CSCI 381/780 Data Analytics -- logistics

Course Description:

- Data science has been one of the fastest growing professions recently. The goal of the Data Analytics class is to prepare students with the necessary skill set and understanding to succeed in this area.
- The first part of the course will go over fundamental concepts spanning across statistics, cross validation, data visualization, data warehousing and python as data manipulation and model development platform.
- Second part of the course will cover common machine learning techniques such as linear and logistics regression, support vector machine, decisions trees and natural language processing.
- Students at the end of the course should be able to carry on to more advanced studies on machine learning and start a career in the data science areas.

CSCI 381/780 Data Analytics -- logistics

Instructor: Dr. Alex Pang

Email: chiuyan.pang@qc.cuny.edu

Lectures: Mon, Wed (8:00pm – 9:15pm)

Pre-requisites:

- CSCI 313 (Data Structures)
- Math 241 (Prob & Stat)

Teaching Assistant: None

Office hours: 9:15 to 9:45 pm after class

Course Objective:

At the end of this course students should

1. have a good overview of the data science professions and modern data analytics platforms.

2. have acquired expertise in using Python as his/her data analysis and model development platform

3. have developed a good analytical mindset in drawing insights on data and making recommendations

4. have understood some of the most common machine learning techniques and feel comfortable in pursuing more advanced skill set in machine learning areas.

CSCI 381/780 Data Analytics -- logistics

Textbook:

Data Analytics Made Accessible:

2021 edition by Anil Maheshwari

Acknowledgement:

I would like to express my special thanks to Dr. Anil Maheshwari for writing such as wonderful textbook in this exciting field as well as his generosity in sharing some of his PowerPoint slides related to his textbook, some of which have been adapted into our course materials

Optional Textbooks for 381:

- Python for Data Analysis by Wes McKinney
- An Introduction to Statistical Learning by Gareth James, Daniel Witten, et al (http://www-bcf.usc.edu/~gareth/ISL/)

Almost required textbook for 780:

- Hands-On Machine Learning with Scikit-Learn and Tensor Flow by Aurelien Geron

CSCI 381/780 Data Analytics Syllabus – Communication

Communication and Class Participation

- Communication is mainly through the following channels
  - Announcement on Blackboard
  - Discussion Forum on Blackboard
  - Direct emails

- Students are expected and highly encouraged to participate in the discussion forum on Blackboard
- Lectures will be given synchronously on Zoom and will be recorded.
- Reading Assignments will be announced and are expected to complete asynchronously

CSCI 381/780 Data Analytics -- logistics

Section 1: Core Business Intelligence and Data Analytics Concepts

1. Data Science / Data Analytics Overview
2. Probability and Statistics Review
3. Exploratory Data Analysis
4. Data Visualization
5. Machine Learning Overview, Linear Regression and Common Data Scientist’s Toolbox

Section 2: Popular Data Mining Techniques

6. Classification and Naïve Bayes Algorithm
7. Classification and Logistics Regression
8. Support Vector Machine
9. Decision Trees
10. Clustering
11. Text Mining
12. Big Data

CSCI 381/780 Data Analytics -- logistics

Section 3: More Advanced Techniques and Application

13. Neural Network and Deep Learning
14. Business Intelligence Applications
15. Amazon AWS, Microsoft Azure

CSCI 381/780 Data Analytics -- logistics

Grade contribution:

- 30% Homework assignments (3 HWs)
- 20% mid-term exam
- 20% final exam
- 25% final Project
- 5% Review Quizzes & Class Participation (Blackboard)

- Some questions may be mandatory for graduate students but optional for undergraduates

Homework format:

- Python 3 Notebook

Exam format:

- Multiple choices and written short answers

Review Quizzes:

- Multiple choices and written short answers

Minimum to pass the course:

65% raw score

CSCI 381/780 Data Analytics -- logistics

Final Course grades may be curved so that the median grade is between B- and C+

CSCI 381/780 Data Analytics - Class Participation

Class Participation is important

I don’t want my class to be like this

I want my class to be like this

There is never such thing as dumb question

CSCI 381/780 Data Analytics -- logistics

Collaboration Policy:

You are allowed and encouraged to discuss homework. Discussion on Blackboard is encouraged, so everyone can benefit; however, do NOT post or share solutions or parts of solutions. Homework and final project must be done and written up independently.

