SIT720 Machine Learning

Task 2

Michael Rideout
Student Id: 225065259

Item 1.1 Main Points Summary

Week 1 Main Points Summary

Foundation of ML (ML)

In week 1 we were introduced to the stages in the machine learning lifecycle. Those being Data Gathering, Preprocessing, Model Construction, Evaluation and Deployment

The high level differing types of machine learning methods were shown.

We learnt that **unsupervised learning** is the method of finding patterns in unlabelled data i.e. there are no predetermined targets. Applications of this type of machine learning method included clustering (grouping of data based on similar metrics), anomaly detection (identification of aberrant data points), visualisation and dimensionality reduction to name but a few.

Supervised learning was introduced as the method by which a data has both an input feature set (independent variables) and an output feature (dependent variable). The two most common task in supervised learning are classification (determining which group or class the input features are a part of) and regression (predicting the target continuous value from the input features)

The final method that was introduced was **reinforcement learning.** This is where learning is accomplished through trial and error, based on rewards or punishment for desirable outcomes.

An integral part of machine learning is **model evaluation**. This is the process by which a trained model's utility is measured via some metric. One common trap for trained models is for them to become **overfitted.** This happens when a model learns the training data too well and when presented with unseen data, its performance metric is significantly below that of the training data.

Revising Knowledge of Linear Algebra

The fundamentals of linear algebra were reviewed in week 1. This included the concept of a vector, an ordered list of numbers. Vector operations were then reviewed including transpose (changing a column vector into a row vector), addition (summation of two vectors of equal size), dot product (a scalar resulting from the summation of all element wise products of two vectors), norm (length of a vector), similarity (how alike two vectors are).

Matrices (a set of numbers organised in rows and columns) were also covered. The differing types of matrices were described those being rectangular (columns not equalling rows), square (columns equalling rows), symmetric (transpose is equal to original), diagonal

(diagonal row of a square matrix populated, all other elements are zeros), identity (diagonal matrix with the diagonal values being 1)

Matrix operations were also described. They were transpose (row become columns), addition and subtraction (summation or subtraction of two equal sized matrices), scalar multiplication / division (matrix multiplied or divided by a scalar constant), element wise multiplication (two identically sized matrices have their corresponding elements multiplied), matrix to matrix multiplication (multiplication of matrices of differing sizes), inverse (product of two matrices is inverse if it results in identity matrix)

Week 2 Main Points Summary

Linear Algebra: Feature Vectors and Matrices

In week 2 the concept of a feature vector was introduced as a way to represent documents as vectors. A vocabulary of features was explained as a predefined list of all possible features that could appear in a feature vector. An example being the set of all words (bag of words)

Probability Concepts

An overview of probability concepts was given. Probability is based on the idea of random experiments i.e. a process or action where the outcome cannot be predicted. The probability of an event is described numerically in the range of [0,1] with 0 having 0% probability of the event occurring and 1, 100% probability of it occurring.
Joint probability was described as the combined probability for two events if they are independent. The joint probability given by $P(A \text{ and } B) = P(A) * P(B)$
Conditional probability is the probability of A occurring if B has already occurred
$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$,
Bayes Rule was covered which is a way to incorporate prior knowledge of events with new data to at a more accurate probability. The formula being: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$,

Random Variable

Random variables were described as a probability score assigned as a result of a random event from an experiment. Discrete random variables are for recording random experiments where the event has a countable result (eg the toss of a coin). The probability mass function

is related to discrete random variables for a given value it provides the probability that value will occur.
Continuous random variables represent random experiments that are not countable (eg the height of a person). The function to provide the probability for a given continuous random variable is called the probability density function.

Distribution of Random Variables

Week 2 also provided an overview of probability distributions, which are mappings between outcomes of random experiments and the chance of those outcomes occurring. A Bernoulli distribution is for a binary random variable (eg coin toss). A normal distribution is for continuous random variables and is determined by the formula

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{1}{2\sigma^2}(x-\mu)^2$$

The central limit theory is the theory that states that if you calculate the means of samples from any population, the distribution of the sample means will itself be a normal distribution

Data Wrangling

Data wrangling is the process of manipulating, clearing and transforming a dataset in preparation for analysis or model training. Scaling (changing the range while maintaining distribution) and normalisation (transforming data to have mean of 0 and standard deviation of 1) were explored

1.2 Summary of Reading List Items

Week 1 Readings

1.4 Defining Machine Learning

Wikipedia ‘Machine Learning’ Article

The wikipedia machine learning article, covers the history, application, and types of methods used in the field.

1.14 Matrix Algebra

Introduction to Eigenvalues (video)

The video introduced eigenvalues and eigenvectors showed the relationship between a matrix, an eigenvector and eigenvalue

Finding Eigenvalues and Eigenvectors (video)

This video demonstrated how to find eigenvalues and eigenvectors using the determinant.

Eigenvectors and eigenvalues (video)

This video explained the foundational aspects of linear transformations required to understand eigenvalues and eigenvectors

Week 2 Readings

Understanding Random Variables (video)

This video focused on the difference between discrete and continuous random variables. It showed how to calculate probabilities from discrete distributions.

Python Libraries

Python libraries utilised in this task include:

- **Random:** Standard library used for generating pseudo-random numbers
- **Hashlib:** Suite of hashing algorithms
- **Tabulate:** Plain text tables
- **Pandas:** Data manipulation and analysis library
- **Numpy:** Library for scientific and numerical computing
- **Mathplotlib:** Visualisation library
- **Seaborn:** Visualisation library
- **Sklearn:** Machine learning toolkit

1.3 Learning Reflection

Week 1 and 2 provided a valuable introduction to the field of Machine Learning. The differing classes and methods available in machine learning were covered. The mathematical foundations to many of these methods were introduced. Those being linear algebra and probability. Data wrangling was also touched upon with regards to scaling and normalisation of datasets.

On the practical side, I learnt about the python libraries for machine learning, pandas, numpy and mathplotlib. These help to reinforce the concepts that were covered in the lectures.

1.4 Quiz Results

Week 1 quiz:

Week-1 quiz

Your work has been saved and submitted

Written 29 March, 2025 10:21 PM - 29 March, 2025 10:23 PM • Attempt 2 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score	**9 / 10 - 90 %**
Overall Grade (Highest Attempt)	**9 / 10 - 90 %**

Week 2 Quiz:

Week 2 quiz

Your work has been saved and submitted

Written 30 March, 2025 1:59 AM - 30 March, 2025 2:06 AM • Attempt 1 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score	9 / 10 - 90 %
Overall Grade (Highest Attempt)	9 / 10 - 90 %

Coding Task 9 Graphs