SIT720 Machine Learning

1.1

Week 1

What is an Algorithm?

An algorithm is a well-defined sequence of instructions used to solve a specific problem or perform a task.

Real-World Applications of Machine Learning

Machine learning uses computer algorithms that enable systems to learn from data and improve over time. Applications include:

- **Robotics** – Autonomous navigation and control
- **Computer Vision** – Image and object recognition
- **Board Games** – AI playing strategic games like Chess and Go
- **Voice Recognition** – Speech-to-text and digital assistants
- **Digit Recognition** – Reading handwritten numbers, such as postal codes

---

Steps in a Machine Learning Workflow

1. **Data Manipulation** – Collecting, cleaning, and preparing data
2. **Analytics** – Applying algorithms to analyze patterns or make predictions
3. **Evaluation and Visualization** – Measuring model performance and interpreting results through visual tools

---

Types of Machine Learning

1. Supervised Learning

Involves training a model on labeled data. Two main types:

- **Regression** – Predicts continuous values (e.g., sales prediction)
- **Classification** – Predicts discrete categories (e.g., spam detection)

2. Unsupervised Learning

Used with unlabeled data to uncover hidden patterns. Key methods:

- **Clustering** – Grouping similar data points
- **Density Estimation** – Modeling the distribution of data
- **Factor Analysis** – Identifying underlying relationships in data

3. Reinforcement Learning

An agent learns optimal actions by interacting with an environment and receiving feedback (rewards or penalties). This approach mimics learning through trial and error.

Model Evaluation

Model evaluation is essential to determine how well a machine learning model performs on unseen (test) data. Common metrics include accuracy, precision, recall, and F1-score.

Model Selection

Model selection involves identifying the best-performing algorithm or hypothesis for a given problem. This includes testing different models and tuning parameters to achieve the best balance between bias and variance.

Python Programming Basics

- **Types of Variables**:

  Python supports various data types such as integers (`int`), floating-point numbers (`float`), strings (`str`), and booleans (`bool`). Variables are dynamically typed.

- **Branching and Decisions**:

  Use `if`, `elif`, and `else` statements to make decisions based on conditions. Example:

  ```
  if age > 18:
      print("Adult")
  else:
      print("Minor")
  ```

- **Iterations**:

  Python supports loops to repeat actions:

  - `for` loop – Iterates over sequences like lists or ranges
  - `while` loop – Repeats as long as a condition is true

    Example:

  ```
  for i in range(5):
      print(i)
  ```

Week 2

---

1. Feature Vectors and Matrices

- In machine learning, **feature vectors** are ordered lists of numerical values representing an object's attributes.
- A **matrix** is a collection of multiple feature vectors, typically used as input data where rows represent instances and columns represent features.

---

2. Probability Concepts

Understanding basic probability is essential for many machine learning algorithms and models.

Random Experiment & Event

- A **random experiment** is a process that leads to uncertain outcomes (e.g., flipping a coin).
- An **event** is a specific outcome or a set of outcomes from the experiment (e.g., getting heads).

Joint Probability

- The probability of two events occurring simultaneously.
- Example: P(A and B)

Conditional Probability

- The probability of event A occurring given that event B has already occurred.
- Written as: $P(A \mid B)$

Bayes' Rule

- A fundamental theorem that allows us to update probabilities based on new evidence.
- Formula:

  $P(A \mid B)=P(B \mid A)\cdot P(A)P(B)$

Random Variable

- A variable that takes numerical values determined by the outcome of a random experiment.
- Can be **discrete** (e.g., number of students) or **continuous** (e.g., height, temperature).

Distribution of Random Variables

- Describes the probabilities of different outcomes for a random variable.
- Common types include **Uniform**, **Normal (Gaussian)**, and **Binomial** distributions.

---

3. Data Wrangling

Data wrangling involves cleaning and transforming raw data into a usable format for analysis or machine learning.

Missing Value Replacement

- Techniques include:
  - Mean/median/mode imputation
  - Forward/backward fill
  - Dropping rows/columns
  - Predictive models for imputation

Scaling or Normalisation

- Converts features to a standard scale.
- Methods:
  - **Min-Max Scaling**: Scales values to a 0–1 range.
  - **Standardisation (Z-score Normalisation)**: Transforms data to have a mean of 0 and a standard deviation of 1.

Non-Numeric Data Encoding

- Converts categorical data into numerical form for model processing.
- Common methods:
  - **Label Encoding** – Assigns a unique number to each category.
  - **One-Hot Encoding** – Creates binary columns for each category.

