BIG DATA, MACHINE LEARNING, AND THEIR REAL WORLD APPLICATIONS

Jonathan Wayne Korn

2021-04-12

Scheduled Meeting Dates:

- Session 1: June 28th, 2021 - July 16th, 2021
- Session 2: July 20th, 2021 - August 6th, 2021

Scheduled Meeting Days and Times:

- M-F @ 9:10am - 11:00am *(Morning Session)*
- M-F @ 1:10pm - 3:00pm *(Afternoon Session)*

Instructor: Jonathan Wayne Korn

Email: "jonathankorn71@yahoo.com"

Virtual office hours: Friday Evenings from 4pm - 6pm @ Zoom

Classroom: "virtual"

Response Policy: Please email me if you need to reach me outside of class. We can always schedule time to have a virtual chat if it is needed. I will usually respond to email within 24 hours. Typically, I am open 24/7 by email during the work week *(M-F)*.

Facilitator/Teaching Assistant: *(TBD)*

Virtual office hours: *(TBD)* "Virtual"

Course Overview

This course introduces students with an overview of big data, machine learning, and their real-world applications. Students focus on the strategic use of data and the innovative technologies utilized to derive actionable business insights. Develop the *"Analytical Mindset"* and start thinking data-centric to solve real-world problems. Students are introduced to R and Python programming language to import/export, manipulate, transform, and visualize data. Also, learn to perform basic data analytics such as statistical summaries. The main focus of the course is to develop students to train and evaluate machine learning models for structured classification, regression, time series and clustering problems and unstructured classification problems. Learn to handle model with both structured and unstructured data using various classical machine learning algorithms and deep learning algorithms. Students will be also introduced to the idea of distributed machine learning systems.

Learning Objectives

At the end of this course students will receive working knowledge in:

1. an introduction into data, analytics, and the research methodology.
2. an introduction into R and Python programming language with basics data analytics.
3. an introduction to machine learning for structured classification, regression, time series, and cluster modeling.
4. an introduction to deep learning for structured and unstructured classification modeling.
5. Conduct/Compile/Communicate information about machine learning within teams, across organizations, and with external stakeholders.

Texts and Programs

1. R, RStudio
2. Distribution of Python 3+: Install Anaconda,
   - Use the R *reticulate* package to knit Rmarkdown reports including python chunks.

Resources

Columbia University Information Technology CUIT provides Columbia University students, faculty, and staff with central computing and communications services. Students, faculty, and staff may access University-provided discounted software downloads

Columbia University Library Columbia's extensive library system ranks in the top five academic libraries in the nation, with many of its services and resources available online.

Program Resources If you do not understand the course content or the instructor's expectations, please either speak up during class or contact the instructor outside of class times.

For other program- and wellness-related needs, contact the Pre-College Program office at hsp-office@columbia.edu or (212)-634-2799.

Course Requirements

- ***Assignment #1:*** *R and Python Programming*
- ***Assignment #2:*** *Processing Data*
- ***Assignment #3:*** *Exploring the Data Visually*
- ***Assignment #4:*** *Structured and Unstructured Modeling*

Note:

- Compile the rmarkdown file labeled "assign.#.rmd.", @ https://github.com/jkorn81/cu-hsp-learning with your responses.

- Make sure to answer all of the questions in the assignment(s) before submitting them.
- All assignments must be knit and submitted in either the set format of html_document, word_document, or pdf_document formatting. *(you are able to change the file format that is generated from the revising the yaml in the rmarkdown file.)*

Course Policies

Participation and Attendance

You are expected to complete all assigned readings, attend all class sessions, and engage with others in online discussions. Your participation will require that you answer questions, defend your point of view, and challenge the points of view of others. If you need to miss a class for any reason, please discuss the absence with me in advance. Class attendance is mandatory. Any disruptive behavior will not be tolerated.

Class Etiquette

To ensure the learning environment is optimal, all students should adhere to the following "netiquette" principles during the online class:

- Log into Zoom in enough time to get set up and ready to commence when the class begins. Test your audio and ensure there are no technical problems.
- Participate in the class from a quiet location with minimal distractions.
- Be visible via your webcam during the entire class, and dress in classroom-appropriate attire.
- Actively participate via mic, online polling, responding in chat, etc.
- Be prepared by completing readings and offline activities.
- Communicate with all fellow students and the instructor respectfully; share perspectives and relevant examples.

