National College of Ireland

Introduction to Artificial Intelligence and Machine Learning

(BSHC3A, BSHC3A, BSHC3B, BSHDS3, BSHCE3)
Release Date: 09/11/2024
Submission Deadline: 02/12/2024 at 23:55

Sumit Tripathi

Harshani Nagahamulla

Administrative Data

This project is an individual Project assignment that is worth 50% of the final grade awarded.
The deadline for submission of this assignment is: 2nd Dec 2024 23:59hrs
The assignment will be marked out of 50%

Extension/Re-run

Should any student miss the assessment deadline with a valid reason, he/she can now apply for an application for coursework Extension/Re-run Form online, via NCI360.

All submissions will be electronically screened for evidence of academic misconduct (plagiarism and collusion).

Project Overview

For the project you are expected to perform the following tasks:

(i) Comparing performance of Logistic Regression and K-Nearest Neighbour (KNN) classification algorithms on a reasonable-sized classification dataset (see Section 5 for different data sources),

(ii) Measuring performance of Linear Regression algorithm on a reasonable-sized regression dataset

(iii) Summarizing characteristics of the dataset chosen for analysis/comparison using descriptive statistics.

(iv) Produce a report that describes your data collection process (if any), data pre-processing methods (e.g. removing noisy instances or variables), data statistics, methodology, evaluation strategy and results.

You should identify as clearly possible the exact topic (related to the chosen topic of interest) relevant to machine learning or artificial intelligence. In other words, you should seek to identify problems (e.g. predicting housing prices, detection of heart disease, spam email detection, identifying anomalous transactions, ...) which can be addressed using relevant machine learning algorithms (classification: K-NN and Logistic Regression, and Linear Regression). You need to identify data sources for collection which will enable a set of repeatable experiments to be carried out. This might be an existing publicly available dataset.

You are required to develop a solution of the problem (of the chosen topic of interest) which you have identified. You should specify data collection strategies when designing and implementing the data driven applications using machine learning techniques and tools.

You should compare the performance of two machine learning methods (Logistic Regression and K-NN) applied to the classification dataset related to the topic. You should measure the performance of the multivariate Linear Regression model on a reasonable-sized regression dataset and present the results. You must use appropriate evaluation techniques and tools in order to measure the performance (in terms of quality) of your methods / models. You should perform a comprehensive manual error analysis on the outputs produced by your models.

Projects will be assessed based on their novelty, technical quality, insightfulness, depth, clarity, robustness, quality of writing and reproducibility. Code and datasets must also be submitted with the report. Algorithms and resources used in a report should be described as completely as possible to allow reproducibility. This includes methodology, empirical evaluations, and results.

Key details, requirements, and definitions

a. **Data Requirements:** The dataset should be for predictive analytics tasks, i.e., it should have a meaningful easily identifiable response variable. It should also be suitably large (at least 10,000 rows).

b. **Deliverables:** There are 2 deliverables for this project:

   i. Final Report (PDF format)

   ii. Source-code and datasets used in the project as a compressed ZIP file

c. **Number of Methods:** In total, you should apply and evaluate two classification methods (Logistic Regression and K-NN) and linear regression method for this project to facilitate your discussion. You should generate descriptive statistics for the data sets you used for your project.

d. **Notions of Performance:** The discussion of performance should be orientated around multiple notions of performance (e.g. RMSE, RSS, Sensitivity/Specificity, Measure).

NOTE: Ensure that your name(s) in full (as per NCI official documents) and student number(s) are clearly visible on the front page.

Project Report

The report must follow the IEEE conference format and should be between 4–5 double column pages in length (this includes all figures and references) (min. 4 pages). For this exercise IEEE style referencing, not Harvard referencing, should be used. Papers over 5 pages will be subjected to a 5% point penalty, i.e., the maximum mark for the paper will be 95%. Microsoft Word and LTEX templates are available at http://www.ieee.org/conferences_events/conferences/publishing/templates.html. The report should include a discussion of the approach, with an emphasis on the critical evaluation of the methods selected. The following structure is suggested for the final report:

Abstract: 100–150 words providing a high-level description of the project, its core findings, and the domain of the datasets (not necessarily in this order).

Introduction: Remainder of 1st page. It should motivate the work, present and discuss the objective(s) of the project and (optionally) provide a concise overview of the following sections (max 1–2 lines per each).

Methodology: This section can be named differently. But it should describe how have you approached to solve the problem. Additional (technical) details can also be discussed here. You should also include here a discussion on key preliminary aspects of the methodology, such as how the datasets have been prepared for study (i.e., the pre-processing).

Results and Discussion: How have you used your method(ology) to your problem (evaluation methodology), i.e., how do you know that a method is good? what performance measures have you selected and why. You should also discuss the results in detail in this section: what are their implications? What do they show / not show? etc. A discussion on sampling methods is expected here too.