1.2 Summary of Reading List

- https://d2l.deakin.edu.au/d2l/le/content/1734011/viewContent/8095182/View
- https://d2l.deakin.edu.au/d2l/le/content/1734011/viewContent/8095321/View
- https://d2l.deakin.edu.au/d2l/le/content/1734011/viewContent/8095323/View

1.3 Learning Reflection

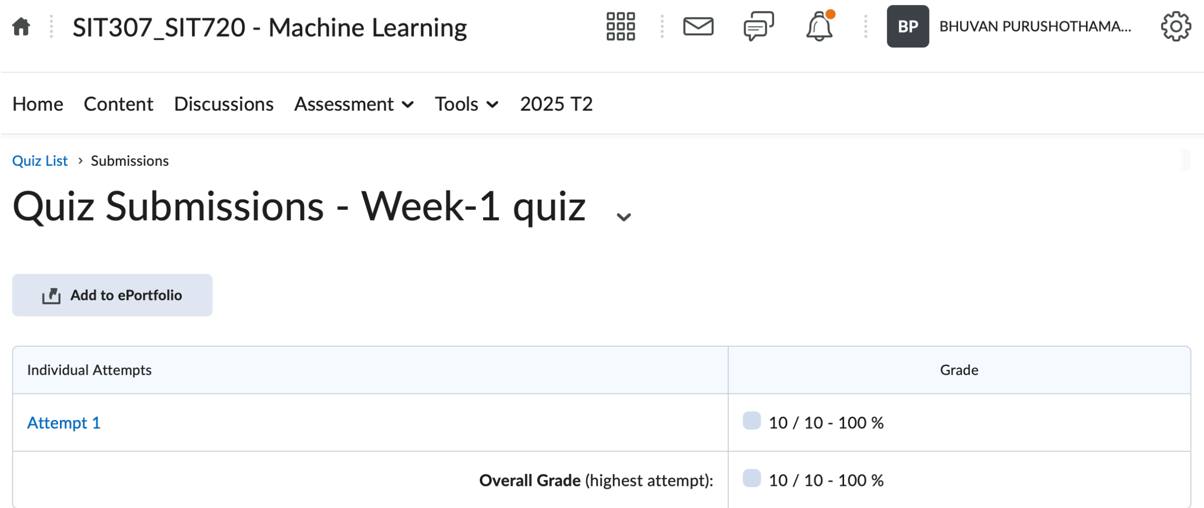
This week deepened my understanding of the statistical foundations behind machine learning. I explored key probability concepts such as **random experiments**, **joint and conditional probability**, and **Bayes' Rule**, which helped me see how models handle uncertainty and update predictions with new data.

I also learned about **random variables** and their **distributions**, which are important when selecting or evaluating models. Understanding **feature vectors and matrices** gave me a better picture of how data is structured in machine learning workflows.

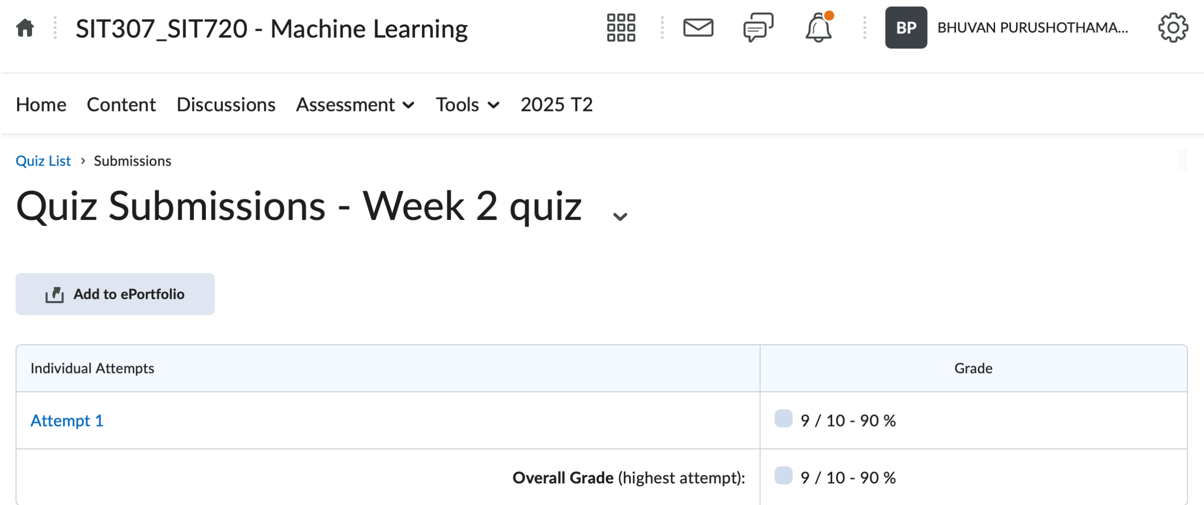
In the practical sessions, I worked on **data wrangling** techniques including handling **missing values**, **normalisation**, and **encoding categorical data**. Applying methods like **mean imputation**, **min-max scaling**, and **one-hot encoding** showed me how vital clean, well-prepared data is for model accuracy.

Overall, this week's content has strengthened my confidence in handling data and applying statistical thinking to machine learning problems.

1.4 Quiz Results



Individual Attempts	Grade
Attempt 1	10 / 10 - 100 %
Overall Grade (highest attempt):	10 / 10 - 100 %



Individual Attempts	Grade
Attempt 1	9 / 10 - 90 %
Overall Grade (highest attempt):	9 / 10 - 90 %