Late Work

All assignments should be submitted by the due date noted in the course syllabus. Late submissions require advance notice and permission from the instructor.

Citation & Submission

All written assignments must cite sources using [citation format] and be submitted to the course website (not via email). Plagiarism, whether intentional or unintentional, will result in dismissal from the program.

School and Program Policies

Student Assessment

Columbia's Precollege Programs for High School Students are academically rigorous; they do not carry college credit, however, nor are they graded. Upon successful completion of the program, students receive an official Columbia University Certification of Participation and written evaluations from their instructors.

Students are evaluated on the basis of the effort they put in, their progress over the duration of the class, and their potential for future work in the pertinent field and in college.

Successful participation is determined by the instructors in consultation with program administration. Attendance, class participation, satisfactory completion of assignments and adherence to the program's community standards are all considered as part of the evaluation process.

Class attendance is carefully monitored. Students must attend all classes unless they are ill. A student who misses multiple class sessions may not receive a Certification of Participation even if those absences are excused.

Copyright Policy

Please note-Due to copyright restrictions, online access to this material is limited to instructors and students currently registered for this course. Please be advised that by clicking the link to the electronic materials in this course, you have read and accept the following:

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted materials. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.

Academic Integrity

Columbia University takes matters of intellectual integrity very seriously. Plagiarism is not tolerated. Plagiarism includes, but is not limited to, submitting work done by another person or purchased from any source; failure to document ideas found in sources, whether print or electronic, with appropriate notes and bibliographical references; failure to enclose borrowed phrases or sentences within quotation marks; and turning in the same assignment for two courses without advance permission from both teachers. Plagiarism, whether intentional or unintentional, will result in dismissal from the program. Students who are unsure about the proper presentation of their work should consult their course instructor.

Class Participation

Class attendance is mandatory. A student who misses multiple class sessions will not receive a Certification of Participation, even if those absences are excused. Unexcused absences can lead to dismissal from the program.

Students are expected to engage seriously in their courses through both class participation and completion of assigned work. Disruptive behavior will not be tolerated.

Community Standards

The Community Standards designed to ensure the safety and well-being of the students and the integrity of the University. They are strictly enforced and failure to abide by them results in dismissal from the program, normally on the first offense.

The determination as to whether a student has violated the Community Standards is made by program staff, instructors, and administrators.

Students who are dismissed from the program do not receive evaluation letters or Certifications of Participation. No portion of the program cost will be refunded to a student who has been dismissed.

Accessibility

Columbia is committed to providing equal access to qualified students with documented disabilities. A student's disability status and reasonable accommodations are individually determined based upon disability documentation and related information gathered through the intake process. For more information regarding this service, please visit the University's Health Services website.

Class Schedule

Here's a best guess at the course schedule - the assignments will not change, but lecture material may be moved around depending on course speed. The course material will be posted, along with all notes and scripts created for you to use.

Week 01, 06/28 - 07/04
Day 1

Monday* (Pre-Session Readings and Assignments)*:

- Participate in the Getting Acquainted discussion forum before the Class Session.
  - Summarize your current thoughts of machine learning for us in a sentence or two. *(... be prepared to discuss in class.)*

Monday* (morning session ~ 9:10am to 11:00am)*:

- Introduce instructor and summarize the course.
- Students introductions. Tell us who you are, your background in machine learning, any specific interests that motivated you into joining the program, etc.
- Review the Syllabus

Monday* (afternoon session ~ 1:10pm to 3:00pm)*:

- Introduction to big data discussion.
  - Structured *(Quantitative, Qualitative)*

Monday* (Post-Session Readings and Assignments)*:

- Search the internet for potential quantitative and qualitative data sources *(include a list of at least 5 data sources, make sure to record the URL so we can locate them later on.)*.
  - Note, in class we discussed a few options of data sources.

