COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

Discussion Section A01
Week 9

Teaching Assistant (TA): Abdullah
Instructional Assistant (IA): Kyra

Where to find all material

COGS 9

Search COGS 9 UCSD Podcast Gradescope

- Home
- Syllabus
- Readings
- Assignments
- Exam
- Final Project
- Office Hours
- Contact Us

This site uses Just the Docs, a

Introduction to Data Science

COGS 9 - UC San Diego - Prof. Kyle Shannon

Spring 2024 | SOLIS 107 | TU & TH 5:00-6:20PM

Welcome 👋

We are all very excited that you decided to join us on this whirlwind tour of data science. All relevant info, e.g. due dates, assignment links, etc. are found on this website. We look forward to teaching and working with all of you and hope to meet you in office hours. Check out the **Getting Started** section so you can hit the ground running when class starts!

NOTE
Week one I try to take as many students from the **waitlist** as I can, please email cogsadvising@ucsd.edu with further questions.

Discussion Sections

	Day	Time	Location	Staff	Materials
A01	Wed	12:00-12:50PM	CENTR 222	TA: Abdullah IAs: Kyra	View
A02	Wed	1:00-1:50PM	CENTR 222	TA: Kaushik IAs: Seshu, Vicky	View
A03	Wed	2:00-2:50PM	CENTR 222	TA: Matthew IAs: Jessica, Wenhua	View
A04	Wed	3:00-3:50PM	CENTR 222	TA: Vineeth IAs: Jiesen	View
A05	Wed	4:00-4:50PM	CENTR 222	TA: Vineeth IAs: Harshita	View

cogs9_TA Public

main | 1 Branch | 0 Tags | Go to file

AbdullahAshfaq Added week3 material

week2	Added week2 material
week3	Added week3 material
README.md	Update README.md

README

Cogs 9 Discussions-Intro to Data Science

Abdullah's discussion section material for COGS9 course

Upcoming Deadlines

Week 9		
Mon, May 27	ASSG	Assignment 3 due
Tue, May 28	GLCT	Guest Lec 1. (prerecorded not in-class)
Thu, May 30	GLCT	Guest Lec 2. (prerecorded not in-class)
Fri, May 31	QUIZ	Reading Quiz 5 due

Week 10		
Tue, Jun 04	LECT	Algorithms & Computability
Thu, Jun 06	LECT	Future of Data Sci & Jobs
Thu, Jun 06	PROJ	Final Project Part 2 due

Week 11		
Wed, Jun 12	EXAM	Take home exam released (@10PM)
Thu, Jun 13	EXAM	Take home exam due (@10PM)

Discussion Sections Outline: Mostly Hands-on

- Week 2: Introductions, Making teams, Reading 1 (Part 1)
- Week 3: Reading 1 (Part 2), Python Basics with Jupyter Notebook
- Week 4: Reading 2, Getting data and wrangling it using Pandas
- Week 5: Reading 3, Assignment 1, Basics of SQL and Visualizations
- ~~Week 6: Reading 4, Final Project Part 1 reviews/discussions~~
- Week 7: Reading 4, Assignment 2, Data Visualization and EDA demo
- Week 8: Assignment 3, Machine Learning demo
- Week 9: Reading 5, Closing thoughts
- Week 10: Final Project Part 2 reviews/discussions

