Machine Learning Ethics Module

Jessica Brown, Ayush Chakraborty

December 2023

Abstract

This document serves as an overview of our ethics module design and implementation, as well as instructions for use. Its intent is to ease the process of integrating the module into the course.

Contents

1 Introduction

1.1 Module Overview

The ethics module will take place at the very beginning of the next iteration of Olin's Machine Learning course. It will provide an ethical preface to the algorithmic and applied topics being covered later in the semester. In particular, this module challenges students to evaluate machine learning models using established evaluation techniques. We wanted to teach them evaluation skills before model implementation skills, because model performance is a prerequisite to understanding the source and extent of model bias. Model bias matters because it remains a pervasive issue in the realm of machine learning ethics.

The module is scheduled to cover three class sessions, or about the first two weeks of the semester, and includes a series of readings, in-class discussions, and a long term homework assignment that assesses each student's individual learning. Based on input from our point instructor Paul Ruvolo, our materials are two fold: first, a high level schedule for the module, and second, a detailed assignment notebook that will be completed over the course of the module. Rather than developing well polished slides for each class, this is what he stated would be most helpful.

2 Learning Objectives

Through this module, students will learn skills that enable them to evaluate models for the duration of the course.

2.1 Objective 1

SWBAT

Students will be able to perform rigorous testing of an existing model to comprehensively understand its performance.

Evaluation

Assessment will be successful completion of the module assignment. The module assignment will consist of ethical content through the analysis of biases, and hands-on content through the implementation of ML evaluation metrics. Therefore, completion of the assignment will indicate:

- The successful implementation of ML metrics
- The successful identification of model bias through the use of those metrics
- That students can describe the relationship between bias in the model output and bias in the input data
- That students have the ability to find a new dataset and manipulate the data in order to run it through the same model for more exhaustive testing

2.2 Objective 2

SWBAT

Students will be able manipulate model performance at an input/output level.

Evaluation

Assessment will be based on students' performance on the assignment tasked for the duration of the module.

Specifically:

- The successful use of a pre-trained model to generate predictions, using a new test dataset
- The ability to understand the nature of the predictions in the context of each model being used

- The ability to manipulate test and training data to influence model output

2.3 Objective 3

SWBAT

Students will better understand the capabilities of ML

Specifically:

- The fact that they can classify images
- The fact that their performance can vary based on input data

Evaluation

Successful completion of the assignment will indicate that these higher level ideas were learned.

3 Course Context

The Machine Learning course at Olin College, last taught in Fall of 2021, intends to "equip students with a multi-faceted and interdisciplinary skill set to understand, implement, and critically evaluate machine learning systems" (Machine Learning 2019 Syllabus). Towards this aim, the course balances developing technical and ethical skills revolving around specific topics in machine learning. There are three primary modules of the course: Neural Networks in the context of Computer Vision, Probabilities/Bayesian Models in the context of Natural Language Processing, and an open-ended final project where students are able to explore topics of interest. The first two modules are set up to achieve a balance between theory, context, and implementation. The theory section involves learning the fundamental mathematics and statistics needed to understand a given algorithm. Context and impact is used to learn the potential implications of ML, especially in regards to ethics, of a given algorithm in a broader context. During interactive activities and implementation is where students get to apply what they learn from the previous two sections. They implement the algorithms in python, using real data. Thanks to this balanced pedagogical approach, there is already a significant amount of ethical content in the course. This means our module is simply an introductory session that prefaces the rest of the course, rather than a holistic module that puts all of course's ethics content in one place.

4 Instructions for Use

Our materials include a module schedule, detailing what to do during each class period. We also created the homework assignment that will be used over the course of the module. What we haven't created are:

- Daily slide decks to use during each of the 3 class days
- The daily in-class activity notebooks- though we have outline of what should be included in each notebook for each of the three days

This means that instructors can use our schedule and assignment materials to guide the creation of lecture materials, slide decks, and the daily in-class activity notebooks. They can use the assignment as is for the module or edit it as they wish.

5 Intended Readings

Here is a list of the readings and resources we intend to use in this module, including in the assignment itself:

- Machine Learning Lifecycle. We felt this would help explain the process people use to evaluate models at a high level.
- Data Quality Intro. We felt this was necessary to include because it speaks to the importance of having quality data when training models. It also describes how to evaluate data quality.
- Google Data Quality. This is an additional data quality article we felt might be useful. It also exposes students to the fact that large companies like Google post educational content in these spaces.
- AI Fairness Checklist. We felt this was necessary to include as it provides a framework for thinking about model fairness and biases, and why its important to mitigate bias to preserve fairness.
- AI Food Assistant. We felt this was important to include so we could introduce people to a well-intentioned, light-hearted application of using ML image classification in the real world.
- Multi-class Classification Metrics. This is a great scientific article about multi-class classification metrics. Not only does it serve to teach students about metrics, but its also a great way to give students experience with reading scientific articles.

- Classification Accuracy. This is just an additional explanation of classification accuracy in the form of a blog post.
- Precision and Recall. We included this to help introduce precision and recall to students. This article is also from Google and introduces students to some educational resources available to them from established companies.