Week 01, 06/28 - 07/04
Day 2

Tuesday* (Pre-Session Readings and Assignments)*:

- Access the course github repo @ cu-hsp-learning to connect to the data source will be using for class. The folder is labeled `data`.
  - Clone the repo and store the root somewhere easy to access. *(I suggest storing it on your desktop for the duration of the course.)*

Tuesday* (morning session ~ 9:10am to 11:00am)*:

- Continued... Introduction to big data discussion.
  - Unstructured *(Text, Image, Video, Speech, etc.)*

***Tuesday* (afternoon session ~ 1:10pm to 3:00pm)**

- Introduction to machine learning discussion.
  - Supervised vs Unsupervised Learning.
  - Classification, Regression, Time Series and Clustering Problems.
  - Classical Machine Learning *(Linear Models, Decision Trees, Random Forest, Naive Bayes, Support Vector Machines, etc.) (sturctured data focused)*
  - Deep Learning *(Tensorflow and Keras in Python and R) (image, text, and structured data focused)*

Tuesday* (Post-Session Readings and Assignments)*:

- Watch the short video @ The future of Machine Learning and its Impact on Your Everyday Life *(3 minutes 55 seconds)* and write down some ideas of how you could use machine learning in your everyday life. *(... be prepared to discuss in class.)*

Week 01, 06/28 - 07/04
Day 3

Wednesday* (Pre-Session Readings and Assignments)*:

- Access the supporting script(s)/file(s) in the following github repo you cloned yesterday @ cu-hsp-learning and review the documents labeled `prepare.tools.guidelines`.
- Watch the following video @ R or Python: Which Should You Learn in 2020? *(19 minutes)* and learn about the iconic battle between R and Python programming languages for dominance in data science. *(Which language you think will be the winner?) (… be prepared to discuss in class.)*

Wednesday* (morning session ~ 9:10am to 11:00am)*:

- R and Python programming discussion.
- Preparing the Tools *(R and Python, Anaconda, Github)*
  - Importing Data:
    - Data Objects *(local/global variables, lists, vectors, matrices, dataframes)*
    - File Types *(cvs, xlsx, SAV, etc.)*
    - APIs *(Discussion on Connecting to API)* at least one project will require to access an api *(i.e. quantmod ~ query stock data)*

Wednesday* (afternoon session ~ 1:10pm to 3:00pm)*:

- Introduction to Processing Data in R and Python ~ *Structured*
  - Subset Variables
  - Data Type Conversions
  - Imputing Missing/NA Values
  - Imputing Outliers
  - Data Normalization Techniques
  - Optional Balancing of the Data *(Randomize Sampling, Automated NoiseFilters)*

Wednesday* (Post-Session Readings and Assignments)*:

- ***Assignment #1:*** *R and Python Programming (… be prepared to discuss in class any pitfalls you may have encountered.)*
  - Access Assignment #1 documentation @ folder labeled assignment1

Week 01, 06/28 - 07/04
Day 4

Thursday* (Pre-Session Readings and Assignments)*:

- Watch the following video on youtube @ https://www.youtube.com/watch?v=d4gGtcobq8M and *(... be prepared to discuss in class your thoughts on some use cases for natural language processing and its importance to machine learning.)*

Thursday* (morning session ~ 9:10am to 11:00am)*:

- Continued... Introduction to Processing Data ~ Unstructured Data
  - Discuss Natural Language Processing Techniques for Text

Thursday* (afternoon session ~ 1:10pm to 3:00pm)*:

- Continued... Introduction to Processing Data ~ Unstructured Data
  - Discuss Image Processing Techniques

Thursday* (Post-Session Readings and Assignments)*:

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Assignment #2: *Processing Data (... be prepared to discuss in class any pitfalls you may have encountered.)*

```
- Access Assignment #2 documentation @ folder labeled [assignment2](https://github.com
```

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com

Week 01, 06/28 - 07/04

Day 5

Friday* (Pre-Session Readings and Assignments)*:

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.
  - Access '06_exploring_data'.
    - Review the code in the script contained in the folder.
      - Do you see any patterns in the code that are recognizable? *(... be prepared to discuss in class.)*

Friday* (morning session ~ 9:10am to 11:00am)*:

- Overview of Exploring the Data using Basic Data Analysis Techniques.