Conclusions and Future Work: Summarize your findings, and discuss limitations / extensions that were you to have more time, you would do next to improve / extend your study. Note the key implications of your findings with respect the methods studied.

References: Include a list of references in your report, if used.

2. Potential Data Sources : Possible sources of datasets include, but are not limited to:
    - Statista https://www.statista.com
    - European Data Portal, EU Open Data Portal, and other https://data.europa.eu/
    - UK's open government data repository https://data.gov.uk
    - Central Statistics Office, Ireland https://www.cso.ie
    - Ireland's open government data repository https://data.gov.ie
    - Run My Code https://www.runmycode.org
    - Amazon's public dataset repository https://aws.amazon.com/datasets
    - Google's Public Data Directory https://www.google.com/publicdata/directory
    - The UCI machine learning repository https://archive.ics.uci.edu/ml/
    - Zenodo https://zenodo.org
    - Dublinked https://data.smartdublin.ie
    - Data.gov https://www.data.gov
    - Quandl https://www.quandl.com

3. Marking Grid

   Total Project Weighting: 50% of the final mark. The project of this coursework will be graded using the marking grid shown in Table.

Marking Grid

CRITERIA	HIGH H1	H1	H2.1	H2.2	Pass	Fail
Objectives and Motivation. (15%)	Very challenging project objectives are well presented, met and thoroughly motivated as well as dis- cussed. It is hard to find a fault in the approach.	Challenging project objectives are well presented, met and thoroughly motivated as well as discussed.	Appropriate project objectives are well presented, met and thoroughly motivated as well as dis- cussed. All steps of project are rigorously studied and implemented. Some minor shortcuts or errors may be present.	Appropriate project objectives are presented, mostly met and motivated as well as discussed	There are clear objectives, which are at least partially met.	Cannot discern project objectives, and/or if project objectives were met
Methodology (30%)	All key decisions are appropriately justified. The project extends well beyond simply applying models to datasets, and thoroughly investigates a diverse range of situations to give a very rich under- standing of performance.	All steps of the project are rigorously studied and implemented.	Most key decisions are appropriately justified. The project extends beyond simply applying models to datasets, and makes a good attempt to investigate a range of situations to give a better understanding of performance.	Some key decisions are appropriately justified, but more depth is needed. The project doesn't (or may only arbitrarily) extend beyond simply applying models to datasets; more depth of differentiated evaluation is necessary to provide a better understanding of performance.	All steps of project are appropriately applied, but the general approach lacks depth. There may be significant mistakes in the approach taken.	All steps not appropriately studied and implemented. The approach taken may also be hard to discern.
Results and Discussion. (30%)	All key decisions are appropriately justified. The project extends beyond simply applying models to datasets, and investigates a diverse range of situations to give a rich understanding of performance	All key decisions are appropriately justified. The project extends beyond simply applying models to datasets, and investigates a diverse range of situations to give a rich understanding of performance	Most Key decisions are appropriately justified. The project extends beyond simply applying models to datasets, and makes a good attempt to investigate a range of situations to give a better under- standing of performance	Key decisions are appropriately justified, but more depth is needed. The project extends beyond simply applying models to datasets, and seeks with some success to investigate a range of situations to give a better understanding of performance.	Some key decisions are appropriately justified, but more depth is needed. The project doesn't (or may only arbitrarily) extend beyond simply applying models to datasets; more depth of differentiated evaluation is necessary to provide a better understanding of performance	Key decisions are not properly justified. The project may also lack depth or complexity in several key aspects.
Conclusion and Future Work. (10%)	Insightful conclusions, which appreciate limitations and implications of the project. Implications of	Insightful conclusions, which appreciate limitations and implications of the project. Implications of	Implications and limitations well understood. Discussion also correctly highlights key takeaways.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Future work lacks depth and creativity, but is	Implications and limitations not well understood. Future work lacks depth and creativity, but is appropriate	Implications and limitations not understood. Future work seems arbitrary or inconsistent with

	the project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	the project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	Appropriate future work is discussed and presented.	appropriate.		project findings.
Quality. (15%)	Exceptionally well writ- ten, and presented, with no mistakes in formatting or referencing (if used).	Well written, with no (large) language errors. All figures are well conceived and readable. The IEEE template is adhered to. Report does not exceed the length limits. References (if used) are appropriately and correctly used.	Main document has a few language or style errors. Figures are well presented. IEEE template and length limit are adhered to. References (if used) are complete, and correctly used.	Main document is readable with some language or style errors. Some figures are mostly well presented. IEEE template is largely adhered to. References (if used) are mostly complete and correctly used.	Main document is read- able with some language or style errors. Some figures may be hard to read or presented in a suboptimal manner. IEEE template is largely adhered to. References (if used) are mostly complete and correctly used.	Littered with typos, or poor use of English. IEEE template may have been broken. Figures may be hard to read. References (if used) are probably incomplete.
	80–100	70–79	60–69	50–59	40–49	<40