MIS S381N 71649/71650/71655 Data Analytics Programming

Schedule: M T W Th { 9:30–11:30 am in GSB 2.120 ; 2:00–4:00 pm in GSB 3.104 }

Instructor: Deepayan Chakrabarti

Office: CBA 6.462

Email: deepayan.chakrabarti@mccombs.utexas.edu

Phone: 232 – 5716

Teaching Assistant: Rohit Arora (arorarohit@utexas.edu)

Office hours: By appointment

Pre-requisites: None

Course Overview

Should I lend to this borrower? Can I detect fraudulent credit card transactions? What are the main types of complaints of my customers? Did my new website design significantly change sales?

Data-driven analysis has wrought a quiet revolution in business. As disk storage and computing power have become cheaper, companies have started maintaining detailed logs of inventories, sales, and customer activity, among others. Yet, this is only half the job; the real need is for *insights*, and this course teaches you the tools for that.

We will learn data analysis in Python, a general-purpose language that lies at the intersection of (a) easy enough to learn, (b) fast enough to scale, and (c) endowed with a wide range of powerful libraries that make data cleaning, visualization, and many common data analysis tasks a cinch. The course is split into five parts:
(1) **Introductory Python**, where we learn the basic language syntax, and gain familiarity with general-purpose tools such as string manipulation,
(2) **Pandas**, which is a powerful data analysis toolkit (similar to R) that makes it easy to explore and visualize data,
(3) **Classification**, where we develop an understanding of how to make predictions,
(4) **Clustering**, where we learn how to discover the major groups or components of a given dataset, and
(5) **Other Topics**, including regression and hypothesis testing.

Course Materials

Books. There are no required books for this course. The following are optional.

For the first two parts of the course, we will use *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, by Wes McKinney. For the remainder, we will use material from a variety of sources, one of which is *Applied Predictive Modeling*, by Max Kuhn and Kjell Johnson. For introductory Python, an additional good reference is *Think Python*, by Downey, available here.

Software. Python, including the following packages:

- IPython
- Numpy
- Pandas
- Matplotlib
- Scikit-learn (sometimes called "sklearn")
- Statsmodels

You can use the Anaconda or Canopy distributions of Python which makes installing packages quite easy.

Grading Policy

The course grade will be calculated as follows.

Work item	Weight
3 Group Assignments	$3 \times 15 = 45$
1 Group Project	25
Final exam	30

Forming groups. I will randomly create groups of 3 students for each assignment, and groups of 4 students for the project.

Homework assignments. There will be three homework assignments. All assignments must be handed in electronically before the beginning of the class (submit online on Canvas).

Project. You must develop a group project on any relevant topic that interests you. The project report is due on 08/12, and each group must present their work on 08/12 and 08/13 (order decided by lottery).

Exams. There will be a final exam. You will need to bring your laptop. The exam is open-book, open-notes, open-slides, and open-web.

Statement on Students with Disabilities

Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259, http://www.utexas.edu/diversity/ddce/ssd/.

Religious Holy Days

By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Policy on Scholastic Dishonesty

The McCombs School of Business has no tolerance for acts of scholastic dishonesty. The responsibilities of both students and faculty with regard to scholastic dishonesty are described in detail in the BBA Program's Statement on Scholastic Dishonesty at http://www.mccombs.utexas.edu/BBA/Code-of-Ethics.aspx. By teaching this course, I have agreed to observe all faculty responsibilities described in that document. By enrolling in this class, you have agreed to observe all student responsibilities described in that document. If the application of the Statement on Scholastic Dishonesty to this class or its assignments is unclear in any way, it is your responsibility to ask me for clarification. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since dishonesty harms the individual, all students, the integrity of the University, and the value of our academic brand, policies on scholastic dishonesty will be strictly enforced. You should refer to the Student Judicial Services website at http://deanofstudents.utexas.edu/sjs/ to access the official University policies and procedures on scholastic dishonesty as well as further elaboration on what constitutes scholastic dishonesty.

Campus Safety

Please note the following recommendations regarding emergency evacuation, provided by the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety:

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement

requires exiting and assembling outside.

- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation should inform the instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors.
- Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- Behavior Concerns Advice Line (BCAL): 512-232-5050
- Further information regarding emergency evacuation routes and emergency procedures can be found at: http://www.utexas.edu/emergency.

Table 1: *Tentative schedule*

Date	Topic	Details
Python refresher		
07/15	Introduction and Python I	Values and variables, control flow, and functions (Think Python chapters 2, 3, and 5)
07/16	Python II and III	Data Structures (Think Python chapters 10, 11, and 12) Data structures detailed example
07/17	Python III (contd.) and IV	Strings and regular expressions Files (Think Python chapters 8 and 14) **First assignment released**
Using Pandas		
07/18	Series and DataFrames	McKinney chapters 5 and 6 Examples using the NYC Complaints dataset
07/22	DataFrames (contd.) and Data wrangling	Merging, concatenation Reshaping, pivoting (McKinney chapter 7)
07/23	Visualization	Plotting Histograms (McKinney chapter 8)
07/24	Grouping data	McKinney chapter 9 **First assignment due** **Second assignment released**

Continued on next page

Table 1 – continued from previous page

Date	Topic	Details
07/25	Time Series and Statistics	McKinney chapter 10 Means and standard deviations Medians and quantiles Correlations
Regression and Classification		
07/29	Regression	R-square Degrees of freedom Relation to correlation
07/30	Regression (contd.) and review	Multiple regression
07/31	Intro to classification	Train/test split Holdout set, cross-validation Accuracy measures **Second assignment due** **Third assignment released**
08/01	Nearest Neighbors	KD-trees (optional)
08/05	Naive Bayes	Probability Conditional Probability The Naive Bayes Algorithm
08/06	Logistic Regression	
08/07	Decision Trees and Ensembles	Basic methodology Information gain and Entropy **Third assignment due**
08/08	K-Means clustering	Distance metric Examples Selecting the number of clusters

Continued on next page

Table 1 – continued from previous page

Date	Topic	Details
Project and Finals		
08/12	Project Presentations I	
08/13	Project Presentations II	
08/14	Review Session	
08/15	(left free for office hours)	