RMIT UNIVERSITY

COSC2673/COSC2793 | Semester 1 2020
Machine Learning & Computational Machine Learning

Assignment 1 (v1.0)
Introduction to Machine Learning

Weight: 15% of the final course mark
Due Date: 5.00pm, Friday 3 April 2020 (Week 5)
Learning Outcomes: This assignment contributes to CLOs: 1, 3, 4

Change Log

1.0
- Initial Release

1 Introduction

1.1 Summary

In this assignment you will explore a real dataset to 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:

- Apply regression algorithm(s) to a real-world dataset.
- Analyse the output of the regression algorithm(s).
- Research how to extend the regression techniques that are taught in class.
- Provide an ultimate judgement of the final trained model that you would use in a real-world setting.

This assignment has three deliverables:

1. A report (of no more than 4 pages, plus up to 2 pages for appendices) critically analysing your approach and ultimate judgement.
2. A set of predictions from your ultimate judgement.
3. Your Python scripts or Jupyter notebooks used to perform your analysis.

An AWS educate classroom (name: RMIT_ML_2020S1_Assignment_1) is setup specifically for this assignment.

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 Relevant Lecture/Lab Material

To complete this assignment, you will require skills and knowledge from lecture and lab material for Weeks 1 to 4 (inclusive). 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.

1.4 Plagiarism

Plagiarism is a very serious offence.

The penalty for plagiarised assignments include zero marks for that assignment, or failure for this course. Please keep in mind that RMIT University uses plagiarism detection software to detect plagiarism and that all assignments will be tested using this software. See the RMIT website for more information about the university policies on Plagiarism and Academic Misconduct.

2 Task

Prediction is an important aspect of machine learning. In this assignment, you will predict the life expectancy of a newborn based on several attributes (features) related to the region which he/she was born in. Your task is to use a ***regression approach*** to predict the life expectancy on test and unseen data. You will also setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data into training and testing data.

As one of the aims of the assignment is to become familiar with the machine learning paradigm, you should evaluate multiple different regression algorithms to determine which one is most appropriate for this task.

2.1 Data Set

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

There are the following files:

- `train.csv`, contains the training dataset. Use this for your own exploration and evaluation of which approach you think is "best" for this prediction task.
- `test.csv`, contains the testing dataset. It has all the independent features but not the dependent one (`TARGET_LifeExpectancy`). For your own evaluation, this file may be useful for exploring the features.
- 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.

The original data is from "Global Health Observatory data repository". The data we provided is based on this, with some modifications.

The above files are also available at: `s3://rmit-ml-2020s1-lab-data/Assignment1_data/<filename>`

2.1.1 Restrictions

As the aim of this assignment is to encourage you to learn to explore different approaches, your 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).

> **!** Take note of the above requirement to ensure your model uses *all* features of the data set.

2.2 Regression

You are required to use a regression approach to find a predictive model. You may use any form of regression techniques, including:

- Linear Regression,
- Polynomial Regression
- Regularisation
- Data normalization

A thorough investigation will consider multiple approaches.

2.3 Ultimate Judgement

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".

2.4 Prediction on Unseen Data

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

2.5 Approach, Critical Analysis & Report

Finally, you must compile a report describing and analysing the approach that you have taken to find a suitable model and make your ultimate judgement. Your report ***must*** be no longer that 4 pages, plus an additional 2 pages for appendices. The appendices must only contain figure, diagram, or data tables that provide evidence to support the conclusions and statements in your report.

> **!** This assignment isn't just about your code or model, but the thought process behind your work.

In this report you should describe elements such as:

- Your final selected approach
- Why you selected this approach
- Parameter settings and other regression approaches you have tried.

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 . . . "*

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)

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

All submission for this assignment regardless of extensions must be submitted by 13.59pm Thursday 9 April 2020. Late submissions will not permitted after this date. Extensions through ELP or assessment adjustment that would extend the submission beyond this date will not be granted. Any special consideration that would extend the submission beyond this date will result in an equivalent assessment.

Assignments submitted after 13.59pm Thursday 9 April 2020 will not be marked.

The reason for this policy is that solutions and approaches for this assignment will be discussed in lectures, tutorials and/or labs after this date.