H9CML: Cloud Machine Learning

Project

THIS DESCRIPTION IS PROVISIONAL. IT IS PENDING REVIEW BOTH INTERNALLY & EXTERNALLY AND MAY BE CHANGED SIGNIFICANTLY!

Submission Deadline: Friday 19th August 2021 23:59hrs

1. Introduction

This project is aimed at achieving the module objectives below:

- **Clean** and **transform** datasets in preparation for data mining, and **build** and **evaluate** machine learning models to extract knowledge from various datasets.
- Evaluate and utilise **cloud computing technologies** and services for data collection, storage and mining when designing and implementing data driven applications.

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.

N.B.

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

2. Project Description

Students are required to propose and execute a research project using **machine learning** techniques as **a team of 3-4 participants**. The project should relate to a real-world problem in any topic. Students should make use of **cloud computing technologies** and services for **data collection, storage and mining**. While doing the project, students should consider the **ethical aspects** with regard to the datasets and machine learning algorithms used.

The overarching focus of the project is to **develop a cloud-based solution of a problem with machine learning methods/techniques**. Your solution should aim at roughly balancing the three paradigms, namely performance, scalability and latency and throughput, of your models. The development pipeline should include the following aspects: data collection process in cloud, transformation strategies applied (locally or in cloud). In order to view data characteristics, you should apply statistical analysis tools. The expectation is that you will be applying a number of ML techniques and models in your experiments, and each method should be applied to **answer-**

specific research question(s) aligned to the overall goal(s) of the project. It is also expected that the application of each method is accompanied by an appropriately sized review documenting pertinent and contemporary approaches in the literature that can both inform the application of a method as well as **justify its potential merit(s)**. You must use appropriate evaluation techniques and tools in order to measure the performance (in terms of quality and latency) of your methods / models. You should perform a comprehensive manual error analysis on the outputs produced by your models and baseline models which have been considered for your investigation.

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 experimental methodology, empirical evaluations, and results. The reproducibility factor will play an important role in the assessment of each submission.

It is essential that the project shows unambiguously evidence of:

1. A critical analysis of fundamental machine learning methodologies to assess best practice guidance when applied to computational problems (e.g. data mining, sentiment analysis, image captioning) in the specific context of the project.
2. The extraction, transformation, exploration, and cleaning of datasets in preparation for building machine learning methods used in the project.
3. Making use of cloud computing technologies and services for data collection, storage and mining.
4. The building and evaluation of machine learning models on a variety of datasets and parameters; producing learning curve of your ML models on a varying sized data sets, data domains and number of parameters.
5. Reporting performance (quality, latency and throughput) and scalability of the machine learning models in varying-sized data sets. The discussion of performance should be orientated around multiple notions of performance and across a variety of parameters (e.g. latency).
6. The extraction, interpretation, and evaluation of information and knowledge that is drawn from the datasets used.
7. The critical review of relevant machine learning research to afford the assessment of research methods applied in the project.
8. A prototype showing facilitating services (e.g. real-time or batch) with your models via cloud to the end-users

3. Deliverables

There are **2 deliverables** for this project:

1. **Report** (PDF format). Ensure that names of all team members in full (as per NCI official documents) and student number are clearly visible on the front page.
2. **Source-code and datasets** used in the project as a compressed ZIP file.

3.1. Report Guidelines

The final report must follow the IEEE conference format and should be **between 4–5 double column pages in length (this includes all figures, but not references)**.

For this report **IEEE referencing style** should be used. Papers with less than 4 or over 5 pages will be subjected to a **5% penalty**, i.e., the maximum mark for the paper will be 95%. Microsoft Word and LaTeX templates are available at http://www.ieee.org/conferences_events/conferences/publishing/templates.html.

The following structure is suggested for the 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 research question(s) / objective(s) of the project and (optionally) provide a concise overview of the following sections (max 1–2 lines per each).

Related Work: Half page -– this should not only summarize the related works, but also **critically evaluate** their key positive and negative aspects with respect to the topic and domain of the project, i.e., how well/badly do the related works artefact address your question(s) / objective(s), what aspects are useful to consider, what are the limitations, etc. Also include here a discussion on the previous uses of the datasets and the methods applied. If you plan to reuse a method already applied to this dataset, discuss what you expect to gain by doing this.

Methodology: This section can be named differently. But it should describe how have you approached answering your question. Additional (technical) details can also be discussed here. Essentially, you should recount how you applied ML models and avail your models in cloud to facilitate your research question(s). 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, and transformation stages).

