COMP3003
MACHINE LEARNING

20 CREDIT MODULE

ASSESSMENT: 100% Coursework **W1: 30% Set Exercises**
W2: 70% Report

MODULE LEADER: Dr. Lingfen Sun

MODULE AIMS

- To introduce the area of machine learning (ML) covering unsupervised, supervised and reinforcement learning from Bayesian perspectives.
- To give students the theory behind a range of learning techniques and how to apply these to build representations of data in systems that make decisions and predictions.
- To enable students to analyse real datasets and control real-time systems.

ASSESSED LEARNING OUTCOMES (ALO):

1. Apply the concepts of unsupervised, supervised and reinforcement learning to tackle machine learning problems.
2. Implement and apply machine learning techniques to make decisions on real data sets and real-time control scenarios.

Overview

This document contains all the necessary information pertaining to the assessment of *COMP3003 Machine Learning*. The module is assessed via **100% coursework**, across two elements: *30% Set Exercises* and *70% Report*.

The sections that follow will detail the assessment tasks that are to be undertaken. The submission and expected feedback dates are presented in Table 1. All assessments are to be submitted electronically via the respective DLE module pages before the stated deadlines.

	Submission Deadline	Feedback
Set Exercises (30%)	**29/11/2022**	Within 20 working days
Report (70%)	**16/01/2023**	Within 20 working days

Table 1: Assessment Deadlines

All assessments will be introduced in class to provide further clarity over what is expected and how you can access support and formative feedback prior to submission. Whilst the assessment information is provided at the start of the module, it is not necessarily expected you will start this immediately – as you will often not have sufficient understanding of the topic. The module leader will provide guidance in this respect.

Assessment 1: Set Exercises

This assignment contributes to **30%** of the overall module mark for COMP3003 and is an **individual assignment**.

The focus of this coursework is to assess your understanding of unsupervised machine learning techniques. You are required to write Matlab code to implement the Kmeans clustering algorithm from 1st principles. This is an extension of Lab 4 on Kmeans clustering. The detailed tasks are as below.

Write Matlab code to implement the Kmeans clustering algorithm from 1st principles. You should demonstrate clustering for 3 clusters with each cluster created based on a 2D uncorrelated Gaussian distributed dataset. You are free to make assumptions on the number of points in each cluster and the degree of overlapping among clusters. The assumptions should be clear enough to demonstrate the concept of Kmeans clustering and clearly show the limitations of Kmeans approach (through your explanation).

You are required to write a report to cover the tasks above. The ***individual*** report should be no more than 1**,500** words (excluding diagrams, images, tables, Matlab code/comments, and references). The report should be organised as follows:

- Abstract: about 150 words
- Introduction: unsupervised learning and Kmeans clustering – about 450 words
- Implementation: implementation of Kmeans clustering (including implementation steps, Matlab code, results/screenshots, performance of clustering, and explanations) – about 600 words
- Discussions and conclusions (including drawbacks of Kmeans clustering) – about 300 words
- References
- Appendix (including all your Matlab code)

Any references should be appropriately cited in the report. Harvard referencing style is recommended. You should write your report as concisely as possible and it is important that you do not exceed or within 10% of the allowed word limit.

Assessment Criteria:

The report will be assessed based on the following criteria.

- Abstract, Presentation, Structure, Conclusions, References (20%): are the abstract and conclusions described appropriately? Are discussions sufficient? Is the report well-presented and structured? Are the references appropriate?

- Introduction of unsupervised machine learning and K-means (20%): are unsupervised learning and K-means clustering concepts explained appropriately? Is it easy to follow? Is it clear?

- Implementation (60%): Are the implementation steps clearly described? Is evidence (e.g. screenshots) sufficient? Have the results been explained and/or analysed appropriately? Is the Matlab code well explained and appropriate?

Please submit the report as a single PDF on the DLE.

