Syllabus | Intro to Data Science

Data Science @ William & Mary

Spring 2021

Course Overview

In this course students will learn the fundamentals of data processing and modeling in the context of Data Science. Emphasis will be placed on careful planning and deliberate decision making when working with data and building models. Programming will be done in the Python language and we will be making extensive use of the scikit-learn (https://scikit-learn.org/stable/) collection.

After learning about the basics of having a good Data Pipeline, students will be introduced to a variety of supervised and unsupervised machine-learning techniques including various methods for regression, classification, and clustering. By the end of the course, students are not expected to be an expert on any particular technique, but should exhibit a solid high-level understanding of the goals of each method, be able to determine when a particular type of model is more or less suitable to a real-world problem and, most importantly, demonstrate a keen attention to detail when working with data.

Throughout the course, there will be a very strong emphasis placed on understanding why we are doing what we are doing.

Catalog Number

- DATA 146
- Section 2
- CRN 25117

Pre-/co-requisites

- “Programming for Data Science” (DATA 141 or CSCI 140) or “Computational Problem Solving” (CSCI 141)

Semester

- Spring 2021 (2021-Jan-27 to 2021-May-18)

Location

- This course is taught online (remotely); therefore, it can be completed from anywhere.

Class Times

- Mo/We/Fr 1100–1150

Instructor

- Dr. Tyler W. Davis
- Email: twdavis@wm.edu (mailto:twdavis@wm.edu)
- Physical Office: Center for Geospatial Analysis, Swem Library Rm 213
- Virtual Office: Nooks (https://nooks.in/goto/7cc0KabvZpnfVKyV?pwd=o162Tu)
- Phone: 757-221-6449
- Website: https://ds-wm.github.io (https://ds-wm.github.io)

Office Hours

- Schedule is posted on virtual office website
- *Appointments are welcome outside normally scheduled office hours; please message or email to set up a time.*
- *There may be certain days when office hours will be either canceled or rescheduled; notifications will be sent ahead of time.*

Delivery

- Course will delivered RSOF (https://www.wm.edu/offices/registrar/facstaff/academicscheduling/instr_del_catgs/index.php), which is fully remote and predominantly synchronous, off campus.
- Some aspects of the class (e.g., lectures, lessons, or demonstrations) will be made available as a recording for to watch either before or after class
- Synchronous class sessions **will not** necessarily be recorded; you are responsible for missed content

Final Exam Period

- Tuesday, 18 May 2021 @ 09:00

Minimum Passing Grade

- D-

Communication

- Instant messaging (e.g., Slack (https://slack.com/))
    - For instant delivery of content or for questions that need quick responses.
    - *You will be invited to our Slack using your W&M email.*
- Course materials delivered using Blackboard (https://blackboard.wm.edu/)
    - This will serve as a primary content hub; all other content will be linked from here.
    - *You need access to Blackboard for this class.*
- Video conferencing (e.g, Zoom (https://cwm.zoom.us/) or Nooks (https://nooks.in/goto/7cc0KabvZpnfVKyV?pwd=o162Tu))
    - Video conferencing is for office hours, video chats, and synchronous class meetings where “face-to-face” communication or screen sharing is required.

    - Zoom room links are posted on Blackboard.
    - *Please note that Nooks presently does not support mobile devices or Safari web browser : (*
- Email (twdavis-at-wm-dot-edu (mailto:twdavis@wm.edu))
    - The new snail mail; use this for personal communication or whenever sharing is inappropriate.

Textbook

- There is no textbook for this class.

Course Materials

- Laptop or desktop computer (*required*)

  * *Your computer should have at least 500 MB of free disk space, have at least 8 GB of memory, and run a modern desktop OS (e.g,. PC, Mac, or Linux). You should have access to headphones and a microphone for virtual class meetings; these will minimize the feedback and allow for an easier time with class discussions.*

Course Structure

Most weeks will be organized as follows:

- Monday: Lecture
- Wednesday: Guided Examples
- Friday: In-class exercises

Not all topics are guaranteed to fit nicely into a single week, so some adjustments to this schedule may be made along the way. Additionally, we have a few short weeks due to the distribution of Spring Break days throughout the semester. Aside from these nuances, the general structure of the class is as follows:

- Since we have SO MANY students interested in Data Science, we are running four concurrent sections of DATA 146 this semester!
- Our Monday lectures will be held as a large Zoom meeting with all four sections, and delivered by one of the four faculty members teaching DATA 146 (Dr. Tyler Frazier, Dr. Tyler Davis, Dr. Daniel Vasiliu, Dr. Ron Smith).
- On Wednesdays, your instructor will walk you through some examples where we will learn how to write the code to apply the techniques introduced in the Monday lecture.
- On Fridays, you will be given exercises to work on in class, with the help of your classmates and instructor.