Evaluation: How have you used your method(ology) to answer the question (evaluation methodology), i.e., how do you know that a method is good?

what performance measures have you selected and why (discuss how the choice of performance measures is appropriate). If you have to parametrize part of an approach how have you done that, and why were these choices made, and what impacts can different parameters have on your results? 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. This section should include a subsection 'error analysis' which will present a thorough and comprehensive error analysis on the outputs of your systems / models including the baselines.

Conclusions and Future Work: Summarize your findings, and discuss limitations / extensions you would do next to improve or extend your study if you had more time. Summarize the (partial) answer to the research question(s) at a high level, and note the key implications of your findings with respect the methods studied.

Contributions: A table should be used in order to show individual contributions in the project. The table should clearly highlight who carried out what part of the project (e.g. design, development, experiments and report preparation).

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

Although not recommended, you may remove, change and/or alter methods, datasets and/or key research question(s) or objective(s) after submission of the Proposal / Interim Progress Report.

4. 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 2. Note that marks in Table 2 represent percentage.

Marking Grid – Project, Cloud Machine Learning **2022**

THE FINAL MARK MUST BE 40% OR ABOVE TO ACHIEVE A PASS

CRITERIA	HIGH H1	H1	H2.1	H2.2	Pass	Fail
Objectives and Motivation. (10%)	Very challenging project objectives are well presented, met and thoroughly motivated as well as discussed.	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 discussed.	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.
Discussion and Related Work. (10%)	Discussion of related work is excellent, and the choice of papers to discuss excellently situates the project within the literature.	Discussion of related work is very good, and the choice of papers to discuss excellently situates the project within the literature.	Discussion of related work is good and the choice of papers to discuss well situates the project within the literature.	Discussion of related work is appropriate and the choice of papers to discuss well situates the project within the literature.	Discussion of related work is appropriate, and the choice of papers appropriately situates the project within the literature.	Discussion of related work lacks depth, or the choice of papers seems somewhat arbitrary.
Choice of Methods. (10%)	The student has studied a selection of complex methods illustrating a well thought out approach to addressing their objective(s).	The student has studied a selection some complex methods illustrating a well thought out approach to addressing their objective(s).	Application of at least two advanced methods.	Application of at least one advanced method.	The student has appropriately selected methods to address their objective(s), but played it safe.	Choice of methods appears arbitrary, or not well justified.
Methodology. (25%)	It is hard to find fault in the approach.	All stages of implementation are rigorously applied.	All stages of implementation are rigorously applied. Some minor shortcuts or errors may be present.	All stages of implementation are appropriately applied, but the general approach lacks some depth. There may be some mistakes in the approach taken.	All stages of implementation are appropriately applied, but the general approach lacks depth. There may be significant mistakes in the approach taken.	methods not appropriately followed or applied. The approach taken may also be hard to discern.
Evaluation. (25%)	All key decisions are justified with appropriate literature. The project extends well beyond simply applying models to complex datasets, and thoroughly investigates a diverse range of situations, parametrizations, and sampling methods to give a very rich understanding of performance.	All key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and investigates a diverse range of situations, parametrizations, and sampling methods to give a rich understanding of performance.	Most key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and makes a good attempt to investigate a range of situations, parametrisations, and sampling methods to give a better understanding of performance.	Key decisions are justified with appropriate literature, 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, parametrisations, and sampling methods to give a better understanding of performance.	Some key decisions are justified with appropriate literature, 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 justified or substantiated with appropriate literature. The project may also lack depth or complexity in several key aspects.
Conclusion and Future Work. (15%)	Insightful conclusions, which appreciate key limitations and implications of the project. Key implications of the project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	Insightful conclusions, which appreciate limitations and implications of the project. Implications of the project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Appropriate future work is discussed and presented.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Future work lacks depth and creativity, but is appropriate.	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 project findings.
Quality. (5%)	Exceptionally well written, and presented, with no mistakes in formatting or referencing. Report provided in PDF format.	Well written, with no major language errors. All figures are well conceived and readable. The IEEE template is adhered to. Report does not exceed the length limits. References are appropriately and correctly used. Report provided in PDF format.	Main document has a few language or style errors. Figures are well presented. IEEE template and length limit are adhered to. References are complete, and correctly used. Report provided in PDF format.	Main document is readable with some language or style errors. Some figures are mostly well presented. IEEE template is largely adhered to. References are mostly complete and correctly used. Report not in PDF format.	Main document is readable 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 are mostly complete and correctly used. Report not in PDF format.	Littered with typos, or poor use of English. IEEE template may have been broken. Figures may be hard to read. References (if any) are probably incomplete. Report not in PDF format.
	80–100	*70–79*	*60–69*	*50–59*	*40–49*	*<40*