RMIT UNIVERSITY

COSC2673 Semester 1 2024
Machine Learning

Assignment 1
Introduction to Machine Learning

Weight: 30% of the final course mark
Type: Individual
Due Date: 5.00pm, Monday 8th of April 2024 (Week 5)
Learning Outcomes: This assignment contributes to CLOs: 1, 3, 4
Note: Marks will be awarded for meeting requirements as close as possible. Clarifications/Updates may be made via announcements / relevant discussion forums, you are required to check them regularly.

1 Introduction

1.1 Summary

In this assignment you will explore a modified real dataset and practice the typical machine learning process. This assignment is designed to help you become more confident in applying machine learning approaches to solving tasks. In this assignment you will:

1. Selecting the appropriate ML techniques and applying them to solve a real-world ML problem.
2. Analysing the output of the algorithm(s).
3. Research how to extend the modelling techniques that are taught in class.
4. Providing an ultimate judgement of the final trained model that you would use in a real-world setting.

To complete this assignment, you will require skills and knowledge from lecture and lab material for Weeks 1 to 4 (inclusive). If you have already started with the assignment quick pick, you have already have the tools for kick starting this assignment. You may find that you will be unable to complete some of the activities until you have completed the relevant lab work. However, you will be able to commence work on some sections. Thus, do the work you can initially, and continue to build in new features as you learn the relevant skills. A machine learning model cannot be developed within a day or two. Therefore, start early.

This assignment has four deliverables:
Please note, you will not receive any mark if you don't submit all 4 assignment deliverables. While you must submit all the 4 deliverables, you will only be marked based on your PDF report and presentation. That means we will not consider any part of your code that your presentation and report does not cover.

1. *A PDF report, preferably converted form of notebook, following bellow criteria*:
   - Bullet point format:
   Bullet point is where you raise each point in one bullet point, using clear topic, then explain the important detail as summary under the title (This description is a clear example). A report must be no more than 4 pages, (plus up to 2 pages for possible references and graphs).
   - Graphs:
   Your report should include the graphs produced by your analysis.
   - Markdown:
   If you are using notebook, it needs to be in the format of the provided tutorials. That means the report should include markdown text explaining the rational, critical analysis of your approach and ultimate judgement.
   - Specification:
   The report needs to be self- explanatory, well structured, and fulfill all the assignment specifications.

2. *A video presentation, following bellow criteria:*
   - Presentation format:
   You need to use your PDF report as the basis of your presentation. In the presentation you will go through each bullet point and explain it in detail. That should include your judgment.
   - Presentation length:
   Your presentation should be10 minutes (minimum of 9, and maximum of 11 minutes). You should not exceed 11 minutes, as you will be marked only based on the first 11 minutes of your presentation. You will lose mark if it is less than 9 minutes.
   - Must cover:
   Fulfill all the assignment specifications, based on your PDF report. You should share your window containing your PDF report while presenting, and have your camera on so your face in also captured in the video.

3. *A set of prediction, following bellow criteria:*
Your prediction must be based on your final method and your ultimate judgement. The sample solution is included, the ID need to include the ID of the selected data from Data_Set that makes up your test set (manual selection is not acceptable).

4. *Your Jupyter notebook, following bellow criteria*:
Your Jupyter notebook, used to perform your modelling & analysis with instructions on how to run them, which need to have embedded explanatory comments. Remember that code is only used for reference, and unless you also include your comments in the report and presentation, you will not receive any mark for them.

Please note, you will not receive any mark if you don't submit all 4 assignment deliverables. While you must submit all the 4 deliverables, you will only be marked based on your PDF report and presentation. That means we will not consider any part of your code that your presentation and report does not cover.

1.2 Learning Outcomes

This assignment contributes to the following course CLOs:

- **CLO 1:** Understand the fundamental concepts and algorithms of machine learning and applications.
- **CLO 3:** Set up a machine learning configuration, including processing data and performing feature engineering, for a range of applications.
- **CLO 4:** Apply machine learning software and toolkits for diverse applications.

1.3 Academic Integrity

Academic integrity is about honest presentation of your academic work. It means acknowledge the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods

- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source

- Copyright material from the internet or databases

- Collusion between students

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/ academic-integrity.

2 Task

In this assignment, you will predict the life span of a human based on several attributes (features) related to the region which he/she was born in.
◦ Roughly 2000 instances and 20 features/attributes for the training data
◦ Metadata describing teach feature is included for more insight into the data
◦ Train regression model and use it to predict the life expectancy
◦ Make an Ultimate Judgement, of the best regression model you would choose
– Remember: "The best model (hypothesis) that you can justify"
You will also setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data into training and validation data (manual split is not acceptable).

You need to come up with an approach (that follows the restrictions in 3.2), where each element of the system is justified using data analysis, performance analysis and/or knowledge from relevant literature.

• As one of the aims of the assignment is to become familiar with the machine learning paradigm, you should evaluate couple of different models (only use techniques taught in class up to week 4 - inclusive) to determine which one is most appropriate for this task.

• Setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data.

