Introduction to Data Science

Course Logistics

Nicole Beckage

Focus of this course

- Goal: rational and analytic thinking about modern data science problems.
- By the end of the semester you should be able to:
  - Evaluate data science workflows (lectures)
  - Apply basic data science steps using R (homework)
  - Learn new technical topics (tech lecture)
  - Conduct Data-drive research (project)

Not part of this course

- Programming (required but not taught)
- Databases
- Specific tools
  - "The databases, software, and workflow management taught in a given Data Science Masters program are unlikely to be the same as those used by one specific employer…in contrast, Data Analysis and Statistics are broadly applicable skills that are portable from organization to organization." Donoho 50 Years of Data Science

Not part of this course

- Programming (required but not taught)
- Databases
- Specific tools
  - "The databases, software, and workflow management taught in a given Data Science Masters program are unlikely to be the same as those used by one specific employer…in contrast, Data Analysis and Statistics are broadly applicable skills that are portable from organization to organization." Donoho 50 Years of Data Science

Course expectations

- This is a graduate level course
- Outside class work should be 3 hours for every credit hour.
- That means you should spend ~9 hours a week on this course.
- You will get the grade you earn
- The higher the quality of work, the higher the grade.
- Often, time-quality, organization can help reduce time without loss of quality
- You must learn on your own.
- The difference between an undergraduate and graduate course is that my job is now to show you what you don’t know, not to teach you everything you need to know
- Office hours are a great place to get help.

Course expectations

- Much of the work is group based
- If issues come up within the group, you need to talk to me.
- Learning to work effectively in a team is a really important skill
- Data science projects are usually a group effort
- Professionalism, civility and respect (10%) 
  - Don’t be late (class and group).
  - Be respectful and helpful to students and instructors.
  - Be accountable to your fellow teammates.
- Prerequisite knowledge
  - Stats, programming, algorithms

Not part of this course

- Programming (required but not taught)
- Databases
- Specific tools
  - "The databases, software, and workflow management taught in a given Data Science Masters program are unlikely to be the same as those used by one specific employer…in contrast, Data Analysis and Statistics are broadly applicable skills that are portable from organization to organization." Donoho 50 Years of Data Science

Course expectations

- This is a graduate level course
- Outside class work should be 3 hours for every credit hour.
- That means you should spend ~9 hours a week on this course.
- You will get the grade you earn
- The higher the quality of work, the higher the grade.
- Often, time-quality, organization can help reduce time without loss of quality
- You must learn on your own.
- The difference between an undergraduate and graduate course is that my job is now to show you what you don’t know, not to teach you everything you need to know
- Office hours are a great place to get help.

Course expectations

- Much of the work is group based
- If issues come up within the group, you need to talk to me.
- Learning to work effectively in a team is a really important skill
- Data science projects are usually a group effort
- Professionalism, civility and respect (10%) 
  - Don’t be late (class and group).
  - Be respectful and helpful to students and instructors.
  - Be accountable to your fellow teammates.
- Prerequisite knowledge
  - Stats, programming, algorithms
Course tools

- Course website: ittc.ku.edu/~beckage/2017ds/index.html
  - Up to date assignment due dates, readings, slides, etc.
- Piazza (piazza.com/ku/fall2017/eecs731)
  - Used for required responses, course questions, team coordination, etc
- CATME
  - Team building, team evaluation
- R, (Anaconda), jupyter notebook (non-trivial setup), GitHub
- Homeworks, projects
- Latex
  - Project write-up

Course book

- Introduction to Statistical Learning
- Free pdf
- If this is too basic, try ESLR (DS bible to some)

Grade breakdown

Homeworks (15%)

- There will be 4 homeworks in R
- The answers are online BUT I know R very very well. I will know if you’re copying
  - Copying==0 and a report to the department
- Additionally, assigned article readings require response on Piazza before 12 noon on class day (marked on website with (*)
- Usually 1 week to read
  - This Thursday there is a reading assignment but it’s only 2 days. Sorry...
  - You have 3 late days to use on homeworks (not responses)

Research talks (%5)

- KU is a tier-1 research school
  - That means “Highest Research Activity”
- Requirement of the course: Go to 3 research talks
  - Two must be outside EECS
  - One must be outside school of Engineering
  - Has to use data and present results from data
- Keep an eye out in Piazza and post potential talks
- One page response due one week after the talk
Research Project

- Next week you’ll get assigned a group
- Next week I will present 4-5 research projects
  - These will be projects put forth by faculty at KU
- First team assignment will be to rank project ideas
- Project work is ~70% of the grade in the course.
- Starting Sept 5, rotating bi-weekly progress meetings last 15 minutes of class. If you’re not meeting with me, you’re meeting with your group.

