

Pattern Recognition

Problem Formulation

ETC3123 Artificial Intelligence for Engineering Applications

Dr Richard Wong

Problem Formulation for Pattern Recognition

Problem Types in Pattern Recognition

Data Representation

Feature Extraction

Model Selection

Performance Metrics

Examples

Challenges and Pitfalls

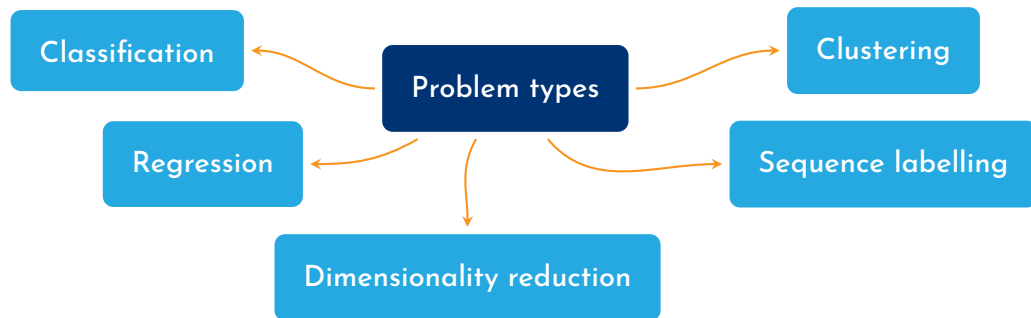
Pattern recognition focuses on the identification and classification of data patterns.



Problem Formulation for Pattern Recognition

- ▶ Problem types
- ▶ Data Representation
- ▶ Feature Extraction
- ▶ Model Selection
- ▶ Performance Metrics

Problem Types in Pattern Recognition

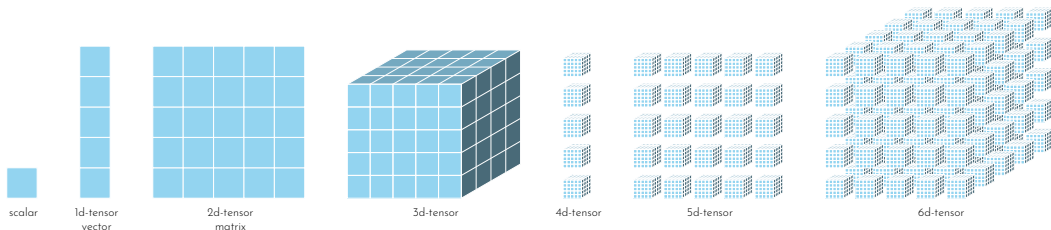


- ▶ **Classification** To assign an input to one of a finite number of categories or classes
- ▶ **Regression** To predict a continuous output based on the input
- ▶ **Clustering** To group data points into clusters based on similarity without predefined labels
- ▶ **Dimensionality Reduction** To reduce the number of input features while retaining essential information
- ▶ **Sequence Labelling** To assign labels to sequences

- ▶ **Numerical Data** Represented in the form of numbers (e.g., sensor readings).
- ▶ **Categorical Data** Represented by discrete categories or classes (e.g., species of plants).
- ▶ **Textual Data** Data in text form (e.g., emails, documents).
- ▶ **Image Data** Represented as pixels (e.g., grayscale or RGB values).

Input Representation

- **Vectors** Most pattern recognition problems use vectorized data, where each data point is represented as a vector of features.
- **Matrices/Tensors** Used for more complex data like images, where data is stored in higher dimensions.



Normalisation and Scaling

Data often needs to be normalized or scaled to ensure uniformity and avoid biasing the model due to large feature values.

- normalisation (min-max scaling)

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- standardisation (z-score scaling)

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

- robust scaling

$$X_{\text{scaled}} = \frac{X - \text{median}}{\text{IQR}}$$

- max-abs scaling

$$X_{\text{scaled}} = \frac{X}{|X_{\max}|}$$

- logarithmic scaling

$$X_{\text{scaled}} = \log(X + \epsilon)$$

Data Augmentation

Techniques used to artificially expand the dataset by generating modified versions of the existing data

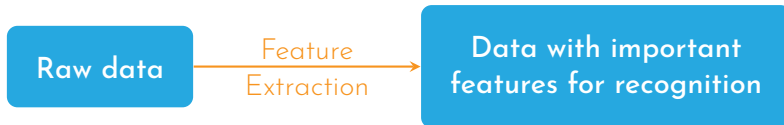
Image

- ▶ geometric transformation
(rotation, flipping, scaling, translation, shearing)
- ▶ color space transformations
(grayscale, brightness adjustment, contrast adjustment)
- ▶ noise injection
(Gaussian noise, Speckle noise)

Text

- ▶ synonym replacement
- ▶ random insertion, deletion, swap
- ▶ noise injection
(typos, spelling errors)

Features are the individual measurable properties of the data that help in distinguishing between patterns.



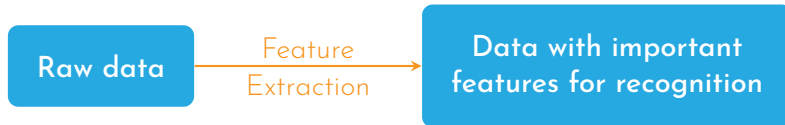
- ▶ **Domain-specific Features** Features derived from domain knowledge. For example, in speech recognition, features could include pitch or frequency.
- ▶ **Learned Features** In deep learning, features are automatically learned by the algorithm, particularly through layers of neural networks.