Friday* (afternoon session ~ 1:10pm to 3:00pm)*:

- Continued... Overview of Exploring the Data using Basic Data Analysis Techniques.
  - Lets create some plots and store them in a particular folder in /.jpg or /.png file format. *(We can use the plots later on to compile into a report.)*

Friday* (Post-Session Readings and Assignments)*:

- ***Assignment #3:*** *Exploring the Data Visually*: *(... be prepared to share with the class Monday.)*
  - Access Assignment #3 documentation @ folder labeled assignment3

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com

Week 02, 07/05 - 07/11
Day 6

Monday **(Pre-Session Readings and Assignments)***:*

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Monday **(morning session ~ 9:10am to 11:00am)***:*

- Application and Reporting of Classical Machine Learning Algorithms for Supervised Problems using Structured Data
  - Architect and train supervised classification and regression *(time series)* models in R and Python.
    - Evaluate the performance of the trained model states.
    - Discuss Use Cases for Potential Machine Learning Project Ideas.

Monday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- Application and Reporting of Classical Machine Learning Algorithms for Unsupervised Problems using Structured Data
  - Architect and train unsupervised classification models in R and Python.
    - Evaluate the performance of the trained model states.
    - Discuss Use Cases for Potential Machine Learning Project Ideas.

Monday **(Post-Session Readings and Assignments)***:*

- Prepare for tomorrow as discussed in class.

Week 02, 07/05 - 07/11
Day 7

Tuesday **(Pre-Session Readings and Assignments)***:*

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Tuesday **(morning session ~ 9:10am to 11:00am)***:*

- Application and Reporting of Deep Learning Algorithms for Supervised Problems using Structured Data
  - Structured Data Applications

Tuesday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- Application and Reporting of Deep Learning Algorithms for Supervised Problems using Unstructured Data
  - Text Applications

Tuesday **(Post-Session Readings and Assignments)***:*

- Prepare for tomorrow as discussed in class.

Week 02, 07/05 - 07/11
Day 8

Wednesday **(Pre-Session Readings and Assignments)***:*

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Wednesday **(morning session ~ 9:10am to 11:00am)***:*

- *Continue. . .* Application and Reporting of Deep Learning Algorithms for Supervised Problems using Unstructured Data
  - Image Applications

Wednesday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- Application and Reporting of Distributed Machine Learning and Deep Learning Algorithms for Supervised Problems using Structured and Unstructured Data
  - Discuss the potential of using structured and unstructured data for distributed supervised machine/deep learning systems.

Wednesday **(Post-Session Readings and Assignments)***:*

- ***Assignment #4:*** *Structured and Unstructured.* Attempt to compile the code in the `assignments` folder @ cu-hsp-learning. *(. . . be prepared to discuss in class any pitfalls you may have encountered.)*

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com

Week 02, 07/05 - 07/11
Day 9

Thursday **(Pre-Session Readings and Assignments)***:*

- Prepare for class by individually writing down ideas for projects that can be conducted using some of the machine learning and deep learning techniques we discussed so far. *(... at least think of 3 solutions that can be developed)*

Thursday **(morning session ~ 9:10am to 11:00am)***:*

- Class Project Discussion and Commissioning of Individual Group Components to contribute to a distributed machine learning system.
- Discuss research and application ideas as a whole:
  - The problem should consist of multiple components that require machine learning to assist in the decision process.
  - For instance, building a distributed machine learning system that consists of two layers to decide whether to buy and sell an asset.
  - One layer could be to predict the classification of the next days weather conditions and the second layer to predict the next day's asset price from the previous day.
  - Two machine learning models are required at minimum to support such a system. We will focus on conducting the research for the individual components on the project idea that the class agrees on.
  - We will divide into groups who will perform the research on the selected components.
  - Each group will present their results and the class as a whole will discuss the work.
- Review the project documentation contained in the course github repo @ https://github.com/jkorn81/cu-hsp-learning in the folder labeled `10_project`.

Thursday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- Reporting Techniques:
  - *(Github)*
    - Create an account.
    - Create a repository *(repo)* for your course work and load any file you would like.
    - We will discuss best practices for managing your repos.

