Wharton Data Science Academy

Summer 2021, Linda Zhao

Contents

Program overview **2**
- Online synchronized . . . 2
- Methods covered (mostly) . . . 3
- Case studies . . . 4

Course Materials **4**
- Software . . . 4
- Canvas . . . 5
- Lecture notes . . . 5
- Textbooks . . . 5
- Communication . . . 5

Program Policies **5**
- Live sessions . . . 5
- Assignments and Exams . . . 6
- Grade Policy . . . 7

Tentative time table (check periodically for updates) **8**
- Pre-Program Preparation . . . 8
- Week 1 (July 12 - 16) . . . 9
- Week 2 (July 19 - 23) . . . 9
- Week 3 (July 26 - 30) . . . 10

Program overview

Run by the Wharton Analytic Institute, the Wharton Academy of Data Science will bring state-of-the-art machine learning and data science tools to high school students. We aim to stimulate students' curiosity in the fast-moving field of machine learning through this rigorous yet approachable program. Building up statistical foundations together with empirical and critical thinking skills will be the main theme throughout. By the end of the program, students will not only be equipped with essential data science techniques such as data visualization and data wrangling but will also be exposed to modern machine learning methodologies which are all building blocks for today's AI field. Along the way, students will develop a working proficiency with the R language, which is among the most widely used by professional data scientists in both academia and industry.

We believe data science is not just a collection of techniques; it is foremost motivated by real world problems. The data scientist of the 21st century must be able to identify relevant problems, provide sensible analyses, and ultimately communicate their findings in meaningful ways. All learning modules are based on real life case studies.

Participants and Prerequisites: 10th through 12th grade high school students with strong interests in data analytic who are willing to be challenged by a rigorous curriculum similar to that of an advanced Wharton undergraduate course. Students should be at least comfortable with the mathematics and statistics delivered at the high school AP/IB level or equivalent. A laptop or a computer is a must. English proficiency is a must.

Uniqueness: Wharton professors who are data science experts will lead the lectures and will also be available to students outside of class. Students will advance their skills with data from real world cases and will be challenged to articulate their findings with a final project.

Learning Experience: The hands-on learning experience begins the moment students arrive at the canvas website. All lectures will be delivered step-by-step interactively with students. Following lectures, students deepen their skills in small group lab settings, led by the best Penn Ph.D. students and undergraduate students.

Deliverable: Final group project and presentation in Data Science Live (DSL4) on **July 30.**

Better Prepared for College and Life Beyond: Special sessions run by Penn admission officers will provide insight into the college preparation and application process. Our outstanding undergraduate teaching fellows will share their entire experience including both pre and current college life.

Online synchronized

Summer 2021 program will be online but live synchronized via Zoom. Here is our everyday schedule.

During program time:

- 10:00 a.m. – 1:00 p.m. Lectures/discussions with breaks (Linda)
- 1:00 p.m. – 3:00 p.m. Breaks
- 3:00 p.m. – 4:00 p.m. Structured sessions (Linda/Teaching fellows)

- 4:00 p.m. – 5:00 p.m. Student hands on group discussions with teaching fellows

After day program time:

- 7:00 p.m. – 10:00 p.m. One hour to finish off the daily homework

Methods covered (mostly)

Part I: Acquiring, preparing, exploring, understanding and visualizing data

- R/Rstudio/Knitr
- Study design and data acquisition/preparation
- Exploratory Data Analysis (EDA)
- dplyr/ggplot/plotly

Part II: Foundations of probabilities and statistics

- Elements of probabilities
  - Definitions
  - Law of large numbers
  - Normal variables
  - Central limit theorem
- Statistics
  - Confidence intervals
  - Hypothesis tests

Part III: Model-based modeling

- Multiple regression
  - Model specification/interpretation
  - Ordinary Least Squared solution (OLS)
  - Prediction
- High dimension regression
  - Training and testing errors
  - k-fold cross validation
  - LASSO (Penalized regression)
- Logistic regression
  - Model interpretation
  - Entropy loss
  - Classification
- Text mining
  - Vectorize texts
  - Bag of words
  - Word couluds

Part IV: Machine learning

- Tree based methods
  - Single tree

- Bootstrap
- Ensemble methods
  - Bagging
  - Random Forest
- Neural network/Deep learning
  - Input
  - Layers/activations
  - Output
- Image processing

Case studies

Most of the following cases will be covered in lectures/homework.