- ▶ **Manual Feature Selection** Based on domain expertise, select features that are most likely to improve model performance.
- ▶ **Automated Feature Selection** Use statistical methods like PCA (Principal Component Analysis) or correlation analysis to reduce dimensionality.

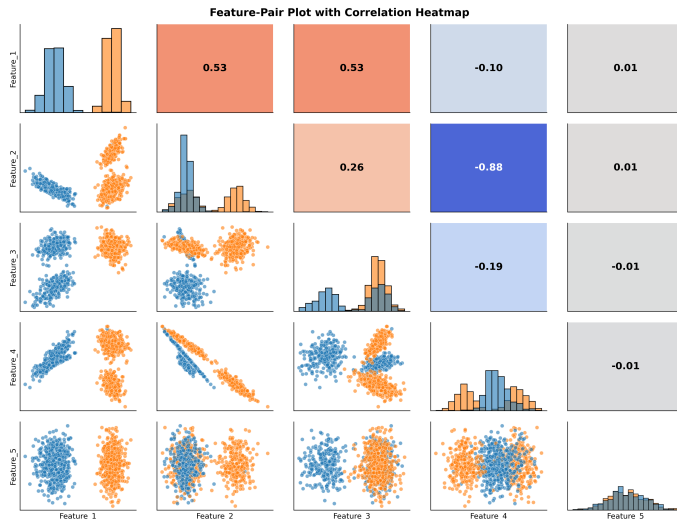
- ▶ The process of creating new features from raw data to improve model performance.
- ▶ Example: Creating interaction terms between features or extracting time-based features from time-series data.

Feature Extraction

Summary



Which are the (un)important features?



Correlation with Target

Feature 1	0.965738
Feature 2	0.555445
Feature 3	0.513271
Feature 4	-0.139160
Feature 5	0.006404

- ▶ **Linear Models** Simple and interpretable models that assume a linear relationship between input features and output labels.
 - Logistic regression
 - Linear regression
- ▶ **Non-linear Models** More complex models that can capture non-linear relationships in the data.
 - Decision trees
 - Support vector machines (SVMs)
- ▶ **Neural Networks** Powerful models capable of automatically learning complex patterns in data.
 - Convolutional neural networks (CNNs)

- ▶ **Data Size and Quality** Simple models (e.g., linear models) may work well with small datasets, while more complex models (e.g., neural networks) require large datasets.
- ▶ **Interpretability vs. Accuracy** Linear models are often easier to interpret, while non-linear models or deep learning methods tend to provide higher accuracy at the cost of interpretability.
- ▶ **Computational Efficiency** More complex models might require significant computational resources for training and inference.

- **Confusion Matrix** A matrix that displays the true positives, false positives, true negatives, and false negatives in classification tasks.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Metrics for Classification

- **Accuracy** The proportion of correctly classified instances.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

- **Precision & Recall** Precision measures the accuracy of positive predictions, while recall measures the ability to capture all relevant positive instances.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

How much of the predicted points are precise?

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

How much of the class (positive) are correctly recalled?

- **F1 Score** The harmonic mean of precision and recall, balancing both metrics.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

- **Receiver-Operating Characteristic (ROC) Curve** Used to evaluate the trade-off between true positive rate and false positive rate.

- ▶ **Mean Squared Error (MSE)** Measures the average squared difference between actual and predicted values.
- ▶ **R-squared** Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

- ▶ **Example 1** Image Recognition - Handwritten Digit Classification
- ▶ **Example 2** Speed Recognition - Keyword Detection

► Overfitting vs. Underfitting

- Overfitting: Model performs well on training but poorly on unseen data.
- Underfitting: Model too simple to capture data patterns.

► Imbalanced Datasets Can bias the model towards majority classes.

► Data Quality Issues Missing, noisy, or redundant data.

► Model Interpretability Complex models like deep neural networks can be difficult to interpret.

Loss Function and Regularisation Techniques

Methods like L1 or L2 regularization can prevent overfitting by penalizing large coefficients in models like logistic regression or neural networks.

Example with single layer neural network

Loss function

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2$$

Actual value Calculated value

Loss function with L1 Regularisation (Lasso Regression)

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

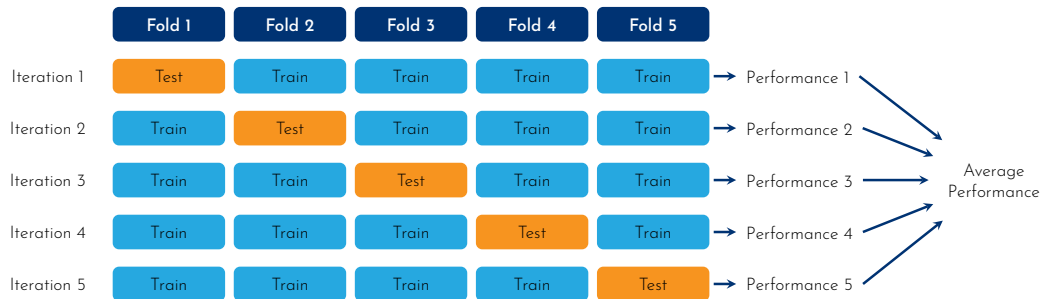
Loss function with L2 Regularisation (Ridge Regression)

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Challenges and Pitfalls

Cross-validation

Used to assess the model's performance on unseen data by dividing the dataset into training and testing sets multiple times.



A large, curved banner with the Sunway University logo and name. The logo features a stylized sun and the text "SUNWAY UNIVERSITY". The banner is supported by several metal poles and is positioned in front of a modern building with a glass facade. The background shows a clear blue sky and some greenery.

SUNWAY
UNIVERSITY

Thank you