Today’s Outline

Participation = Extra Credit 😀

- Reading 5 Summary
- Questions about Project Part 2

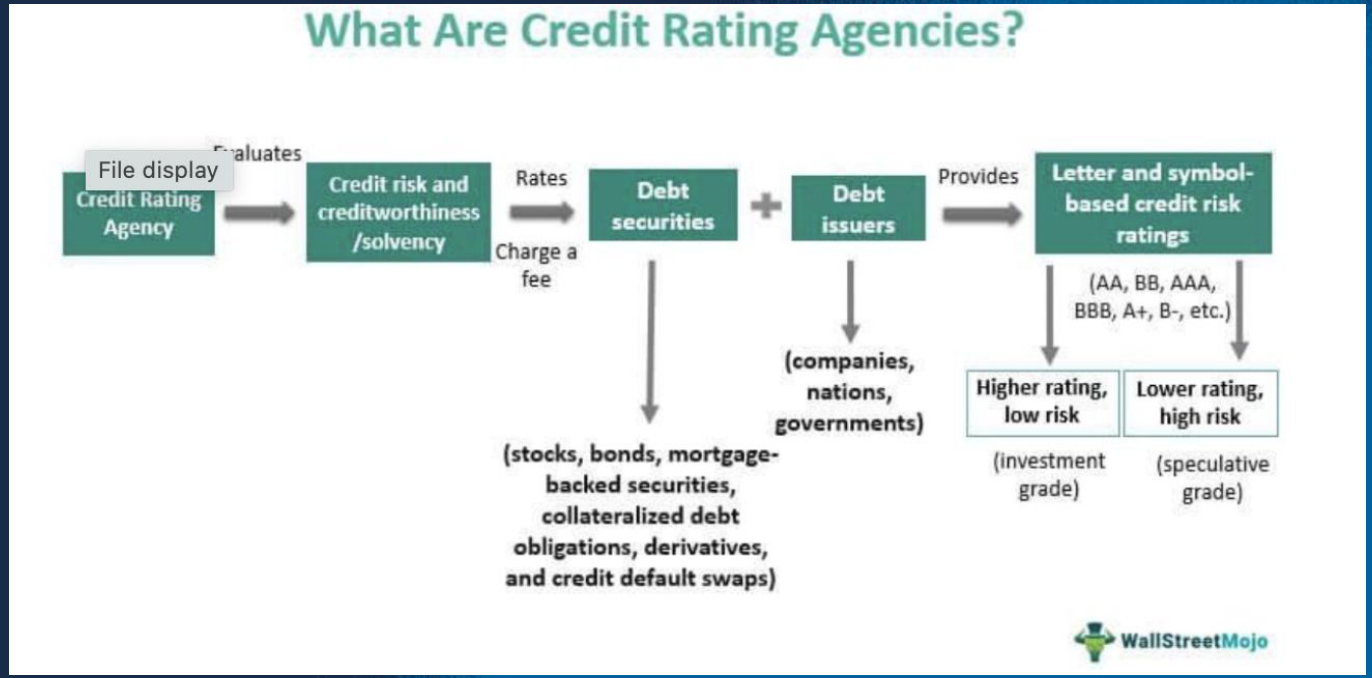
Reading 5
Accountability in Algorithmic Decision Making