Academic Integrity Policy:

Absentees are solely responsible for catching-up. Academic dishonesty, such as plagiarism or cheating - taking other people's work with or without their permission in order to get credit for yourself, will be dealt with seriously, including an "F" grade for the course and/or disciplinary action according to the University's policy on academic integrity

CSCI 381/780 Data Analytics -- logistics

Recording Consent for online session

Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live.

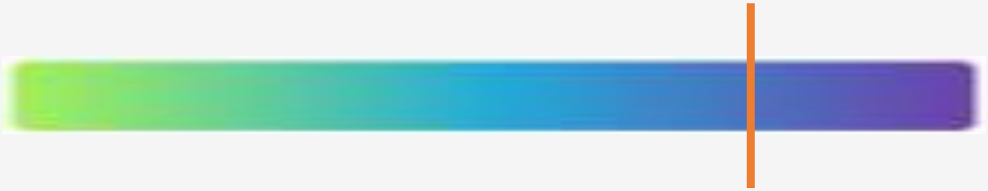
CSCI 381/780 Data Analytics – on the subject of cheating

- It is not a lie if you believe it
- You are not cheating if you are not caught
- I am just trying to help my friends, no harm no foul
- I know the materials, just not have enough time for working on the HW
- …..

- You think your professors are dumb?
- When you work full-time after school, can you call your friends to see if you can copy-and-paste from him or her or look for course material from previous semester?
- What’s your goal in signing up for a class? A fake GPA or learn some real skills
- You will be graded based on the whole class

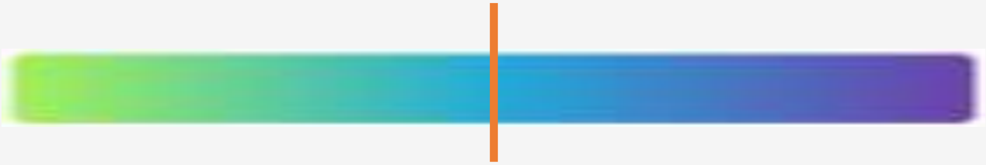
CSCI 381/780 Data Analytics -- logistics

Teaching Style:



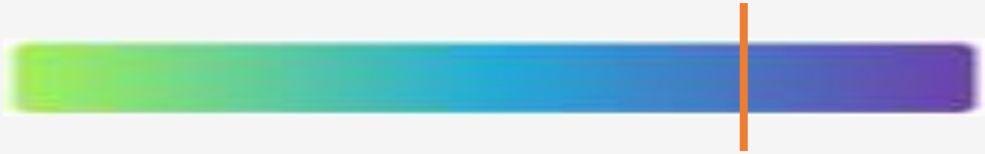
Theory

Practical



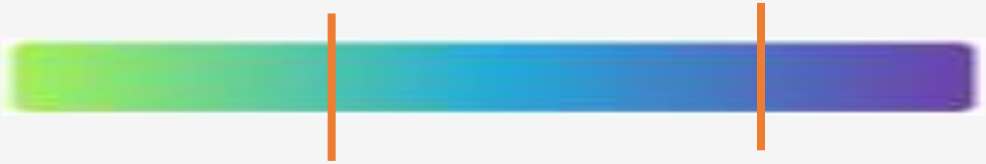
No homework

20 hours per week



Dry
Only me talking

Stand-up comedy
Highly Interactive



follow textbook

No textbook

- I do NOT shy away from using various free resources from the Internet. Remember DRY principle.
- When you want to learn something, where and how do you start?

CSCI 381/780 Data Analytics -- logistics

Weekly Routine:

- Monday:
  - Theory, bi-weekly homework due (if any)
  - Finished reading assignments from the textbook

- Wednesday:
  - Application, homework description,
  - Weekly summary, next week preview, assign reading from the textbook

- Saturday
  - Informal online office hours, will announce before hand,
  - Email is the best way to communicate

- Will post the PowerPoint after class on Blackboard

CSCI 381/780 Data Analytics -- logistics

Python:

- Lectures as well as homework will be based on Python 3 notebooks
- You need to make sure you have a PC or laptop where you can run Python notebooks
- Recommended distribution and installation is Anaconda (https://www.anaconda.com/)
- Make sure you are familiar with the syntax as well as the Pandas and NumPy library for data manipulations
- The textbook has a chapter on Python !