Durham University

Computer Science

Summative Assignment

Module code and title	COMP2261 Artificial Intelligence
Academic year	2023-24
Coursework title	Machine Learning
Coursework credits	10 credits
% of module's final mark	50%
Lecturer	Yang Long
Submission date*	Thursday, January 18, 2024 14:00
Estimated hours of work	20 hours
Submission method	Turnitin (report) and Ultra (code)

Additional coursework files	*Dataset:* https://archive.ics.uci.edu/
Required submission items and formats	*Code:* *[CIS_username].zip* *Report:* *[CIS_username].pdf*

* This is the deadline for all submissions except where an approved extension is in place.

Late submissions received within 5 working days of the deadline will be capped at 40%.
Late submissions received later than 5 days after the deadline will receive a mark of 0.
It is your responsibility to check that your submission has uploaded successfully and obtain a submission receipt.

Your work must be done by yourself (or your group, if there is an assigned groupwork component) and comply with the university rules about plagiarism and collusion. Students suspected of plagiarism, either of published or unpublished sources, including the work of other students, or of collusion will be dealt with according to University guidelines (https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/).

Summative Assignment

COMP2261 Artificial Intelligence – Machine Learning

Deadline for submission: 18 January 2024 14:00 (GMT)

1 Introduction

This project focuses on developing machine learning skills for independent work. The project can be chosen from a variety of topics and datasets. Your first task is to scope your project. That includes finding a dataset and defining the machine learning tasks. Your project should be feasible to achieve before the submission deadline. It should have a significant amount of experimentation, such as data exploration and pre-processing, model training and evaluation, comparisons, analysis, and discussions. There are three options for the project as follows.

First option: Use a dataset that interests you from https://archive.ics.uci.edu/. In this task you will need to explore the datasets and choose one for your project. Following the machine learning workflow we introduced in the lecture, you will need to download data, frame the problem, prepare the data, choose models, evaluate using quantitative and qualitative methods, according to the marking scheme below.

Second option: An example is provided http://www.cs.ubc.ca/projects/local-nbnn/ and you can build your project upon it. You can reuse the code but make sure you put your own work in a separate main file and import the reused code if possible. Although the task is specific to image classification, the marking scheme is the same as the first option. You will need to report your machine learning workflow according to the marking scheme below.

Third option: You can choose your own dataset. The challenge, compared to the first two options, is that you need to clearly introduce your dataset including key statistics, problem framing etc. Again, the marking scheme is consistent with the first and second options. Note that only a link to the dataset is needed to be submitted, not the raw dataset.

2 Submission and Marking Scheme

2.1 Code file (0%)

The file should have the name of [CIS_username].zip (or .tar), and it should include all the notebook (.ipynb) files you used for the project. In each notebook file, comments should be given when necessary. Your code files are not going to be marked, but it is your responsibility to make sure the submitted code is working. Plagiarism and collusion checks will be performed on your code files. You may be subject to a brief viva in February, to ensure it is your own work.

2.2 Project report in PDF (100%)

The PDF file should have the name of [CIS_username].pdf, 4 pages in total. You are encouraged to draft your report using IEEE journal Latex template. Compile the latex and make the PDF submission on Ultra.

2.2.1 Report Summary (30%)

Provide a brief introduction to the project. Explain why this task is interesting and important. What is the background of real-world problems? How could the machine learning project potentially solve these problems? Precisely define the problem you are solving, i.e., formally specify the inputs and outputs. Frame the problem as a machine learning task and use a figure to illustrate the proposed ML system.

Summarise the project: what the project was about, what you did, what the results were, the major limitations of the approach you used, what could have been done to improve the procedure and the result, the lessons you learned through this project.

List all the external references.

>> Specific assessment criteria:

- Provides a clear, concise, and interesting introduction, highlighting the importance and relevance of the project.
- Clearly defines the problem with formal specifications of inputs and outputs.
- Effectively frames the problem as a machine learning task and includes a well-designed figure illustrating the proposed ML system.
- Summarises the project thoroughly, including key results, limitations, and lessons learned.
- Lists all external references in a proper format.

2.2.2 Data and Experimental Setup (30%)

Summarise the main characteristics of the dataset, using tables and statistical graphics, and/or other data visualisation methods. Describe how you split the dataset and present stats such as count, mean, etc. Describe how you constructed and/or transformed the data/feature. Describe the evaluation metrics and explain why they are appropriate.

Note: for any decision you made, provide clear and appropriate justifications. Marks are generally awarded for a good experimental procedure that supports the results, not for achieving the best performance. Deduction of the final mark may be made for not submitting files as requested, e.g., file names/extensions, report formatting (e.g., captions, etc.)

>> Specific assessment criteria:

- Provides a comprehensive summary of the dataset, using tables, statistical graphics, and other data visualization methods.
- Clearly describes the dataset split and presents relevant statistics such as count, mean, etc.
- Thoroughly describes data construction/transformation and the rationale behind these decisions.
- Selects appropriate evaluation metrics and justifies their relevance to the project.
- Follows all required file naming/extensions and report formatting guidelines.

2.2.3 Model Evaluation (30%)

Choose two machine learning algorithms from this submodule. The choice of the two models are based on the data exploration from 2.2.2. One model is chosen as a baseline approach whereas the other is chosen as the proposed model. Technical justification is needed to explain why the proposed model could perform better than the baseline model.

Your main results and evaluation should focus on machine learning methods. Explain why the chosen algorithms are appropriate for the project. Deep learning or reinforcement learning are not assessed but can be as a part of the process, e.g. feature selection or dimension reduction. For example, you use CNN for visual feature extraction and then you compared linear regression, and knn models. In this case, CNN will be part of the data preparation. Only machine learning algorithms are assessed in this section.

Describe the training process, including the parameters involved and how they fit, concerns about underfitting and/or overfitting, and concerns about the convergence of the optimisation. Describe the hyperparameter selection and tuning process, including the hyperparameters involved and how they were selected and tuned, the candidate values that were considered, and the performance metrics that were used for optimisation. Describe how you compared the trained models, using tables and statistical graphics and/or other data visualisation methods.

>> Specific assessment criteria:

- Selects two appropriate machine learning algorithms based on data exploration, with one as a baseline and the other as a proposed model.
- Provides technical justification for the proposed model's potential superiority over the baseline model.
- Describes the training process, including parameters, concerns about underfitting/overfitting, and optimization convergence.
- Explains the hyperparameter selection and tuning process, including candidate values and performance metrics used for optimization.
- Compares trained models using tables, statistical graphics, or other data visualization methods.

2.2.4 Self-Evaluation Report (10%)

Use up to 1 page (included in the four pages) at the end of the whole report to answer the following five questions.

- What have you learned from the lectures?
- What have you learned from the coursework?
- What are your difficulties in the module?
- What would you do differently if you were to do it again?
- Are there unique contributions or novel ideas that make your project different from existing machine learning work.