Algorithmic Decision Making

1. Prioritizing
2. Classification
3. Association
4. Filtering

Government vs. Privacy

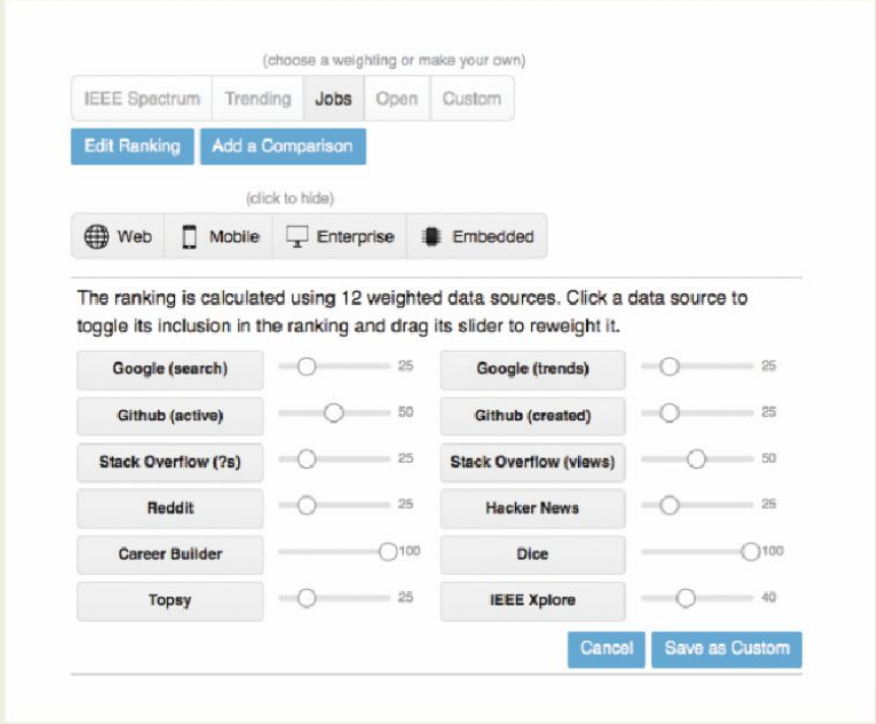
Sector Accountability



An Algorithmic Transparency Standard

1. Human involvement
2. Data
3. The model
4. Inferencing
5. Algorithmic presence

IEEE top programming languages ranking and reweighting interfaces.



Reading 5
Biased Machine

Machine Learning Algorithms

- Used in predicting future criminal behavior
- Analyze historical data to make prediction
- Touted as a way to make more objective decision

The biggest weakness

- Objectiveness depends on the data
- Biased against black people, leading to higher rate of false positive

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Call the fairness in the criminal justice system

- Greater transparency and accountability in the use of algorithms
- Involvement of experts and community members in the development and evaluation of algorithms
- The need for increased awareness and understanding of the potential for bias in algorithms

Project Part 2

New Sections in Report

- Analysis Proposal (20 pts)
  - Data Collection (3 pts)
  - Data Wrangling (3 pts)
  - Descriptive & Exploratory Data Analysis (3 pts)
  - Data Visualization (3 pts)
  - Analysis Type 1 (4 pts)
  - Analysis Type 2 (4 pts)

- Discussion (5 pts)

Data Collection

Obtaining Data...

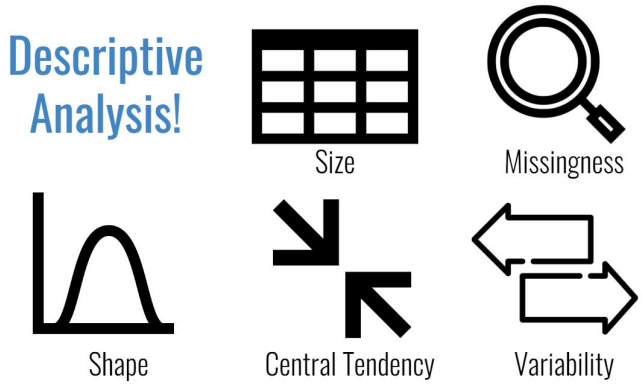
- Buy it.
- Source it internally from your data
- Collect it externally from your users
- Freely download it from the web
- Request it from an API (paid/open source)
- Scrape it from a website
- Steal it (Intentionally or unintentionally. Don't do this…)

Data Wrangling

Rules for Tidy Spreadsheets

1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD (remember ISO-8601)
4. No empty cells (what does empty mean…..??)
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files (i.e. → .csv != .xlsx)

Descriptive and EDA



Dimensions of Analysis within EDA

Univariate: Seeking to explore, plot, and measure **one** variable

Bivariate: Seeking to explore, plot, and measure **two** variables

Multivariate: Seeking to explore, plot, and measure **many** variables

Data Visualization

- Make some charts
- Explain why you chose these chart
- How will you interpret the results from the chart

Analysis Type 1 - Inferential

Approaches to Inference

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

i.e. simple regression, multiple regression

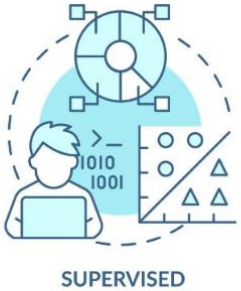
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

Analysis Type 2 - Predictive

Supervised Learning



- Labeled data
- Make predictions
- Classification or Regression!

Unsupervised Learning

- Unlabeled data
- Find structure
- Reduce Dimensions

Reinforcement Learning

- Learn a set of actions
- Reward feedback system
- Agent explores a world

Discussion

- Interpret results
- Discuss limitations/pitfalls/biases (Slightly related to ethics modeling and deployment part)
  - Societal and Ethical implications
- Discuss how you would address the limitations

Thanks!