Projects

- Goals
  - Do something you enjoy, find something interesting
  - Go beyond class, be creative, feel like a Graduate student
- Types
  - Data exploration and inference
  - Fundamental data-science paper review
  - P&S engagement - new topic, problems, demo, game
- Guidelines
  - One or two students per project
  - Due Sunday after exam week, at noon
  - Roughly 20 hours per project member
  - Please innovate

Deadlines

- Saturday 11/11
  - Tentative project type, topic, and team (≤2) selection
  - Google form, link will be sent in a canvas announcement
- Tuesday 11/21
  - Final selection, including a 250-word plan
  - Google form
- Sunday 12/17, noon
  - Projects due, , please upload earlier if possible

Data Exploration

- Pick interesting dataset
- Identify important question or inference task
- Visualize and understand data
- Shed light on question or task
- Use something learned in class
- Can use techniques beyond this class
- Be creative
- Examples next

kaggle.com

- Platform hosting diverse datasets and competitions
  - Wine review
  - Pollution data
  - Speed dating
  - TMDB 5,000 movies
  - ….

Wine

- 10 Features (cost, rating, region, winery, variety,…)
- Most correlated features
- Most predictive feature for rating
- Predict rating based on features

Speed Dating

- Self descriptions, peer descriptions, 2nd dates
- Most / least popular interest
- Most / least correlated features
- Correlation between data and likelihood of 2nd date
- Predict probability of 2nd date

Bit Coin

- Time, price, volume, per minute
- Correlate price, time, and volume
- Predict price based on previous minutes
- Long-term prediction

California Housing Prices

- Housing Data in a California district from 1990 census
- Identify relevant features: rooms, bedrooms, ocean proximity
- Predict median value of houses using regression
- Dataset

Walmart Sales

- Walmart weekly sales data from 45 US stores
- Sales trends during the year, holiday period
- Correlate with temperature, fuel price, consumer price index
- Regression with these features and sales volume to predict sales
- Dataset

Used Cars Database

- Cars on German eBay
- Regression to identify depreciation rates of each model
- Compare different models
- Hypothesis: manual and automatic transmission depreciate same
- Dataset

Climate Change

- 1.6 billion temperature measurements
- Hypothesis: Temperatures at two periods have same mean
- Is climate change real?
- Dataset

Traffic Data

- CA Traffic data http://pems.dot.ca.gov
- Visualized: http://trafficpredict.com/current_traffic/?location=los-angeles
- Model traffic flow between SD and LA
- Predict traffic patterns
- Find best routes
- Estimate travel times

Data-Exploration Report

- 3-4 pages, well written, pdf, Google Colab, Jupyter notebook
- Describe data
- Problem and why it is interesting
- Preprocess data
- Model and methodology
- Visualization
- Insights gained
- Code
- On your own, cite all external resources

Paper Review

- Select insightful paper
- Read, understand, write report
- Eight papers follow, you can choose others

Some Paper Examples

- PCA I
  - Zou, Hui, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis." Journal of computational and graphical statistics 15.2 (2006): 265-286.
- PCA II (Applications in Bioinformatics)
  - Ma, Shuangge, and Ying Dai. "Principal component analysis based methods in bioinformatics studies." Briefings in bioinformatics 12.6 (2011): 714-722.
- Bayesian I
  - Tipping, Michael E. "Sparse Bayesian learning and the relevance vector machine." Journal of machine learning research 1.Jun (2001): 211-244.
- Bayesian II (Applications in Neuroscience)
  - Darlington, Timothy R., Jeffrey M. Beck, and Stephen G. Lisberger. "Neural implementation of Bayesian inference in a sensorimotor behavior." Nature neuroscience (2018): 1.
- Regression I
  - Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.
- Regression II (Applications in Bioinformatics)
  - Wu, Tong Tong, et al. "Genome-wide association analysis by lasso penalized logistic regression." Bioinformatics 25.6 (2009): 714-721.
- Hypothesis Testing I
  - Javanmard, Adel, and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression." The Journal of Machine Learning Research 15.1 (2014): 2869-2909.
- Hypothesis Testing II
  - Malek, A., Katariya, S., Chow, Y., & Ghavamzadeh, M. (2017, April). Sequential multiple hypothesis testing with type I error control. In Artificial Intelligence and Statistics (pp. 1468-1476).

Paper-Review Report

- 3-4 pages, pdf, well written
- Emphasize concepts over formulas
- Do on your own
- Cite all external resources
- Include
  - Abstract summary
  - Problem addressed
  - Significance
  - Ideas
  - Contributions
  - Results and visualization
  - Implications

P&S Engagement

- Review new topics
    - Theoretical or applied
    - Describe topic, importance
- Problems
    - 10-15, at least 5 in Python, include solutions
    - If taken from elsewhere, state source
- Demo
- Game
- Interesting, exciting, fun

Past Projects

- Sample projects will be uploaded to Canvas