Thursday **(Post-Session Readings and Assignments)***:*

- Keep writing down ideas that you'd like to pitch to the class and associate those ideas with a neat project name.

Week 02, 07/05 - 07/11

Day 10

Friday **(Pre-Session Readings and Assignments)*:***

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Friday **(morning session ~ 9:10am to 11:00am)*:***

- *Continue. . .* Reporting Techniques
  - *(Rmarkdown)*
    - ∗ Create Rmarkdown generated documents in pdf, html, word format.
    - ∗ Create Rmarkdown presentations in pdf, html (ioslides, slidy), or powerpoint format.

Friday **(afternoon session ~ 1:10pm to 3:00pm)*:***

- *Continue. . .* Reporting Techniques
  - *(tfruns and Jupyter Notebooks)*
    - ∗ Create reports in Jupyter Notebooks.
    - ∗ Use tfruns to automate reporting.

Friday **(Post-Session Readings and Assignments)*:***

- Create a presentation using one of the reporting techniques discussed in class. *(I suggest making a presentation containing your project ideas.)*

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com

Week 03, 07/12 - 07/18
Day 11

Monday **(Pre-Session Readings and Assignments)***:*

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Monday **(morning session ~ 9:10am to 11:00am)***:*

- Final Commission of the project and initial construction of the project components.
- Discuss the initial steps to conducting machine learning research in a group.
  - Step 1: Discuss the component of machine learning research that your group was tasked to complete.
  - Step 2: Set a plan for the research for an individual in your group to complete.
  - Step 3: Distirbute engineering tasks for portions of the code. *(I suggest breaking into subgroups within your group so that a pair can work on one problem together.)*
  - Step 4: Create a repo in one of the groups githubs and provide access to all group members. *(A cloud based enviroment is important to utilize in remote based work. We will use the repo to manage pushes in your groups code and also track issues throughout the development. Think of it as home base for your group to share files.)*
  - Step 5: Each group will divide and conquer to build a component of the distributed machine learning system as comissioned by the class in a majority vote.

Monday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- Discuss the importance of keep track of the packages/modules being utilized in the development.
- *Continue. . .* Constructing project components in groups.

Monday **(Post-Session Readings and Assignments)***:*

- Work on your group project components. *(Make sure the tasks are divided equally among the group members.)*

Week 03, 07/12 - 07/18

Day 12

Tuesday **(Pre-Session Readings and Assignments)***:*

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.
  - Work on your group project components.

Tuesday **(morning session ~ 9:10am to 11:00am)***:*

- Discuss working on the ReadME.md file on the group repos so that the components of the project can be easily explained to external parties.
- *Continue...* Constructing project components in groups.

Tuesday **(afternoon session ~ 1:10pm to 3:00pm)***:*

- *Continue...* Constructing project components in groups.
  - Each group provide a quick presentation constructed in an rmarkdown to update the class on the group's progress.

Tuesday **(Post-Session Readings and Assignments)***:*

- Prepare for tomorrow as discussed in class.

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com

Week 03, 07/12 - 07/18
Day 13

Wednesday* (Pre-Session Readings and Assignments)*:

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Wednesday* (morning session ~ 9:10am to 11:00am)*:

- Compile the Group work into final presentations.

Wednesday* (afternoon session ~ 1:10pm to 3:00pm)*:

- *Continue. . .* Compile the Group work into presentations.

Wednesday* (Post-Session Readings and Assignments)*:

- Prepare for tomorrow as discussed in class.

Week 03, 07/12 - 07/18
Day 14

Thursday **(Pre-Session Readings and Assignments)*:***

- Access the supporting scripts in the following github repo you cloned yesterday @ cu-hsp-learning.

Thursday **(morning session ~ 9:10am to 11:00am)*:***

- Group presentations.

Thursday **(afternoon session ~ 1:10pm to 3:00pm)*:***

- *Continue. . .* Group presentations.
- Determine if each groups components can be utilized to create a reliable distributed machine learning system and justifies further development.

Adapted from The Course Syllabus, A Learning-Centered Approach, 2nd Edition, Judith Grunert O'Brien, Barbara J. Millis, Margaret W. Cohen. ISBN 978-0-470-60549-3. Available as an E-Book from Wiley at https://www.wiley.com