Ask CDSS: Course Selection

November 29, 2017

Overview

- Computer Science = COMS
  - Electrical Engineering and Computer Science = EECS
- Statistics = STAT
- Industrial Engineering and Operations Research = IEOR
- Your Questions!
  - This presentation isn't an end-all say about which classes you should take
  - Definitely explore in other departments (Operations Research, Applied Math, Mathematics, Economics, Quantitative Methods in the Social Sciences, etc.)
  - Feel free to message me (Ashutosh Nanda) with your opinions so that we can grow this presentation

{Computing in Context, Intro to Python (for SEAS)} = COMS {1002, 1006}

- Programming Language for Course: Python
- Will help you learn Python
- Computing in Context is more domain specific (various domains like political science and economics)
- Intro to Python will be closer to scientific computing (usage of `numpy`, `matplotlib`, etc.)
- Homework is programming focused

{Intro to Java, Honors Intro to Java} = COMS {1004, 1007}

- Programming Language for Course: Java
- Will help you learn Java
- Honors isn't really worth it…
    - The design practices covered are kind of out of date
- Homework is a mix of theory and programming
    - You will end up learning how to read Java documentation = a good skill to have!

Data Structures in Java = COMS 3134

- Programming Language for Course: Java
- Basics of data structures and algorithms
- Not really necessary for data science work, but helps build comfort with programming
- Homework is a mix of:
    - Theory: describe an algorithm, do this step of an algorithm manually
    - Programming: implementation of algorithms covered in class

Advanced Programming = COMS 3157

- Programming Language for Course: C/C++
- Will make you a really **advanced** programmer
    - Gets you comfortable with debugging code
    - Helps you learn using the terminal
    - Teaches you how to use Git
- Certain skills (terminal, Git, etc.) are key to being a productive data scientist
- Homework is essentially all programming
    - This class will take a lot of your time!

Introduction to Databases = COMS 4111

- Programming Language for Course: SQL, Python
- Covers both database theory and SQL queries
- Homework is a few projects

Natural Language Processing = COMS 4705

- Kathy McKeown!
    - Director of Data Science Institute
    - Has done lots of research in the field
- Useful and fundamental class because text is a huge and messy part of data science!

Computer Vision = COMS 4731

- Shree Nayar!
    - Great professor
- Good way to go into traditional image processing
    - Convolutions
    - Edge Detection
- Not particularly relevant to data science
    - but lots of people say it's a good class!

Machine Learning = COMS 4771

- New professor
    - …but was a postdoc with David Blei
- A good first-time machine learning course
    - Need to be comfortable with probability + linear algebra + (multivariate) calculus
    - Covers a lot of basic ML techniques (but not too much depth in any particular topic)
- Homework is a mix of theoretical derivations and implementation of ML algorithms

Causal Inference = COMS 4995 (section 4)

- Adam Kelleher: chief data scientist at Buzzfeed, relatively new at teaching, smart but disorganized
- Class material is interesting and useful, mostly theoretical
- Not much math

Ubiquitious Genomics = COMS 6998 (Section 1)

Advanced Topics Projects in Deep Learning = COMS 6998 (Section 3)

- Iddo Drori
  - Professor isn't a deep learning researcher
- Fast moving field
- A project that weights about 50% of your grade

Topics in Learning Theory = COMS 6998 (Section 4)

- Daniel Hsu
  - Great at his content, not an amazingly nurturing professor

Algorithms Geometric Lens = COMS 6998 (Section 5)

Bandits Reinforcement Learning = COMS 6998 (Section 6)

Fundamentals of Speech Recognition = COMS 6998 (Section 7)

IoT - Intelligent & Connected Systems = EECS 4764

Bayesian Models for Machine Learning = EECS 6720

- John Paisley!
    - Great professor
- Covers very modern and powerful ML techniques
- Homework is a mix of theoretical derivations and implementation of the derived algorithms
- Lectures consist of heavily theoretical derivations

Big Data Analytics = EECS 6893

- Hands-on experience with practical big data tools (Hadoop, Spark, etc.)
- Professor does research on graph computing at IBM

{Intro to Statistical Reasoning, Introduction to Statistics, Calc-Based Intro to Statistics} = STAT {1001, 1101, 1201}

- Take Calc-Based Intro to Statistics!
    - Need to get comfortable with some math when doing data science
- Good for someone who has never taken a statistics class before

Introduction to Probability & Statistics = STAT 4001

- This is what the CS majors are supposed to take
    - You'd be better served by taking the next two classes in sequence
    - Or take the IEOR two classes in sequence
- Homeworks are all theory

Probability Theory = STAT {4,5}203

- Will get you comfortable with how to work with probability distributions
    - Take this class if you want to take ML
- Homeworks are all theory

Statistical Inference = STAT {4,5}204

- Builds the theoretical foundation behind hypothesis testing and confidence intervals
    - Not the most modern techniques
- Homeworks are all theory

Probability for Engineers = IEOR 3658 fall

- Antonius Dieker-good lecturer
- Homeworks are all theory, relatively hard, but serves you well
- Open to SEAS undergrad