- COVID-19: Lock-down and Compliance
- Billion dollar Billy Beane
- Discrimination against women in STEM fields?
- Wharton Business Radio Audience Estimation via Amazon Mturk
- Is the Roulette game a fair one?
- How to improve gasoline consumption?
- Actions recommended to reduce crime rates
- Framingham heart disease study: Identifying risk factors
- Would customers leave a good review: Yelp text mining
- Handwriting recognition (image recognition)
- Diabetes/Health care (Predicting Readmission Probability for Diabetes Inpatients to Save Health Care Cost)
- IQ=Success?
- Hunting for important gene expression positions to help out with HIV positive patients

And more!

Course Materials

Software

R: The free and open source statistical computing language R is used through RStudio.

Rstudio: Throughout of the semester, we use the free RStudio, an interface for writing R documents and working with data.

Rmarkdown: We will use RMarkdown for all materials to ensue reproducibility.

Canvas

All materials needed for the program will be available in canvas. Files will be uploaded to Canvas, including datasets, homework, and lecture notes/slides, sample case studies...

Lecture notes

Detailed self-contained lecture notes accompanied with set of slides will be available on Canvas. They are organized by topics and written in reproducible RMarkdown format which combines R codes, visualizations, and narrative text. Students are **required** to read through **slides** before classes.

Textbooks

It is not required, but we suggest you get the following two e-books as references:

- Garrett Grolemund & Hadley Wickham, *R for Data Science*, 2016, O'Reilly. Available freely online.
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Application in R (ISLR)*, Available freely online, First Edition, 2013, Springer New York. (7th printing).

Additional optional readings are recommended for getting familiar with and working with R. These include:

- The R Core Team, *An Introduction to R*, available from CRAN.
- Hadley Wickham, *Advanced R*, available online

Communication

Communication will be through Piazza. Piazza is a useful forum for students to ask and answer questions and to make suggestions. While the teaching team closely monitor activities there, active students participation is highly encouraged. Asking questions by itself is the most important step of learning. It is more important than being able to answer one's question.

Program Policies

Live sessions

Due to Covid19, we have turned the program online. But each session will be a live synchronized one. To maximize our social interactions you are required to show your face through-out the entire live sessions. Contact your teaching fellow if you can not make some sessions in advanced.

- Make sure no noisy background (or mute yourself)

- Cell phone is prohibited

Assignments and Exams

Homework: A few individual/group homework are expected. While we aim to get homework started during the regular time, you will need to budget time to finish off the homework after the regular program time.

Quizzes: A few scheduled short, simple in class quizzes will be given for the purpose of reinforcement learning.

Final Project: The ultimate goal of the program is to prepare/expose students to techniques that are suitable for modern data. The final project is designed so that each of you will bring a problem of your interest to the program. You will need to identify a problem to tackle with a data set that either you collect/extract or find.

This project is done in groups of five members.

- A well-motivated, relevant topic is most desirable.
- Originality, complexity, and challenge will be another plus.

Deliverables:

- **Presentation** at Data Science Live 4 (DSL4)
  - July 30th: **15-minute** group presentation
- **Slides:**
  - July 29th: Submit it to canvas for the presentation
  - August 2: Final submission
- **Report**:
  - August 2: A final group report with details
  - **Maximum of 8 pages**.

Possible data sources:

- Kaggle is a good place to find a data set.
- Google provides public dataset through BigQuery on Google Cloud Platform.
- gapminder
- UCI Machine Learning Repo

Data Science Live (DSL)

DSL is a place that we showcase students projects. DSL 4 will be for you to show the world what a high school student can do to help having a better world. Each group gets a time slot to share the project with the entire class.

