ImpCML Coursework Instructions

Three documents make up the full course work for Mathematics for Machine Learning (70015) that we are titling "ImpCML."

(1) - This instruction document
(2) - Coursework marking scheme and specifications
(3) - Kryptonite-*n* Mock Paper and Dataset

This document gives instructions and expectations, the coursework specification goes over formatting details for student write-ups (in conference format), and the prompt is a mock research paper. We describe each here.

1.) Coursework Instructions (This Document)

ImpCML (Imperial Coursework on Machine Learning) is meant to give students some practical experience with the tools we have seen throughout the course. The coursework objective is for groups of 3-4 students to read the Kryptonite-*n* mock paper, explore the proposed challenge dataset with a variety of methods (using valid experimental design), and write a response paper with Abstract, Introduction, Methodology, Experimental Design, Experimental Results, and Discussion sections (with expectations for each outlined below).

The submitted content is comprised of two components: (A) a written report and (B) code to reproduce experiments. Expectations for these components are given below followed by a mark scheme.

1A) Written Report Expectations:

The written report should be between 3-6 pages of content with unlimited additional space for references.
Each report should have the following sections:

- Abstract - A short paragraph describing the main aims of the paper, method investigated for achieving that aim, and briefly mentioning the key results.
- Introduction - Sets a broader stage than the introduction. Spends a paragraph outlining the importance of the problem (in our case challenge datasets), a paragraph describing how the paper seeks to solve the problem, a paragraph giving the reader some details of the solution structure (e.g., what experiments

are carried out) and a paragraph summarizing the main takeaways of the experiments.

- Methodology - The mathematical foundations of any models used. This uses notation similar to what we have seen in class to orient the reader. E.g., if describing gradient descent writing it out formally and pointing out the hyper-parameters you are interested in like learning rate and number of epochs.
- Experimental Design - Typically in research papers this is left to the Appendix, but in this case we are asking students to dedicate a section of their paper to describing a hypothesis they seek to test, the experiment they will run to test this, and how they will ensure their results are valid (e.g., including error bars, different data splits)
- Experimental Results - Students will present the results of their experiments with tables or figures and describe how these plots validate or invalidate their hypothesis. *Importantly: text in figures should be no smaller than font size in the main text.*
- Discussion - Students will use this section to discuss what they were able to do, highlight any experiments they wish they could have done, and provide an environmental impact assessment of the experiments that they ran.

Together, each of these sections make up the full submission for the group's written report.

1A) Code and Labeling Expectations

In addition to the written report students are expected to submit the following artifacts:

- Code to reproduce each of their figures. The code should be readable and interpretable to GTAs and myself. Ideally, students would create a script to reproduce each experiment individually.
- Test labels - The challenge dataset comes with a hidden_kryptonite_n_X.npy which is set of features that does not have a corresponding labels. Students are expected to submit a numpy file: y_predicted_n.npy for each dataset (n corresponds to the feature dimension). Each .npy should load into a binary vector of length 10000 and will be used to evaluate the final accuracy of each reported model. While students may explore several types of models in the context of their report, they are asked to only submit what they believe is their "best" predictions for each dataset.

Written document, GitLab Hash, and zip file containing labels must be submitted no later than November 21

2.) Coursework Specification

As this coursework aims to mimic an academic conference or workshop, we have modified the ICML formatting instructions to fit with this coursework. Students can duplicate the provided Overleaf project in order to produce their write-up using LaTeX.

3.) Mock Paper

The mock paper "Kryptonite-*n*" lays out the perspective of challenge datasets as valid ways to audit machine learning models and puts forward a dataset that it claims proves machine learning is all hype and that the universal function approximation theorem does not really hold in practice. Additionally, this paper provides a github corresponding to the mock paper which contains the challenge datasets, code to reproduce the described experiments e.g., polynomial regression, and the unlabelled grading data.