Statistics and Data Analysis =IEOR 4307 spring

- Antonius Dieker-good lecturer
- Homeworks are all theory, relatively hard, but serves you well
- A group project uses R shiny app to analyze a data set
- Open to SEAS undergrad

Linear Regression Models = STAT {4,5}205

- Entire course about the first model you'll usually try on a dataset
- Helps introduce some of the concepts that you'll cover in Machine Learning courses
- Homeworks are mix of programming and theory

Statistical Computation & Intro to Data Science = STAT {4,5}206

- Meant to be a primer to Statistical Machine Learning (StatML, same kind of content as Computer Science's Machine Learning)
- Mainly just getting comfortable with programming in R (covering R syntax, loops, functions, data frames, etc while doing stat problems)
- Class is very long (all morning Friday and bimonthly lab)
- Programming Language for Course: R

Bayesian Statistics = STAT {4,5}224

- Covers Bayesian version of probability + statistical inference
- Some advanced topics could be covered depending on time: Gibbs sampler, Markov Chain Monte Carlo (MCMC)
- Good primer for higher level Bayesian courses (Bayesian Models for Machine Learning, etc.)

Sample Surveys = STAT {4,5}234

- More of a traditional statistics topic
  - How to design + analyze a representative survey?
  - How to correct for the biases you find?

Advanced Data Analysis = STAT {4,5}291

- Covers a variety of miscellaneous statistical techniques
- According to Chris Mulligan (helped found CDSS): "Don't do it"
    - Mostly revolves around a group project
    - If you've taken Probability, Statistical Inference, and Linear Regression, you're fine
- Textbook is good: "The Statistical Sleuth"
    - Nice "What do I do now?" approach to traditional statistical inference and regression

Statistical Machine Learning = STAT ????

- Not currently on syllabus
- Has been taught by new professors for the last two semesters
- Another good introduction to ML class like COMS 4771
    - Same caveats about comfort with probability, linear algebra, and calculus apply
    - Essentially identical list of topics

Advanced Machine Learning = STAT 5242

- New class: assumes you have taken Statistical Machine Learning
- John Cunningham is a great professor!
    - Previously taught StatML

Foundations of Graphical Models = STAT 6701

- David Blei
    - Turing Award Winner in 2013
- Starts from fundamental material in probability and graph theory but progresses quickly
- Similar content as Bayesian Models for Machine Learning
- Ph.D. Level class
    - Homework is part of the class, but the major focus is on the end of term project

Translational Bioinformatics = BINF 4006

- No bio background needed
- Lectures consist of discussing papers and learning about bio and how data science applies
- Most of grade is based on a semester-long project (most people use R)
- Pretty good class for people who have little knowledge about bioinformatics and want to learn more on the applied side

Intro Data Science Industry = APMA 4990

- Dorian Goldman relatively new professor
  - ok-ish lecturer
- Covers machine learning in python
- Group projects

Data Analytics for OR = IEOR 4523 (former IEOR4572)

- Great course to work with packages in python for analytics
- Hardeep Johar / Yair Avgar
- Some homework requires quick self-taught html, sql etc.
- Use python, covers Obtain data from files (csv, html, json, xml) and databases (Mysql, PostgreSQL, NoSQL), machine learning packages,brief overview of natural language processing, network analysis, and big data tools available in Python
- A group project analyzing data set of your own choice, deliver in flask (optional)

Data Mining for Engineers = IEOR 4540

- Krzysztof M Choromanski former google researcher
- projects, midterm, final
- present some fundamental techniques used in data mining and machine
- learning (dimensionality reduction mechanisms, neural networks, etc)
- Does not assume any machine learning background knowledge all the things will be explained along the way

Analytics on the Cloud = IEOR 4574

- Hardeep Johar (great professor! explains things well)
- Focus on Spark
  - Cloud essentials (unix basics, hadoop, map/reduce)
  - Cloud programming (functional programming, scala, spark)
  - Machine learning using Spark and PySpark

Social Network Analysis = QMSS G4062

- Greg Eirich (QMSS program director, good professor, very lively and approachable)
  - Intro level course
  - How to manipulate, analyze and visualize network data themselves using statistical software
  - Using R

Data Visualization = QMSS G4063

- Designed to the interdisciplinary and emerging field of data science. It will cover techniques and algorithms for creating effective visualizations based on principles from graphic design, visual art, perceptual psychology, and cognitive science to enhance the understanding of complex data.
- Positive feedback from QMSS students. Although professor may be a bit messy.

Bayesian Statistics for the Social Sciences = QMSS G4065

- Professor can seem a bit crazy, but is a big fan of Bayesian Stats and a contri
  STAN package.
- Any non-QMSS students interested in taking this course should have a compa
  background to a QMSS student in basic probability.

APPLIED DATA SCI FOR SOC SCIENTISTS = QMSS GR5069

- Professor is Head of Data Science at NBCUniversal
- They will say it’s hard to get in as non-QMSS students but can always try

Questions?

- Thanks for your attention!
- To get in touch with us: cdss_executives@columbia.edu