6 Module Outline

Now that we've introduced our module, here is the detailed schedule we have created to go along with it:

This module will take place over the course of 3 in-class periods and an assignment. The schedule for the module is as follows:

6.1 Day 1: Intro to the Blackbox

6.1.1 Agenda

- Discussion: Impacts of machine learning (15 min)
- Lecture: The life cycle of an ML model (15 min)
- In-class activity: Playing with input data (40 min)
- Reflection (15 min)
- Brief course overview and assignment intro (10 min)
- Fill out course entrance survey and ask remaining questions (5 min)

6.1.2 Homework

Read this article about the ML development lifecycle.

6.1.3 Rationale

Our goal was to "win day one" (Paul) by incorporating technical information, contextual information, and hands-on implementation, so that students get the sense that they will have ownership over the course content, and that the content will be relevant to them in the real world.

6.2 Day 2: Model Evaluation

6.2.1 Agenda

- Discussion: Debriefing the homework reading (10 min)
- Lecture: Evaluation Metrics (20 min)
- In-class activity: Interpreting metrics practice (40 min)
- Reflection (15 min)
- Work time: Assignment 1 (25 min)

6.2.2 Homework

Finish through section 1 of assignment 1 (in the A1 notebook). Read these two articles (one, two) on evaluating dataset quality for machine learning models.

6.2.3 Rationale

Our goal here was to delve into evaluation metrics in a scaffolded and systematic way so the learning curve wouldn't be too steep out of the gate. We want to ensure, as its still early in the class, that people feel the course content is accessible and informative at the same time.

6.3 Day 3: Dataset Evaluation

6.3.1 Agenda

- Discussion: How might you use evaluation metrics to assess bias? (15 min)
- Lecture: Dataset splitting (10 min)
- In dataset vs. out of dataset evaluation (10 min)
- In class activity: Playing with datasets and dataset splitting (40 min)
- Reflection (5 min)
- Work time: Assignment 1 (20 min)

6.3.2 Homework

Finish Assignment 1.

6.3.3 Rationale

We wanted to wrap up this module by teaching students about dataset evaluation and the impacts that training data has on models. This is the first chance they have to play around with training data for themselves and see how their choices influence output, as oppose to other people's choices (which is what they get expose to via the assignment).

7 Activity/Discussion

The activity we developed is assignment 1, but we also have more detail about the types of prompts and questions that could be used during the first class day.

For item 1 on the agenda (Discussion: Impacts of machine learning (15 min)), we propose the following questions (many copied from the 2019 iteration of the class). These questions will correspond to the food prompt that will be the theme over the course of the module.

7.1 Day 1 Discussion Prompts

- How are ML systems typically evaluated?
- Are my "classes" truly discrete? Or is there actually a continuum (e.g., blood pressure), where clinical thresholds are in reality just cognitive shortcuts? If so, how far beyond a threshold is the case I'm "classifying" right now?
- Are you using the right metric? When your sample is not evenly distributed (unbalanced data), then always choosing 1 solid accuracy but provides no useful information.
- How diverse and representative is the test data?
- What if the consequences of being wrong is not the same for 1 vs 0? (This would be modified to adhere to the food context)
- What is the model trying to solve?

7.2 Day 3 Discussion Prompts

We believe the following discussion questions would serve as great closing prompts for the module on day 3 (many copied from the previous iteration of the course), for the reflection component of the day 3 schedule:

- How well would your system need to work to generate value for the user?
- Are there types of errors that are permissible or those that would be catastrophic?
- Would you use this model? Would you deploy it? Why or why not?
- If you could collect data to improve the system, what sort of data might you collect as additional training data (remember, machine learning systems can improve based on training data)?

8 Assignment

8.1 Overview

The full assignment can be found in this repository. In summary, we will be providing students with a jupyter notebook, A1.ipynb, that they must complete over the course of the two weeks. The assignment is broken into two parts. In the first part, students will have to implement different machine learning evaluation metrics based on the theoretical overview provided to them during class. The second part of the assignment involves having students pretend to be a ML model auditor, where they are given three different food classification algorithms and need to determine if any are fit for production. The specific prompt is as follows:

8.2 Prompt

Welcome to the first Assignment of Machine Learning! In this assignment, you are the head of the R and D department at Pie-thagoras Labs, a new FoodTech start-up that is aimed at helping members of the visually impaired and blind community in different food-related situations. The first product you are making is an image-based dessert classifier that enables users to obtain detailed information about foods in front of them by simply taking a picture. As a part of developing this product, three of your employees have trained their own classification models (‘model0‘, ‘model1‘, and ‘model2‘)

on different components of the dessert dataset. It is now your responsibility to evaluate each of them. By the end of your analysis, you need to give a thorough assessment of each model, including which (if any) should be integrated into your company's consumer products.

This will require students to use their algorithms from the first part in order to support the conclusions they draw. Additionally, students are provided with the full training and test datasets and will need to make assessments on whether these datasets are sufficient for training and evaluating using qualitative frameworks that are discussed in class.

8.3 Rubric

The rubric for assessment is as follows:

- Depth of analysis- do they reference appropriate accuracy metrics covered in class and make relevant conclusions specific to the situation?
- Do they use relevant data sources to estimate accuracy?
- Clarity of writing (basic grammar and writing conventions)
- Scientific Communication - are their ideas conveyed effectively, with well-developed visuals?
- Code quality. Are their results reproducible? Can we run cells in the notebook and reproduce their results?
- Completion. Did they attempt all components of the assignment? Was anything left unfinished?