7th November 2024

PRACTICAL COURSE WORK FOR MACHINE LEARNING (H) COMPSCI 4061

Code of Assessment Rules for Coursework Submission

Deadlines for the submission of coursework which is to be formally assessed will be published in course documentation, and work which is submitted later than the deadline will be subject to penalty as set out below. The primary grade and secondary band awarded for coursework which is submitted after the published deadline will be calculated as follows:

1. In respect of work submitted not more than five working days after the deadline
   (a) the work will be assessed in the usual way;
   (b) the primary grade and secondary band so determined will then be reduced by two secondary bands for each working day (or part of a working day) the work was submitted late.
2. Work submitted more than five working days after the deadline will be awarded Grade H.

Penalties for late submission of coursework will not be imposed if good cause is established for the late submission. You should submit documents supporting good cause via MyCampus. Penalty for non-adherence to Submission Instructions is 2 bands You must complete an "Own Work" form via `https://studentltc.dcs.gla.ac.uk/` for all coursework

Submission and deadline

Your submission will be:

- a report in PDF which contains the answers to all questions including discussions, algorithms, plots and figures.
- the source code of all the implementations required to conduct the experiments.

Deadline: You must submit this on Moodle by 6.00 PM, Friday 13th December 2024

Applying Machine Learning for Breast Cancer Detection

- **Dataset**: The **dataset** that we are going to use is the `Breast Cancer Wisconsin (Diagnostic)` dataset[1]. The dataset is comprised of **569 instances** and **30 features**. You can find more details on the dataset page, including sample code on how to load the dataset up in Python.
- **Prediction Objective**: On this dataset, the **prediction task** is to estimate the target variable `diagnosis`, which is either 'B' (benign) or 'M' (malignant) for each set of input features (there are 30 of them, e.g., radius, smoothness etc.).

As a part of the coursework, you need to conduct the following tasks.

Submission Guidelines:

- This is **not just a coding exercise**. Rather, you have to submit a **PDF-formatted report** with functionally correct code.
- Marks will be deducted if you don't submit code or your submitted code is incomplete, functionally incorrect, or errors out.
- In your report, you have to clearly describe **how you have worked towards the solution** of each task (preferably with examples). Your report should be easily readable and sufficiently clear for the examiners to understand conceptually.
- **The examiner should also be able to execute your code** (ideally a Jupyter/Colab notebook) for validation purposes.

Tasks

Complete each of the following tasks (marks appear alongside each task).

Q 1 (10 marks)**.** *Apply any classification model (e.g., SVM, logistic regression etc.) for this task of cancer detection from the input features. You are free to use **any programming language library** for this. Note that the dataset doesn't come with a train:test split, which means that you either will need to create your own split (clearly mention the split ratio) or conduct a k-fold validation (again, these details must be present in your report). Clearly describe your classifier model settings and hyper-parameters.*

Q 2 (10 marks)**.** *How should you evaluate your model for this task? Is this a precision or a recall oriented task? How does your model perform in terms of per-class precision, recall and overall accuracy?*

Q 3 (10 marks)**.** *Report an investigation (with empirical findings) on how you can calibrate your model towards achieving a higher precision or recall **by changing the hyper-parameters of your model** (e.g., by changing the kernel function if you're using SVMs, or by using a different activation function/reguralisation for logistic regression etc.).*

Q 4 (10 marks)**.** *Report an investigation (with empirical findings) on how you can calibrate your model towards achieving a higher precision or recall by changing **only how the predictions map to the two classes**, i.e., **model calibration**.*

[1] `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`