Assessment 2: Project Report

This assignment contributes to **70%** of the overall module mark for COMP3003 and is an **individual assignment**.

The focus of this coursework is to assess your understanding of machine learning techniques, with a focus on supervised learning and reinforcement learning, and their practical applications. You are required to (1) undertake a literature review on the above two machine learning techniques and identify two real-world applications for each approach; (2) write Matlab code to develop supervised learning models (neural networks) for intrusion detection in network security applications. The detailed tasks are as below.

Task 1: Literature review
Undertake a review of machine learning techniques, with a focus on supervised learning techniques. Understand differences between supervised learning and reinforcement learning. Identify two real-world applications for each approach.

Task 2: Write Matlab code to build a network intrusion detector based on supervised learning. You will develop neural network models for intrusion detection based on a dataset obtained from real network traffic.

Download a ZIP file on the DLE, called "**COMP3003_CW.zip**". This archive contains a dataset, which is extracted from a 10% subset of KDD full dataset (Knowledge Discovery and Data Mining - KDD99 Intrusion Detection Dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html). The KDD99 dataset contains 41 features linked with over five million TCP/UDP connection records. Each connection is labeled as either a 'normal' (good) connection, or a 'bad' connection (called an intrusion or attack). In the dataset, columns 1 – 41 are for features, the 42nd column is for traffic type, either normal, or an attack. For the dataset provided for this coursework, it contains 3,000 connections with 2000 'normal', 500 for 'neptune' and 500 for 'portsweep', where 'neptune' belongs to Denial of Service (DoS) attack, and 'portsweep' for probing attack. Two data files are provided as below:

(1). cw_dataset_original.txt (the dataset extracted from KDD99 including symbolic features)
(2). cw_dataset_processed.txt (Symbolic features are converted to numerical labels, e.g., the 2nd column, protocol type, tcp=> 1, udp=>2, and icmp=>3; the 3rd column, 32 service types such as ftp and http, are converted to 1 to 32, respectively. Column 42 remains unchanged.

Tasks:

a) Carry out Principal Component Analysis (PCA) for 41 input features and convert them to an appropriate number of Principal Components (PCs) for developing intrusion detection classifiers for 2 classes (i.e., 'normal' and 'attack') and 3 classes (i.e., 'normal', 'neptune' attack and 'portsweep' attack).
b) Same as a), develop intrusion detection classifiers for 2 classes and 3 classes without PCA.

c) Performance analysis and comparison between Task a) and Task b). Performance needs to include accuracy and recall rate with confusion matrix etc.

You are free to make assumptions about dataset splitting between training, validation and testing datasets. You are also free to choose a neural network classifier, e.g., pattern recognition neural network (patternnet) and set relevant learning parameters in Matlab.

You are required to write a report to cover the tasks above. The ***individual*** report should be no more than **3,000** words (excluding diagrams, images, tables, Matlab code/comments, and references). The report should be organised as follows:

- Introduction – about 200 words
- Literature review (on supervised and reinforcement machine learning techniques and two real-world applications for each) – about 1000 words
- Implementation of supervised learning (neural network models) in Matlab for intrusive detection (methodology, datasets, classifier development with/without PCA, including key Matlab codes with explanations, and key steps for model development) – about 1000 words
- Results and performance analysis/comparison – including accuracy and recall rate, confusion matrix etc. for performance analysis/comparison – about 500 words.
- Discussions and conclusions – about 300 words
- References
- Appendix (including all your Matlab code)

Any references should be appropriately cited in the report. Harvard referencing style is recommended. You should write your report as concisely as possible and it is important that you do not exceed or within 10% of the allowed word limit.

Assessment Criteria:

The report will be assessed based on the following criteria.

- Introduction/Conclusions, Presentation, and Structure (25%): are the introduction and conclusions described appropriately? Is the report well-presented and structured? Are the references sufficient and properly cited in the report?