Course Grading Policy

Most weeks you will be given a lab to work on outside of class. These labs will usually be assigned on Mondays and will due the following Sunday at midnight.

The midterm and final projects will be similar to the labs, but more substantial (and cumulative). All assignments will consist of a mixture of conceptual questions, as well as adapting the techniques we have learned to new data sets and reporting/interpreting the results. All assignments will be completed on Blackboard.

Important note: You may only make one submission per assignment! This is to encourage you to carefully double-check your work before you submit.

The grade breakdown is as follows:

- Labs: 60%
    - Assigned on Monday and due the following Sunday by midnight
- Midterm & Final Projects: 30% (15% each)
- Participation: 10%

Below is a Python function to compute your final letter grade, based on your numerical grade.

```
def GetLetterGrade(pct_score):
  """
  Name:     GetLetterGrade
  Inputs:   float, final percentage score (pct_score)
  Outputs:  str, final letter grade
  Features: Returns your letter grade given your percentage score
  """
  if pct_score > 0.91:
    return "A"
  elif pct_score > 0.9:
    return "A-"
  elif pct_score > 0.89:
    return "B+"
  elif pct_score > 0.81:
    return "B"
  elif pct_score > 0.8:
    return "B-"
  elif pct_score > 0.79:
    return "C+"
  elif pct_score > 0.71:
    return "C"
  elif pct_score > 0.7:
    return "C-"
  elif pct_score > 0.69:
    return "D+"
  elif pct_score > 0.61:
    return "D"
  elif pct_score > 0.6:
    return "D-"
  else:
    return "F"
```

Collaboration

Collaboration is both allowed and encouraged! Most of the time in life we are not working in isolation, and it is both wise and efficient to use all resources available to you, including professors, coworkers, other students, the internet, etc.

> The only thing that is off limits is the discussion of specific answers to any graded assignment.

Additionally, although we are conducting this course remotely - you may very well make some lifelong friends along the way, and the chances of this become greater the more you choose to interact with others!

Never underestimate the power of community, collaboration, and cooperation, and don't allow pride, ego, or shyness to interfere with the development of your own creative potential. Always ask for help when you need it, and help others when you are able!

Communication

Our preferred method of communication will be Slack and Piazza (http://piazza.com/wm/spring2021/data146). You will receive a Slack invitation prior to our first class, and I encourage you to make a post and introduce yourself!

I encourage you to post any general questions about course material here first. You are free to use either one or both of these tools, choose whichever works best for you. Just note that you may need to adjust your preferences so that you don't get inundated with notifications every time someone makes a post.

You may ask questions of each other, as well as contact me either publicly or privately. Regular participation (both asking and answering questions) is strongly encouraged.

Resources

The in-class examples will be presented in a Jupyter notebook format (.ipynb files).

When coding in class, I typically use the JupyterHub (https://jupyterhub.wm.edu/); however, you are free to use whatever editor/IDE you prefer (we will discuss several options in class).

Before contacting me with technical issues, I only ask that you test your code on JupyterHub first, as this ensures that you and I will both have the exact same installation, version numbers, etc. (*Note*: last semester, a few students were getting different answers than I was even when running identical code, which ended up being due to a difference in the random number algorithm that the computer was using caused by running an older version of one of the code libraries we were using. This took a bit of time to troubleshoot!).

Some useful links are below:

- Anaconda: https://www.anaconda.com/ (https://www.anaconda.com/)

Course Schedule

This course will be presented in several modules. Expect for each module to span somewhere between 1-3 weeks; this will depend both on the progress of the class as well as the amount of material in the module.

For our first day of class I will review the syllabus and course expectations with you, and give you a brief introduction to the history of Data Science.

