RMIT UNIVERSITY

COSC2673 Machine Learning semester.1 2025

Assignment 1
Introduction to Machine Learning

Weight: 30% of the final course mark

Type: Individual

Due Date: 5.00pm, Monday 4th of April 2025 (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

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. Performing EDA
2. pre-proccessing the data (if needed)
3. Experimenting with different regression approaches and their effectiveness in making predictions
4. Analysing the performance of such methods by choosing a correct metric, and comparing them accordingly
5. Selecting the appropriate ML techniques and applying them to solve a real-world ML problem.
6. Analysing the output of the algorithm(s).
7. 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). You already have the tools for kick starting this assignment. When you start (No later than the start of Week 3) 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 two deliverables:
Please note, you will not receive any mark if you don't submit all 2 assignment deliverables. While you must submit all the 2 deliverables.

1. *Your Jupyter notebook, following bellow criteria*:
   - A fully commented Jupyter notebook with completed mark-down sections, and coding sections, displaying the output of each code section. A template is provided for you in the assignment page. You should not delete or add any section to the template.
   - Bullet point format:

   Your mark down must be:
     - self contained
     - Detailed
     - in bullet point format

   In the bullet point is where you raise each point in one bullet point, using clear topic, then explain the important details ummary under the title (This description is a clear example). A report must be no more than 10 pages.

   - Graphs:

   Your code should create graphs produced and then the graphs should be analysed in your following markdown section.
   - Markdown:

   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 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).

Please note, you will not receive any mark if you don't submit all 2 assignment deliverables.

2 Task

You are working as a data analyst for an AI research organization that aims to predict two key performance metrics of chatbot systems based on various technological and operational indicators:

1. **TARGET_Capacity**: The chatbot's capacity in terms of the length of meaningful conversation.
2. **TARGET_CaseCount**: The number of AI tasks that can be run in parallel.

Your task is to build predictive models (For this assignment TARGET_Capacity and TARGET_CaseCount need separate models), using different regression techniques and evaluate their performance.

Task Details

Students are required to perform bellow task (**note that they are not mentioned in the correct order, and you need to use the knowledge you have learned in the class to figure out the order**) :

1. **For TARGET_Capacity:**
   - **a.** Apply **Linear Regression**
   - **b.** Apply **Polynomial Regression**
   - **c.** Apply **Generalized Linear Models (GLM)**
   - d. Compare their effectiveness in predicting the AI capacity
2. **For TARGET_CaseCount:**
   - a. Apply **Poisson Regression** (since it deals with count data)
   - **b.** Apply **Generalized Linear Models (GLM)**
   - c. Compare their performance in predicting AI task execution
3. **Perform Exploratory Data Analysis (EDA)** before model development.
4. **Perform desired pre-processing techniques**
5. **Ensure a balance** between model complexity and error using techniques discussed in **Lecture 2 and Lecture 3**.
6. **Compare the models**, analyzing their assumptions, limitations, and performance.
7. **Provide a final judgment** on the best-performing models and justify their choices.

Your baseline method should include all the features (no feature selection at this stage)

- For feature selection, you should only rely on your EDA and data analysis. If you decide to remove a feature for your second method, you have to explain your reasoning clearly. Then, you need to compare your method performance before and after the feature removal.
- You are only allowed to use methods taught in Week 1 – 4 (inclusive)
- There are two tasks for each, you need to develop regression models
- If you want to include learning rate in parameter tuning. It is allowed as it was thought in lecture. If you decide to do so. Check the python document to see how you can do that specifically for the regression function that you are using
- Please ensure to check all the previous announcement that help you gain a better understanding of the assignment.

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.csv**, contains the training part of dataset. You need to divide this data set into training and **testing** (don't divide the dataset manually, use the method taught in the lab instead to ensure randomness), then perform your analysis and tasks on them. you will use the **testing set that you have generated(by dividing the Data into training and testing)** here, to evaluate your methods and compare their performances.
- **Eval.csv**, this dataset has no target value. You will only use it to generate predictions based on your best-performing methods. Please ensure to have two methods each for predicting one of the target values. Your methods will be evaluated by us based on your test set error.
- The file metadata.pdf contains a 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).

Dataset Overview (Modified Metadata)

- **ID**: Unique row identifier
- **TARGET_Capacity**: The chatbot's capacity in terms of the length of meaningful conversation
- **TARGET_CaseCount**: The number of AI tasks that can be run in parallel
- **Country**: Country identifier
- **Year**: Year of release

- **Status**: Last updated less than 1 year ago (0 = No, 1 = Yes)
- **RL**: Reported errors
- **AReLM**: Average reported error rate last month
- **AReLW**: Average reported error rate last week
- **SLS**: System Latency Score (measured in milliseconds)
- **Alcohol**: Alcohol consumption per capita (litres)
- **PercentageExpenditure**: AI infrastructure expenditure as a percentage of total research budget
- **Measles**: AI security breach incidents per 1000 user sessions
- **BMI**: Bot Model Integration score (indicating efficiency in integrating multiple AI models)
- **Under5LS**: System uptime reliability over the last five major releases
- **Polio**: AI learning stability metric (%)
- **TotalExpenditure**: General AI development expenditure as a percentage of total budget
- **Diphtheria**: Security patch coverage rate (%)
- **HIV-AIDS**: System vulnerability score
- **GDP**: Global Development Popularity index (indicative of AI adoption trends)
- **Population**: Total number of active chatbot users
- **Thinness1-19years**: AI model efficiency rate in handling user queries aged 10-19 (%)
- **Thinness5-9years**: AI model efficiency rate in handling user queries aged 5-9 (%)
- **IncomeCompositionOfResources**: Funding diversity index for AI research
- **Schooling**: AI training dataset diversity scor

2.3 Marking guidline

A detailed rubric is attached on canvas.

All the elements of your approach should be justified . The justifications you provide may include:

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.

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.

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/

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 skill.