CS 475/675 Machine Learning: Homework 1
Supervised Learning 1
Due: Wednesday, September 22, 2021, 11:59 pm US/Eastern
70 Points Total Version 1.0

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

1 Homeworks

Homeworks will typically contain two parts:

1. **Analytical:** These analytical questions will consider topics from the course. These will include mathematical derivations and analyses. Your answers will be entirely based on written work, i.e. no programming.
2. **Practicum:** In the practicum portion of the assignment, you will apply machine learning concepts to gain experience working with data from different domains. Practicums could involve Python notebooks, applied explorations of topics covered in the class, or programming assignments.

   Click here for the Practicum Google Colab Notebook

The point total for each portion of the homework will be listed in the assignment. Written assignments will be submitted as PDFs. See below for more details about what to submit.

1.1 Collaboration Policy

The course policy is that, *unless otherwise specified*, all work must be your own. See the about page on the course website for more details.

For this assignment, both analytical problems and the practicum is intended to be done with a partner (e.g. teams of two students). However, you may also choose to work on and submit the assignment by yourself. If you choose to 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's information when you submit). You and your partner will receive the same grade, so please choose your partner carefully.

1.2 What to Submit

For this assignment you will submit the following.

1. **Analytical**. You will submit your analytical solutions to Gradescope. **Your writeup 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 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: Supervised Learning 1: Analytical".

2. **Practicum 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: Supervised Learning 1: Practicum".

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

You will need to create an account on gradescope.com and signup for this class. The course is `https://gradescope.com/courses/297335`. Use entry code `GEG23D`. **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`.

1.3 Questions?

Remember to submit questions about the assignment to the appropriate group on Piazza: `piazza.com/jhu/fall2021/cs601475675`.

2 Analytical (35 points)

Please see the accompanying `homework1_template.tex` file for the analytical questions for this assignment. There is space provided in that file for you to type your answers in LaTeXafter each question. **Do not edit the file in any way except to add your answers.** Gradescope assumes that the PDF will exactly match our template except for your solutions.

In addition to completing the analytical questions, your assignment for this homework is to learn LaTeX. All homework writeups must be PDFs compiled from LaTeX. 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 they are written by hand or if you 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 `pdflatex`. It's available for nearly every operating system. 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`

For each homework analytical we will provide you with a LaTeX template. You **must use the template**. The template contains detailed directions about how to use it.

Please open the template to view the analytical questions.

3 Practicum (35 points)

In this assignment you will be exploring a dataset for supervised learning and training a decision tree classifier.
You should find a dataset for a problem suited for supervised machine learning. 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.

3.1 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 or just one source, and you should have a clear idea of the problem you are trying to solve with the data you are collecting.

[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`

3.2 Working with a dataset

Open the Jupyter notebook `homework1_practicum.ipynb`. This notebook will walk you through defining your problem, exploring your data, and training a decision tree. 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.

Click here for the Practicum Google Colab Notebook