- Literature review (20%): Has literature review been carried out and described appropriately? Are two real-world applications for each machine learning approach (i.e., SL and RL) well -presented?

- Methodology and Implementation (40%): Is the methodology described appropriately? Are assumptions made clearly and appropriately? Are the implementation steps clearly described? Is evidence (e.g. screenshots) sufficient? Is the Matlab code well explained and presented? Are all tasks completed?

- Results and performance analysis/comparison (15%): Have the experimental results been explained and analysed appropriately? Are the results correct and illustrated appropriately? Is the performance comparison clear?

Please submit the report as a single PDF on the DLE.

Threshold Criteria (these are indicative only):

< 40% Little or no analysis. Minimum of results, and results are largely incorrect. Little understanding of the subject. Almost no evidence of investigation, evaluation and research on answering the questions. The report is poorly written and structured.

40–49% (Third): Brief discussion and little analysis/investigation. Results are partially correct and/or complete. Little evidence of investigation, evaluation and research on answering the questions. The content and style of the report are mostly appropriate.

50–59% (Lower Second): The majority of the results are correct and complete. Answers are given at an appropriate level of detail and are explained clearly. Some evidence of investigation, evaluation and research on answering the questions. The report is reasonably well structured and the content and style are appropriate.

60 – 69% (Upper Second): A significant majority of the results are correct and complete. Answers are given at great details and are well explained. Clearly and concisely description of how results are obtained. Good evidence of investigation, evaluation and research on answering the questions. The report is of a good standard and structure.

> 70% (First): The results are correct and complete. Especially clear, ambitions and well justified analysis and description. Demonstrating ideas for original thoughts and stretched work. There is strong evidence of investigation, evaluation and research on answering the questions (e.g. deep analysis, fully investigation, good summarise of the investigation). The report is well presented and organised (focused and concisely).

General Guidance

Extenuating Circumstances

There may be a time during this module where you experience a serious situation which has a significant impact on your ability to complete the assessments. The definition of these can be found in the University Policy on Extenuating Circumstances here:
https://www.plymouth.ac.uk/uploads/production/document/path/15/15317/Extenuating_Circumstances_Policy_and_Procedures.pdf

Plagiarism

All of your work must be of your own words. You must use references for your sources, however you acquire them. Where you wish to use quotations, these must be a very minor part of your overall work.

To copy another person's work is viewed as plagiarism and is not allowed. Any issues of plagiarism and any form of academic dishonesty are treated very seriously. All your work must be your own and other sources must be identified as being theirs, not yours. The copying of another persons' work could result in a penalty being invoked.

Further information on plagiarism policy can be found here:

Plagiarism: https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations/plagiarism

Examination Offences: https://www.plymouth.ac.uk/student-life/your-studies/essential-information/exams/exam-rules-and-regulations/examination-offences

Turnitin (http://www.turnitinuk.com/) is an Internet-based 'originality checking tool' which allows documents to be compared with content on the Internet, in journals and in an archive of previously submitted works. It can help to detect unintentional or deliberate plagiarism.

It is a formative tool that makes it easy for students to review their citations and referencing as an aid to learning good academic practice. Turnitin produces an 'originality report' to help guide you. To learn more about Turnitin go to:
https://guides.turnitin.com/01_Manuals_and_Guides/Student/Student_User_Manual

Referencing
The University of Plymouth Library has produced an online support referencing guide which is available here: http://plymouth.libguides.com/referencing.

Another recommended referencing resource is Cite Them Right Online; this is an online resource which provides you with specific guidance about how to reference lots of different types of materials.

The Learn Higher Network has also provided a number of documents to support students with referencing:
References and Bibliographies Booklet:
http://www.learnhigher.ac.uk/writing-for-university/referencing/references-and-bibliographies-booklet/
Checking your assignments' references:
http://www.learnhigher.ac.uk/writing-for-university/academic-writing/checking-your-assigments-references/