STA 5934-0003: Statistical Data Science with Python

Fall 2021

Instructor:	Chao Huang	**Time:**	MoWeFr 12:00PM - 12:50PM
Email:	chuang7@fsu.edu	**Place:**	OSB 0327

Course Pages:

1. All the documents can be found in Canvas
2. Some codes and datasets can be found on my GitHub page

Office Hours: After class, or by appointment, or post your questions in Canvas.

Main References: There is no required text for the course; however, lecture notes will be regularly distributed. Some interesting and useful books will be touched during the course, which are listed here for your information:

- Trevor Hastie, Robert Tibshirani, Daniela Witten, and Gareth James, *An Introduction to Statistical Learning: With Applications in R*, New York: springer, 2013. http://faculty.marshall.usc.edu/gareth-james/ISL/
- Peter Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, 2012. http://people.cs.bris.ac.uk/~flach/mlbook/
- Peter Bruce and Andrew Bruce, *Practical statistics for data scientists: 50 essential concepts*, "O'Reilly Media, Inc.", 2017. https://github.com/mbeveridge/Bruce_Practical-Statistics
- Jake VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, "O'Reilly Media, Inc.", 2017. https://jakevdp.github.io/PythonDataScienceHandbook/
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016. https://www.deeplearningbook.org/

Course Description: This course is primarily designed for graduate students as an introduction of statistical learning models used by data scientists for prediction and inference. Also, this course is designed to help students get prepared for both research work and job interviews, where students with different motivations will all get benefits from this course.

In this course, we emphasize the tools useful for tackling modern-day data analysis problems. Many of these are essential statistical learning tools (e.g., regression model, sampling techniques, model regularization, unsupervised learning, and supervised learning), but we also include techniques at the cutting-edge of technology for handling big-data problems (e.g., deep learning and reinforcement learning).

Since the Python programming language is a very popular and powerful platform for scientific and statistical analysis and visualization, it is introduced and used throughout the course, where all the methods, examples, and projects are developed in Python. In particular, several important Python packages (e.g., Numpy, Scipy, Pandas, Matplotlib, and Scikit-Learn) and the parallel computing technique in Python are both covered.

Prerequisites: An undergraduate-level understanding of probability and statistics is assumed.

Tentative Topics:

- Elementary probability and statistics (brief overview)
- Python: basics (installation, program's flow control (conditionals, loops), variables and functions)
- Python: important packages (Numpy, Scipy, Pandas, Matplotlib, and Scikit-Learn)
- Python: parallel computing (multiprocessing)
- Statistical learning: regression model (Linear, GLM, Cox, Nonparametric, etc.)
- Statistical learning: model regularization (Dimension reduction, PCA, ICA, Ridge, and Lasso)
- Statistical learning: sampling & model assessment (Bootstrap, MCMC, and Cross validation)
- Statistical learning: unsupervised learning (K-means, Hierarchical clustering, and Mixture model)
- Statistical learning: supervised learning I (Logistic, Naive Bayes, LDA, and SVM)
- Statistical learning: supervised learning II (Decision trees, Random forests, and XGBoost)
- Deep learning: Convolutional neuronal network (CNN)
- Deep learning: Tensorflow & Pytorch
- Deep learning: Autoencoder (AE)
- Deep learning: Recurrent neural network (RNN)
- Deep learning: Generative adversarial network (GAN)
- Reinforcement learning

Tentative Course Outline:

Week #01 ..... Elementary probability and statistics
Week #02 ..... Basics in Python
Week #03 ..... Important Python packages
Week #04 ..... Regression model
Week #05 ..... Dimension reduction
Week #06 ..... Sampling techniques
Week #07 ..... unsupervised learning
Week #08 ..... supervised learning - I
Week #09 ..... supervised learning - II
Week #10 ..... Neuronal networks & CNN
Week #11 ..... Tensorflow & Pytorch
Week #12 ..... AE & RNN & GAN
Week #13 ..... Reinforcement learning
Week #14 ..... Final project presentation

Grading Policy: It will be determined by a weighted average of the following items: three mini-projects (3*15%=45%) and one final project (30% (report) + 25% (presentation) = 55%). Final grades may be adjusted. However, you are guaranteed the following:

- If your final score is 90–100, your letter grade will be at least A-;
- If your final score is 75–89, your letter grade will be at least B-;
- If your final score is 60–74, your letter grade will be at least C-;
- If your final score is 50–59, your letter grade will be at least D-;
- If your final score is below 50, your letter grade may be an F.

