CS 475/675 Machine Learning: Homework 1

Due: Monday, February 14, 2022, 11:59 pm

50 Points Total Version 1.0

Make sure to read from start to finish before beginning the assignment.

1 Introduction

The goal of the homework assignments in this course is to learn about how machine learning algorithms work through hands-on experience. This includes both implementation and analysis of various methods.

This semester we will have 5 homework assignments for a total of 400 points. The first and last homework will be worth 50 points, and the other homeworks will be worth 100 points.

There are three different types of work that we will incorporate in the homework assignments.

1. **Analytical:** These questions will ask you to consider questions related to the topics covered by the assignment. You will be able to answer these questions without relying on programming. Many will require mathematical work.

2. **Lab:** In the lab you will utilize Python notebooks (Google colab) to explore data and algorithms, and answer questions. You can work on this as a standalone Python notebook (`https://jupyter.org`) or using Google Colab (`https://colab.research.google.com/`).

3. **Programming:** You will implement a machine learning algorithm in Python and evaluate it on a provided dataset. We will provide detailed instructions to guide your implementation, so behaviors will remain consistent across the class.

Most assignments will focus on either analytical or programming work, in combination with a lab. For the first assignment, you will complete analytical work and a lab.

Each homework will have a master document (this document) that overviews all of the work in the assignment. For analytical work, there will be a separate LaTeXtemplate that contains the actual questions. For labs, we will distribute a Python notebook for you to complete. For programming, we will (often) distribute starter code and data.

Each assignment will contain a version number at the top. While we try to ensure every homework is perfect when we release it, errors will happen. When we correct these, we'll update the version number, post a new PDF and announce the change. Each homework starts at version 1.0.

Let's get started!

2 Collaboration Policy

The course policy is that, unless otherwise specified, all work must be your own. See the course information Google document for more details.

For this assignment, we strongly recommend you work with a partner. You and your partner will make one submission for the two of you on Gradescope (make sure to include your partner when you submit). You and your partner will receive the same grade, so please choose your partner carefully.

You can only work in teams of one or two (not more). Your partner can be anyone from either section (01/02/03/04) or course (475 or 675) provided that both of you are taking the course for credit (not audit). We *highly* recommend that you do every part of the assignment together instead of splitting it up. You can work on the same Overleaf document and think through the questions together. You probably want to work with the same partner for the semester (*only* for assignments where collaboration is allowed) but it is not a requirement.

3 LaTeX

All solutions for the analytical problems must be PDFs compiled from a LaTeXtemplate we will distribute for each assignment. Why learn LaTeX?

1. It is incredibly useful for writing mathematical expressions.

2. It makes references simple.

3. Many academic papers are written in LaTeX.

The list goes on. Additionally, it makes your assignments much easier to read than if you try to scan them in or complete them in Word.

We realize learning LaTeX can be daunting. Fear not. There are many tutorials on the Web to help you learn. We recommend using Overleaf.com. We have provided you with the tex source for this PDF. You **must use the template**. If you do not use this template, we will not grade your assignment. As the semester progresses, you'll no doubt become more familiar with LaTeX, and even begin to appreciate using it.

Be sure to check out this cool LaTeX tool for finding symbols. It uses machine learning! `http://detexify.kirelabs.org/classify.html`

4 Analytical (25 Points)

The analytical questions will be included in a separate template, where you can fill in your answers. Please open the file `homework1_template.tex` and respond to the analytical questions.

5 Lab (25 points)

In this assignment you will be creating a dataset for supervised learning. You will also become familiar with some popular off-the-shelf machine learning tools people use in practice. You will evaluate your dataset using the framework we introduced in class:

1. Is there a well-defined problem?

2. Does an easy solution exist for the problem?
3. Do you have large amounts of high quality data?
4. Can you meaningfully evaluate results? What would the loss function measure?
5. Is using machine learning for this problem justified?

You should create a dataset for a problem where applying machine learning is challenging, *but still possible.* In other words, a supervised machine learning algorithm should be able to generalize from a training set of (x, y) pairs to make predictions for unseen x examples.

Make sure to think through the ethical implications of the data you are collecting[1]. Beyond this course, as future researchers and practitioners of machine learning, you must consider ethical implications of your work. We'll learn more about this over the semester.

5.1 Identifying the data source

You are free to use data from any domain of interest. We provide some examples of sources of data.

- If you are interested in text data, Wikipedia is a great starting place (`https://meta.wikimedia.org/wiki/Datasets`). Let's consider a Wikipedia document as our example x. Then, we may be interested in predicting y, where y is the number of page revisions, the number of authors, the number of page views, the topic of the page, the language the page is written in, etc.
- If you're interested in image data, consider exploring this collection of open image datasets for inspiration: `https://blogs.ntu.edu.sg/openimagecollections/browse/#collections`. Let's consider an image as our example x. Then, we may be interested in predicting y, where y is the year the image was created, the artist who created the image, the medium of the image, etc.
- If you are interested in public policy, consider exploring datasets produced by the US government (`https://www.data.gov/`) and by the Baltimore City government (`https://data.baltimorecity.gov/`). There are a number of directions to take that address social problems.
- If you are interested in health, consider exploring datasets produced by the CDC (`https://www.cdc.gov/nchs/data_access/ftp_data.htm`) or related to COVID-19 (`http://www.socialmediaforpublichealth.org/covid-19/resources/`).
- Also see this repository of structured data (`https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk`). There are many domains and data formats represented.

You may use data from a combination of sources you identify, and you should have a clear idea of the problem you are trying to solve with the data you are collecting.

You may not just download an existing dataset to use! We want you to create a new dataset for supervised learning. Pick something of interest to you, and that others

[1] Not sure how to think through these ethical implications? Start by reading this Medium article: `https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de`

in the class would find interesting. The labels in your data can be automatically derived from the online source (e.g. how many Wikipedia page views?) or manually applied by you (e.g. did I like this song on Spotify?). If you find some features you like for x, you should at least find some other labels to predict for y.

5.2 Creating the dataset

Open the Jupyter notebook for this assignment. This notebook will walk you through defining your problem, creating the dataset, exploring your data, and running some basic machine learning algorithms. There are questions that should be answered inline within the notebook.

You will hand in both the Python notebook, which contains answers to the questions, and the dataset you create.

6 What to Submit

For this assignment you will submit the following items to Gradescope.

1. **Analytical**. Your analytical solutions **must be compiled from LaTeX and uploaded as a PDF**. The writeup should contain all of the answers to the analytical questions asked in the assignment. Make sure to include your name and your partner's name in the writeup PDF and to use the provided LaTeX template for your answers following the distributed template. You will submit this to the assignment called "Homework 1: Analytical".

2. **Lab Python Notebook** You will submit your Python notebook as a PDF by going to File → Export via PDF or File → Export via PDF via LaTeX. Once you download the pdf, open the file to ensure that the plots show up. You will submit this to the assignment called "Homework 1: Lab".

3. **Lab Data** You will submit your data and associated files as a zip file. You will submit this to the assignment called "Homework 1: Lab Data".

You will need to create an account on gradescope.com and signup for this class. The course is `https://www.gradescope.com/courses/362364`. Use entry code `ZR2Z6B`. **You must either use the email account associated with your JHED, or specify your JHED as your student ID.** See this video for instructions on how to upload a homework assignment: `https://www.youtube.com/watch?v=KMPoby5g_nE`.

Seriously, this is important: **You must either use the email account associated with your JHED, or specify your JHED as your student ID.**

7 Questions?

Remember to submit questions about the assignment to the appropriate group on Piazza: `https://piazza.com/jhu/spring2022/601475` with access code `J29F`.