• Finally, you need to analyse the model and the results from your models using appropriate techniques and establish how adequate your model is to perform the task in real world and discuss limitation if there are any (ultimate judgement).

• Predict the result for the test set.

2.1 Data Set

The data set for this assignment is available on Canvas. It has been modified and pre-processed to some extent, such that all the attributes/features are integers or floats, and missing values has been estimated and filled in.

There are the following files:

- Data-set.csv, contains the entire dataset. You need to divide this data set into training and testing (don't divide the dataset manually), then perform your analysis and tasks on them.
- The file metadata.txt contains some brief description of each of the fields (attribute names).
- The file sample_solution.csv shows the expected format for your predictions on the unseen test data (reminder: test set is the result of randomly dividing your entire dataset into train and test).

2.1.1 Restrictions

As the aim of this assignment is to encourage you to learn to explore different approaches, while you can explore feature impotency, and regularization, your approach must not explicitly perform feature selection. That is, your models should have all features as input (except the "ID" field which is not an attribute).

2.3 Marking guidline

A detailed rubric is attached on canvas. In summary:

• Approach and ultimate judgment 80%

• Report and Presentation structure 20%

Approach: You are required to use a suitable approach to find a predictive model. You may use any ML technique taught in class during week 1-4, including: linear, non- linear and regularization techniques. Each element of the approach need to be justified using data analysis, performance analysis, your analytical argument and/or published work in literature. This assignment isn't just about your code or model, but the thought process behind your work. The elements of your approach may include:

• Performing EDA

• Setting up the evaluation framework

• Selecting models, loss function and optimization procedure.

• Hyper-parameter setting and tuning

• Identify problem specific issues/properties and solutions.

- Analysing model and outputs.

All the elements of your approach should be justified and the justifications should be visible in the PDF version of the notebook (inserted as Markdown text), and your video presentation. The justifications you provide may include:

- How you formulate the problem and the evaluation framework.

- Modelling techniques, you select and why you selected them.

- Parameter settings and other approaches you have tried.

- Limitation and improvements that are required for real-world implantation.

This will allow us to understand your rationale. We encourage you to explore this problem and not just focus on maximising a single performance metric. By the end of your report, we should be convinced that of your ultimate judgement and that you have considered all reasonable aspects in investigating this problem.

Remember that good analysis provides factual statements, evidence and justifications for conclusions that you draw. A statements such as:

“I did xyz because I felt that it was good”

is not analysis. This is an unjustified opinion. Instead, you should aim for statements such as:

“I did xyz because it is more efficient. It is more efficient because . . . ”

Ultimate Judgement & Analysis: You must make an ultimate judgement of the “best” model that you would use and recommend in a real-world setting for this problem. It is up to you to determine the criteria by which you evaluate your model and determine what is means to be “the best model”. You need to provide evidence to support your ultimate judgement and discuss limitation of your approach/ultimate model if there are any in the notebook as Markdown text.

Performance on test set (Unseen data): You must use the model chosen in your ultimate judgement to predict the target for unseen testing data (provided in test data.csv). Your ultimate prediction will be evaluated, and the performance of all of the ultimate judgements will be published.

Implementation

Your implementation needs to be efficient and understandable by the instructor.

Should follow good programming practices.

You must use your the model chosen in your ultimate judgement to predict the TARGET_LifeExpectancy on unseen testing data (which is a result of dividing the dataset into train and test set). Your ultimate prediction will be evaluated, and the performance of all of the ultimate judgements will be published

3 Additional Information

3.1 Getting Started

To help you get started, we suggest the following:

- Load dataset into your Jupyter or your favourite Python IDE
- Do some preliminary data exploration, to understand it better (this will help you later on with trying to figure which regression approach is ideal and how to improve it)
- Setup your data into training and testing datasets
- Select the basic linear regression algorithm and train it then evaluate it
- Analyse the results and see what is going on (to help you determine what needs to be changed to improve the regression model)
- Now you can continue with your method development, discussion and ultimate judgment, etc.

3.2 Sources of Help

Most questions should be asked on Canvas, however, please do not post any code. There is a FAQ, and anything in the FAQ will override what is specified in this specifications, if there is ambiguity.

Your lecturer is happy to discuss questions and your results with you. Please feel free to come talk to us during consultation, or even a quick question, during lecture break.

3.3 Marking Rubric

The rubric is attached on Canvas.

3.4 Submission Instructions

Submission instructions will be placed on Canvas.

3.5 Late Assessment Policy

A penalty of 10% of the maximum mark per day (including weekends) will apply to late assignments up to a maximum of five days or the end of the eligible period for this assignment, whichever occurs first.

Assignments will not be marked after this time.

3.5.1 Extensions and Special Consideration

A penalty of 10% per day is applied to late submissions up to business 5 days, after which you will lose ALL the assignment marks. Extensions will be given only in exceptional cases; refer to the Special Consideration process. Special Considerations given after grades and/or solutions have been released will automatically result in an equivalent assessment in the form of a test, assessing the same knowledge and skills of the assignment (location and time to be arranged by the course coordinator).