Important Dates:

Mini project #1	Week #04
Mini project #2	Week #09
Mini project #3	Week #13
Final project presentation	Week #15
Final project report submission	Week #16

Course Policy:

- **Classroom policies:** The classroom environment is an important factor for effective learning. In order to not distract other students' attention please follow these classroom policies. The first one of these is the university policy. Remember that no food or drinks are allowed in the classroom. Turn off all audible alarms (cell phones, pagers, calculators, watches etc.) Do not use cell phones in the class. Come to the class on time. Opening and closing the classroom door in the middle of a class cause distraction to the students and the teacher. Do not talk to other students without permission while the professor is teaching. More than one conversation creates noise and makes it difficult for the students to pay attention to the lecture.

- **Attendance:** You are required to attend all classes. The class activities will help you assimilate the lessons more easily, giving you an opportunity for active learning. Do not let this opportunity slip away. Any foreseen absence must be cleared with the instructor. If the absence is due to emergencies, it is the student's responsibility to notify the instructor at the earliest opportunity of the emergency.

- **In-class discussions:** There will be in-class group discussions. The class will be divided into small groups. One problem will be given in class and each group should present their solution after a discussion (about 15 min). The in-class discussions will be graded.

- **Mini projects:** There will be three mini projects covering different topics including Python programming, unsupervised and/or supervised learning, and deep learning. The student is required to work on the mini project independently, and a project report is expected with the python code attached. The Python code should be reproducible and some comments or help documents are needed for explanations.

- **Final project:** There will be one final project selected from the *kaggle machine learning and data science competitions* (`https://www.kaggle.com/competitions`). It will be released at the beginning of this semester. The problems in this project are quite open, and any statistical learning or deep learning tools are welcome. Students will be divided into several small groups (typically three people each group) and work on the project together. Each group needs to given an in-class presentation to describe their methods and results at the end of this semester, and some suggestions will be given after the presentation. A final project report and the Python code needs to be submitted one week later. The detailed contribution for each group member should be mentioned in the report. The Python code should be reproducible and some comments or help documents are needed for explanations.

- **Contacting the instructor or TA outside the class:** You are strongly encouraged to come to the instructor or TA during their office hours. If your schedule conflicts with the office hours, you can make an appointment. You may ask the instructor or TA brief questions by e-mail, but you may be asked to come to office hours if the instructor or TA thinks that the questions are better answered in person. When you send e-mails, remember the following: always send e-mails from your FSU accounts. The e-mails from non-FSU accounts may not reach me due to filters. Always write your full name at the end of each e-mail message you send.

- **Academic honor policy:** The Florida State University Academic Honor Policy outlines the University expectations for the integrity of students' academic work, the procedures for resolving alleged violations of those expectations, and the rights and responsibilities of students and faculty members throughout the process. Students are responsible for reading the Academic Honor Policy and for living

up to their pledge to "... be honest and truthful and ... [to] strive for personal and institutional integrity at Florida State University." (Florida State University Academic Honor Policy)

- **Students with disabilities:** Students with disabilities in need of academic accommodation should: 1. must register with and provide documentation to the Office of Accessibility Services (OAS); 2. must provide a letter from OAS to the instructor indicating the need for accommodation and what type; 3. should communicate with the instructor, as needed, to discuss recommended accommodations. A request for a meeting may be initiated by the student or the instructor. This should be done during the first week of class. See https://dos.fsu.edu/sdrc/ for more information.