Module 0: Python Review

Some preliminary review and practice with Python using JupyterHub.

Module 1: The Data Pipeline

After some preliminary review and practice with Python, we will talk about what it means to have a good Data Pipeline (which will be a recurring theme throughout the semester).

As an example, we will write a program to retrieve data from The Covid Tracking Project (https://covidtracking.com/ (https://covidtracking.com/)), perform some basic preprocessing, and create a plot of selected COVID-19 statistics for states in the US. Once complete, we will be able to get up to date data and produce new output with only a few lines of code, using functions we have written.

Keywords

- pandas
- matplotlib
- DataFrame
- data preprocessing
- data pipeline

Module 2: Data Preprocessing and Descriptive Analysis

Before building any model, it is always a good idea to “get to know” your data. Summary statistics and visualizations are often a good way to go about this. Additionally, it is not uncommon to have to make corrections to your data, either due to inaccuracies/missing data, or to facilitate future modeling steps.

The topics In this module will include some common data preprocessing steps, basic data visualizations, how to handle missing data, and dimensionality reduction techniques such as PCA and tSNE. The topics discussed in this module will be used throughout the remainder of the semester.

Keywords

- scatter plot
- box plot
- histogram
- central tendency

- variability
- percentile
- quantile
- summary statistic
- dimensionality reduction
- Principal Component Analysis (PCA)
- tSNE

Module 3: Intro to Modeling and Model Validation

In this module we will be introduced to modeling via linear regression (ordinary least squares). We will also discuss how to assess things like whether a model is overfit using a procedure known as K-fold cross validation. We will then look at types of regularization, namely Ridge, Lasso, and Elastic Net regression. These processes can, among other things, sometimes help to prevent overfitting. We will also discuss the importance of "feature scaling" in this module.

Keywords

- linear regression
- ordinary least squares
- internal validity
- external validity
- K-fold cross validation
- regularization
- hyperparameter
- Ridge regression
- Lasso regression
- Elastic Net regression

Review and Midterm

The midterm will be completed on Blackboard, similar to our weekly lab assignments but more substantial. We will devote at least one day to review, and you will be given one class period's time off to work on the project.

Module 4: Classification

In this module we will be introduced to several classification methods, all of which are types of supervised learning. While the overall goal of each method is similar, the methods and results can be quite different. We will explore the differences between these methods when applied to the same data sets and discuss some of the pros and cons of each.

Keywords

- Logistic Regression
- K-nearest neighbors
- supervised learning

Module 5: Decision Trees and Random Forests (Classification & Regression)

In this module, we will learn about Decision Trees and Random Forests. Primarily motivated for the purposes of classification; however, we will also see how they can be used for regression.

Keywords

- Decision Tree Classification/Regression
- Random Forest Classification/Regression
- gini impurity
- ensemble models

Module 6: Clustering

In this module we will be introduced to several clustering methods, all of which are types of unsupervised learning. We will also discuss the difference between supervised and unsupervised learning, by contrasting these approaches with the classification methods seen in the previous module.

Keywords

- K-means
- DBSCAN
- Agglomerative Hierarchical Clustering
- unsupervised learning

Module 7: Neural Networks

In our final module we will be introduced to neural networks, starting with the classic multilayer perceptron, and then moving into more state-of-the-art techniques such as convolutional neural networks.

Keywords

- neural network
- hidden layer
- multilayer perceptron (MLP)
- convolutional neural network (CNN)

Review and Final

The final project will be similar in length to the midterm, and you will need to leverage techniques learned throughout the entire semester.

Other Important Dates

- Last day of add/drop: February 5
- Last day to withdraw: March 29

Accommodations

If you need accommodations, **you have a right to have these met**, so it is best to notify the instructor as soon as possible.

It is the policy of William & Mary to accommodate students with disabilities and qualifying diagnosed conditions in accordance with federal and state laws. Any student who feels s/he may need an accommodation based on the impact of a learning, psychiatric, physical or chronic health diagnosis

should be referred to Student Accessibility Services (http://www.wm.edu/offices/deanofstudents/services/studentaccessibilityservices/) staff at 757-221-2509 or at sas@wm.edu (mailto:sas@wm.edu). SAS staff will work with you to determine if accommodations are warranted, and if so, to help you obtain an official letter of accommodation.