Technical Lecture (20%)

- Each research project requires different skills
- Each group will be responsible for delivering a 1 hour lecture on a technical skill related to the project (early Oct)
  - Possible examples: GIS, sentiment analysis, topic analysis, time series modeling, dagger algorithm, EEG signal analysis
- Grade based on lecture and lecture notes
  - Practice makes perfect!
  - Classmate evaluations

Research Presentation and Paper (45%)

- Last few lectures are lectures. One lecture per class, 2 during final slot.
  - You must attend the final: Dec 13 1:30-4pm. You cannot pass the class if you cannot make the final.
- Grade for research presentation and paper will also include
  - Paper workshoppping
  - Bi-weekly meetings
  - Code (5%)
  - Team evaluations

Clash of the titans: R vs Python
Why we’re using R

- It’s good to learn a new language!
- The book uses R
- R has a richer view of data analysis
  - But it has a steep learning curve
- R plays nice with LaTeX (more to come on that)
- R can be used with jupyter notebooks
  - Though it’s non-trivial to set up.

R help

- R has a large supportive community. Almost every issue you have will be on stackoverflow.
- Another important reference is the R help pages.
  - ?Function

Quick R demo

- How to interpret help files

Latex

- Powerful text editor
- Useful skill to many companies
- Integrates R code directly within a document– compile document == run code (Sweave)
- It has a learning curve, but it’s good to know about it.
- Only required for final project.
- DO NOT wait until paper is written to put it into LaTeX

Jupyter notebook and R

- Jupyter notebook (ipython notebook) now has R capability
- Non-trivial setup so please start ASAP.
- My experience:
  - download anaconda
  - it comes with (or is easy to download) R-studio (optional)
  - I still recommend downloading R from CRAN
  - then go here
  - My mac needed curl updated among other things
  - Make sure you install things where recommended

Quick Jupyter demo
Due Thursday

- Introduction on Piazza (part of HW0)
  - Introduce yourself to your classmates and me with a follow-up post in the introduction thread. Include your name/nickname, your affiliation, why you are taking this course, and tell us something interesting about yourself (e.g., an industry job, an unusual hobby, past travels, or a cool project you did etc.). Also tell us whether you have experience with data science, statistics courses and research projects.
- Team survey (CATME)
- Read 50 Years of Data Science (on course website)
  - Post a comment/question related to the article
  - Be ready to discuss in class tomorrow.
- Look over first homework assignment (on course website)
- Start setting up Jupyter notebooks

Basic idea behind data science

- Collect or find data related to a specific scientific question
- Analyze it to find patterns
- Use the patterns to do something

Disciplines that have benefitted from data science:
- natural sciences (e.g. genomics, neuroscience, astronomy)
- web technology (e.g. Google, Netflix)
- finance (e.g. stock predictions)
- policy (e.g. predicting what intervention X will do)

Possible analysis tools:
- linear models
- deep neural networks
- decision trees
- and many others
Basic idea behind data science

- Collect or find data related to a specific scientific question
- Analyze it to find patterns
- Use the patterns to do something
  - prediction
  - analytics
  - recommender system
  - model architectures

6 divisions of GDS

- Data Exploration and Preparation (homework 0-2)
- Data Representation and Transformation (lectures, homework)
- Computing with Data (lectures, project)
- Data Modeling (project, technical lectures, lectures)
- Data Visualization and Presentation (project, lectures)
- Science about Data Science (lectures)

Another perspective

Adapted from CS109 at Harvard

20th Century Innovation

- Cars
- Airplanes
- Power grid
- Digital computers
- Internet

20th Century questions

- Does fertilizer increase crop yields?
- Does Streptomycin cure Tuberculosis?
- Does smoking cause lung-cancer?
What's the difference?

- Deterministic versus random
- Deductive versus empirical
- Solutions deduced (derived) mostly from theory versus solutions deduced most from data

21st century

“I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?”

- Hal Varian, Google’s Chief Economist

Data Science in the wild

“The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data.”

- Hal Varian

For your enjoyment:

- Partially Derivative podcast
- Linear Digressions podcast
- Data Skeptic podcast
- Andrew Gelman’s blog
- Simply Statistics blog
- Edwin Chen’s blog