INT104 – Artificial Intelligence

Coursework 2

Introduction

A spreadsheet of mark is provided where the mark for each question is listed for each student. There are more than 500 students in total who comes from 4 majors. The majors in the spreadsheet are labelled as '1', '2', '3' and '4'. There are also a few students that belong to other majors, which are labelled as '0'. This coursework requires the student to design a classifier to classify students to the program they belonged to, according to the mark earned for each question.

There are three lab sessions planned for this coursework, where you must demonstrate your progress to your assigned TA. Each lab session corresponds to a task in this coursework, which will be described in the following sections. After all tasks are finished, you are expected to submit a lab report documenting the whole process of your attempt on design a classifier.

Though Python is recommended to be used as the implementation tool for this coursework, other tools / programming language such as Microsoft Excel, Weka, MATLAB, Java and C++ are also acceptable. However, the student should use only **_ONE_** type of tool throughout three tasks in this coursework. The lab report should be submitted in PDF format, which could be produced by any typesetting software such as Microsoft Word, Latex (recommended) and Markdown.

This coursework will be marked by the lab report submitted with a given marking criteria. However, the demonstration of progress over lab session is **_compulsory_** and failing to demonstrate your progress to the assigned TA over **_ANY_** lab session will result to **_a mark of 0_** in this coursework. The demonstration process during each lab may include showing your code and report draft, which demonstrates that the time management ability of the students and avoids possible cheating on programming.

Task 1: Data observation

With the presented data, the student is required to make reasonable data observation in this task. The results of data observation could be presented as data visualisation. As we have learned from the lectures, all data contains bias that is caused by different reasons. The bias of some features in data could lead to a good classifier to a dedicated machine learning task.

This task requires the student to find their own ways to show the data bias and encourages the student to find data bias that is closely related to the given labels. Despite commonly used, data features such as average value, median value, range and standard deviations are not necessary to be the features showing data bias. Data features resulted from more complex mathematical operations such as Principal Component Analysis (PCA), Discrete Fourier Transform (DCT) and Non-negative Matrix Factorisation (NMF). For different features, there might be a bias of distribution hence a process of removing bias in features should also be considered and presented.

In this task, the student should try different ways to extract data features of the given dataset and decide which data feature will be used as the input of classifier that to be developed in the next task. In the lab report, a full and detailed justification on feature selection (including bias removal process) should be presented with no less than five features as candidates.

Task 2: Build Classifiers

With the selected data features, this task requires the student to design ***three*** classifiers to classify which program a student is enrolled, according to the marks awarded for each question in an exam. For a better performance of the proposed systems, data features may be used as the input of the candidate system, which replace the raw data given in the dataset.

The candidate classifier should be built in a supervised way. The classifier could be the method taught in the lectures such as Support Vector Machine (SVM) and Decision Tree. However, the student is also encouraged to try methods beyond the scope of the lectures delivered such as deep neural networks and Bayesian graphical models.

In the lab report, the student should specify what method is used to build the candidate classifiers, what data feature is used as the input of the candidate system and what result is obtained. The key part of program such as data pre-processing, the training process of the model and the inference process of the system should be stated with details. However, the use of screenshot should be avoided, where a text-based description (such as text-based source code with line number) is preferred when necessary. (NOTE: source code presented in the lab report should be in `Courier New` or similar fonts.)

The student is then required to recommend ONE classifier among the three candidate classifiers. The process of classifier evaluation should be fully demonstrated (e.g., cross validation process). The decision of recommendation should also be fully justified. The process of classifier evaluation and the justification of classifier choice should be documented in the lab report.

Task 3: Unsupervised Classification for Student Classification

In this task, the student is required to classify students to different groups *in an unsupervised manner*, according to marks awarded for each exam question. The classification of groups should make the delivery of lecturers in the next semester easier by classifying students in the same group sharing comparable properties of studying. There are no dedicated ways of classification hence the student should make their choice of principle applied to classification (i.e., ***there are no dedicated number of groups that the student should be classified into***).

In the lab report, the student is expected to present the full details of the unsupervised classification process and fully justify the final decision made for student classification. Specifically, the student should interpret the principle that is followed for the classification. As the case of earlier tasks, screenshots of source code should be avoided.

Lab Report

The lab report should follow an IEEE-like format in double column with the following sections:
- Abstract
- Introduction
- Data Observation
- Candidate Classifiers
- Evaluation and Selection of Candidate Classifiers
- Conclusion

Please note, a title should also be included in the lab report. (Please do NOT use "lab report" as the title of the report.) ***There are no word limit or page limit for the lab report*** if the student is happy with the content of the lab report. The majority of marks in this coursework will be awarded ***according to the content of lab report***. Please be aware that ***a lengthy report does not guarantee a high mark for this coursework***.

Marking criteria

Marking Criteria	Item	Marks	Description
Editorial and Language Issues (20)	Formatting	5	Format should follow IEEE-like style with double columns: • Font size: 9 or 10 • Line spacing: 1 • Alignment: Justified • Font: Times New Roman • Margins: Conventional

	Organization	5	The paper should be well organised showing a clear structure of the report as instructed in this material.
	Spelling	5	The spelling in the paper should be completely correct. Any incorrect spelling can result in a loss of marks.
	Grammar	5	The language used in the essay must be academic language.
Task 1 (20)	Feature Extraction	10	Showing reasonable amount of feature extraction attempts.
	Data Visualisation	10	Visualise the extracted features of raw data to show the bias of distribution.
Task 2 (30)	Description of Classification Methods	10	Describe the selected methods of classifications with justification.
	Training Classifiers	10	Training three classifiers and obtain predictions.
	Classifier Selection	10	Using the obtained classifiers to classify given data and evaluate performance of the candidates
Task 3 (20)	Unsupervised Classification	10	Classify students into several groups via an unsupervised classification method
	Critical thinking for Classification	10	The students should interpret the classification results with a reasonable interpretation of language.
Lab Sessions (10)	Time Management	10	The assigned TA will judge whether the coursework has been proper progressed with lab session. Last-minute rush will be discouraged.
	Demonstration	Critical	***Missing live demonstration to assigned TA over any lab session will result a 0 mark in this coursework***

Repetition check	This coursework will be checked by *Turnitin* for the repetition rate of your submission. If a repetition rate ***higher than 20%*** is founded, a 0 mark is likely to be awarded.

Submission

1. Only submissions in PDF format are accepted.
2. Submit your lab report via the dedicated Learning Mall coursework link.
3. Please name your submission file as ID_FirstName_LastName_C2.pdf (e.g., 1234567_FirstName_Surname_C2.pdf).
4. Late submission is NOT allowed.