To give you a flavor of what can be done, please take a look at some sample projects our students at Penn have done before.

- DSL 1, Spring 2019
- DSL 2, Fall 2019

- DSL 3, Spring 2021 (slides included)
- DSL 4, Summer 2021
- A highschool student's project
  - Katherine, a high school student, did an independent study with Professor Zhao in summer 2020. We have uploaded her compelte case study and her presentation in /Canvas/Sample Projects.

Grade Policy

While the program does not require assigning a grade, we would like to give you a final unofficial grade based on

- Class/Piazza participation
- Homework
- Quizzes
- Final project

Important note: This program is an unaccredited program. The unofficial grades can not be used for college credit.

Tentative time table (check periodically for updates)

It is very important that we are all on the same page. While the syllabus listed below is a tentative one, we will try to follow it. Once again all materials will be available on our canvas site.

Pre-Program Preparation

June 25th to July 10th:

1. Activate Canvas and Piazza:

- Get familiar with Canvas and Piazza
- Be able to download course materials from Files on Canvas
- Read through this syllabus carefully
- Be able to submit your work to Canvas
- Ask questions or comments on Piazza

2. Install R/RStudio/RMarkdown following our tutorial and pre-recorded video:

- A basic tutorial: **Get_staRted.html**, **Get_staRted.pdf**, and Get_staRted.rmd (the source file of Get_staRted.html).
  - File location: *Files/Module 0 Get Ready* on Canvas
  - Instructions on installing R/RStudio/RMarkdown
  - Some basic R functions
- Pre-recorded video to go through **Get_staRted.rmd**
  - File location: *Files/Video/Basic R tutorial.mp4* on Canvas
  - A video that our previous TA shows steps by steps instructions

3. Your homework 0:

- Part I:
  - Get familiar with the R/RStudio and Rmarkdown
  - Be able to execute codes in Get_staRted.rmd in R-studio
  - Go through chunk by chunk to study R-commend. You are not required to memorize most of them at all. Rather, try to understand what each chunk does.
  - (Optional): study `dplyt/ggplot` in **advanced_R_tutorial.Rmd**
- Part II:
  - **Due 11pm on July 10**
  - In Section **Your Summary** at the end of Get_staRted.rmd, summarize what you have learned in a short paragraph (no more than 100 words)
  - Compile the source code Get_staRted.rmd into Get_staRted.html
  - Turn in your compiled html to Canvas under Homework 0.

Week 1 (July 12 - 16)

Day 1: Data acquisition, preparation, exploration and visualization

- R/Rstudio/Knitr
- Study design and data acquisition/preparation
- Exploratory Data Analysis (EDA)
- Packages: `dplyr/ggplot/plotly`

Day 2: Data Preparation and EDA

- Machine learning workflow
- Data wrangling
- dplyr/ggplot

Day 3: Elements of probability and statistics

- Events, probability
- Law of large numbers
- Central limit theorem
- Confidence intervals/Hypothesis tests

Day 4: Simple regression

- Linear equation revision
- Graphical perspective of OLS
- Confidence interval/Hypothesis testing
- R^2, RSS, goodness of fit
- Model diagnoses

Day 5: Multiple Regression

- Interpretation
- Categorical variables

Week 2 (July 19 - 23)

Day 6: LASSO

- K-fold cross validation
- Training/Testing errors

Day 7-8: Logistic Regression/Classification

- Model and estimation
- Classifier
- Criterion: MCE
- Classification with LASSO

Day 9-10: Trees/Random Forest

- Single trees
- Ensemble methods
- Random Forest

Week 3 (July 26 - 30)

Day 11: Text-mining

- Vectorize texts
- Bag of words/n-gram
- Words clouds

Day 12-13: Image processing/Neural Network

- Elements of neural network
- Digitize images
- Implementation of Neural network

Day 14: Leeway

Day 15: DSL final projects presentation