Module 2: Introduction to Machine Learning

Learning outcomes

1. Identify the foundational concepts associated with machine learning.
2. Determine whether a given variable is an input or output in a machine learning context.
3. Classify the goal of a machine learning project as either forecasting or inference.
4. Differentiate between machine learning and statistics.
5. Analyse fundamental machine learning concepts, including variable classification, functions, forecasting vs inference and the distinction between machine learning and statistics.
6. Identify data types in machine learning applications.
7. Classify machine learning problems as either prediction or classification.
8. Identify the characteristics of a parametric approach vs a non-parametric approach.
9. Differentiate between supervised vs unsupervised machine learning.
10. Analyse the first three steps of the machine learning process.
11. Analyse the steps required to handle missing data in Python.
12. Implement the steps to handle missing data in Python.
13. Examine the key concepts of machine learning, specifically the major dividing lines in the machine learning landscape and the machine learning process.
14. Identify a real-world machine learning problem for a specific industry.
15. Determine if machine learning is the suitable solution for a specific business problem.

What is machine learning?

- Machine learning involves mapping input variables (X_1, ..., X_p) to an output variable (Y) based on data.
- Key challenges of machine learning include:
    - Limited data points that can complicate the relationships.
    - Noise and stochastic nature of data that can impact accuracy.

Machine learning vs statistics

- Statistics follows defined stochastic models to estimate parameters.
    - Stochastic data models treat data generation as a random process (e.g. normal, binomial, t-distributions).
    - Goodness of fit is a measure that summarises discrepancies between

observed and expected values (e.g. root mean square).
    - Summary statistics describe a data set (e.g., mean, standard deviation, median).
    - Residual analysis observes the difference between observed values and model predictions and is used to evaluate model effectiveness.
- Machine learning treats data as complex, unknown processes to develop predictive functions.
    - Predictive accuracy evaluates how well predicted values match actual values in a test data set.
    - A black box is a system or process whose internal workings are not fully understood or visible.
- The following terms are used in both machine learning and statistics.
    - Model validation is the process of confirming that a model achieves its intended purpose.
    - Generalisation indicates the model's ability to perform well on new, previously unseen data.

Key distinctions in the machine learning landscape

Prediction vs classification
- Prediction – continuous output variable (e.g. estimating house prices)
- Classification – categorical output variable (e.g. identifying spam emails)

Parametric vs non-parametric
- Parametric – assumes a specific form for the function (e.g. linear function)
- Non-parametric – makes no strong assumptions about the function's shape, requiring more data to learn effectively

Supervised vs unsupervised learning
- Supervised learning – involves both input and output variables for learning
- Unsupervised learning – involves only input variables to identify patterns or clusters within the data

Ten steps of a machine learning pipeline

- Clarify the goals and scope of the machine learning project.
- Gather relevant data from internal and external sources.
- Explore, scale and preprocess data.
- Handle missing data and outliers through removal, manual filling or algorithmic methods.
- Remove irrelevant or unavailable variables.
- Transform categorical variables into numeric values if necessary.
- Create new features based on domain knowledge.
- Determine if the task is classification, prediction or unsupervised learning.

- Split the data into training, validation and test data sets for supervised learning.
- Select appropriate machine learning methods based on the defined task.

Examples of machine learning applications

- Predicting patient diagnoses based on medical data
- Detecting fraudulent transactions in financial records
- Analysing customer data to identify clusters for targeted marketing