusters are dense regions separated by low-density regions; can find clusters of any shape and detect noise. Example: DBSCAN.
- **Model-based** – assumes the data comes from a mixture of distributions. Example: Gaussian Mixture Model (GMM).

K-Means Clustering

K-Means is the most popular clustering algorithm. It divides the data into **K clusters**, each represented by its centre called the **centroid**.

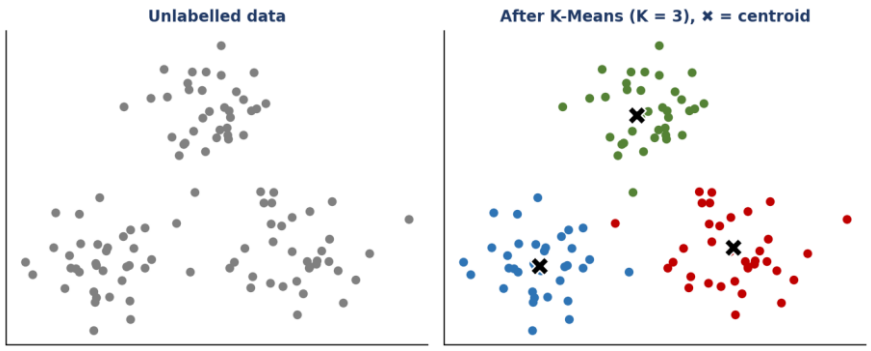


Fig 5.1 K-Means clustering (K = 3)

Algorithm

- **Step 1:** Choose the number of clusters K.
- **Step 2:** Select K points randomly as the initial centroids.
- **Step 3:** Assign every data point to the nearest centroid (using Euclidean distance).
- **Step 4:** Recalculate each centroid as the mean of all points assigned to it.
- **Step 5:** Repeat steps 3 and 4 until the centroids stop changing (convergence) or the maximum number of iterations is reached.

$$\text{Objective: minimise } \text{WCSS} = \Sigma_k \Sigma (x_i \in C_k) \; \| x_i - \mu_k \|^2$$

Choosing K – The Elbow Method

The algorithm is run for different values of K and the **WCSS (within-cluster sum of squares)** is plotted against K. WCSS falls as K increases; the point where the fall becomes slow, like an elbow, gives the best K. The **silhouette score** can also be used – a value near +1 means well-separated clusters.

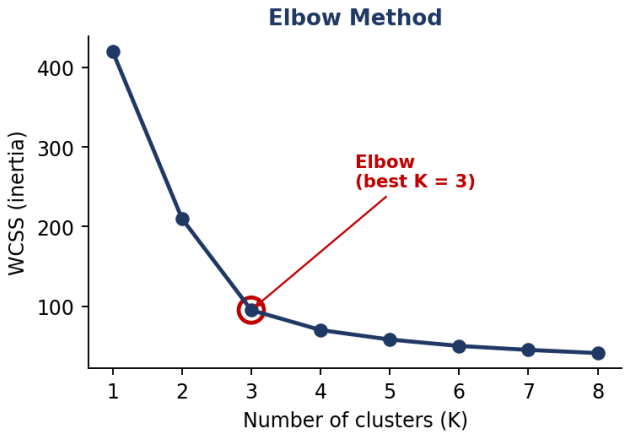


Fig 5.2 Elbow method

Merits and Demerits of K-Means

Merits	Demerits
Simple, fast and easy to implement	K must be decided in advance
Works well for large data sets	Sensitive to initial centroids and outliers
Gives compact, spherical clusters	Cannot find clusters of irregular shape

Hierarchical Clustering

Hierarchical clustering creates a **tree of clusters** called a **dendrogram**. In the **agglomerative (bottom-up)** method, every point starts as its own cluster and the two closest clusters are merged step by step until only one cluster remains. In the **divisive (top-down)** method, all points start in one cluster which is split repeatedly. Cutting the dendrogram at a chosen height gives the required number of clusters. No need to fix K in advance.

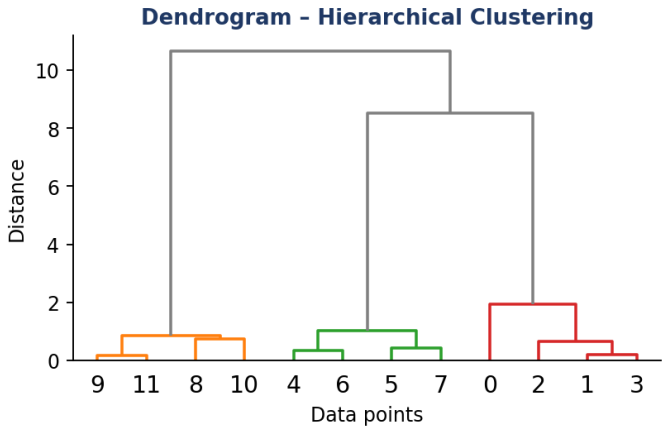


Fig 5.3 Dendrogram

- **Linkage methods** decide the distance between clusters: single (minimum distance), complete (maximum distance), average and Ward's method (minimises increase in variance).

DBSCAN (Density-Based Clustering)

DBSCAN groups points that are closely packed together and marks points in low-density regions as **noise (outliers)**. It needs two parameters: **ε (epsilon)** – the radius of the neighbourhood, and **MinPts** – the minimum number of points required to form a dense region. It does not need K and can find clusters of any shape.

Applications of Clustering

- Customer segmentation for marketing.
- Image segmentation and compression.
- Grouping of similar documents, news articles and genes.
- Anomaly and fraud detection.
- Grouping similar machines or sensor patterns in smart factories.