Scalable Data Analysis (CIS 602-02)

Clustering

Dr. David Koop
Statistical Modeling

- Breiman in 2001: Two cultures
  - The black box
  
  \[
  \begin{array}{c}
  y \\
  \end{array} \quad \begin{array}{c}
  \text{nature} \\
  \end{array} \quad \begin{array}{c}
  x \\
  \end{array}
  \]

  - Data Modeling
  
  \[
  \begin{array}{c}
  y \\
  \end{array} \quad \begin{array}{c}
  \text{linear regression} \\
  \text{logistic regression} \quad \text{Cox model} \\
  \end{array} \quad \begin{array}{c}
  x \\
  \end{array}
  \]

  - Algorithmic Modeling
  
  \[
  \begin{array}{c}
  y \\
  \end{array} \quad \begin{array}{c}
  \text{unknown} \\
  \end{array} \quad \begin{array}{c}
  x \\
  \end{array}
  \]

  \[
  \begin{array}{c}
  \text{decision trees} \quad \text{neural nets} \\
  \end{array}
  \]
Two Cultures

- Breiman: Data Modeling Culture has
  - "Led to irrelevant theory and questionable scientific conclusions"
  - "Kept statisticians from using more suitable algorithmic models"
  - "Prevented statisticians from working on exciting new problems"

- Most articles start "Assume that the data are generated by the following model"
  - Is the model being fit to the data?
  - If multiple models work (goodness-of-fit), what do we learn?
Algorithmic Modeling

• Instead focus on solutions and predictive accuracy, "Live with the data"

• Lessons:
  - Multiplicity of Good Models: close in accuracy, distinct in form
  - Simplicity vs. Accuracy: a tradeoff between interpretability and accuracy (e.g. decision trees vs. random forests, ensembles)
  - Dimensionality: Bellman says this is a curse, but sometimes more dimensions lead to better solutions
Responses to Breiman

• "Data looking for a question" — Cox

• "Sample sizes have swollen alarmingly while goals grow less distinct ("find interesting data structure")" — Efron

• "Algorithms often appear in the form of black boxes with enormous numbers of adjustable parameters, …sometimes more knobs than data points" — Efron

• "High performance (predictive accuracy) on the test sample does not guarantee high performance on future samples; things do change" — Hoadley
Assignment 2

- [http://www.cis.umassd.edu/~dkoop/cis602/assignment2.html](http://www.cis.umassd.edu/~dkoop/cis602/assignment2.html)
- Bikes and Weather
- Bike sharing data and weather for Washington DC
- Hypothesis might be that people are more likely to use the program when the weather is "nicer"
- Integrate data, visualize it, do regression
- Use pandas, matplotlib, and scikit-learn libraries
- Can generate a solution with relatively little code (~50 lines) but will require experimentation to get there
- Due Tuesday, October 20
Project Proposal

• Type of project: (a) Analyze real data, (b) Improve a technique

• Analyze real data
  - What are the datasets?
  - How much data is there (size/scope/time span)?
  - What questions would you be interested in?
  - What has already been done? Related Work
  - What techniques do you plan to use to answer these questions?

• Improve technique
  - What is the problem?
  - What ideas do you have to solve the problem?
  - What has already been done? Related Work
  - How do you plan to evaluate your improved technique? (e.g. data)
Project Proposal

• You may work on the same dataset but must have independent questions, investigations, code, and reports
• Datasets must be large enough to provide scalability challenges
• Potential datasets:
  - https://github.com/caesar0301/awesome-public-datasets
  - https://github.com/fivethirtyeight/data
  - http://datahub.io/dataset
  - http://www.37billionmilechallenge.org
  - https://data.cityofboston.gov
Class Attendance and Participation

- This is part of your grade...

- Reading Quiz
Machine Learning

A. Zisserman
Machine Learning Problems

- **Supervised**: have input data and known labels, build solution that will predict labels from new inputs
  - Classification: predict class (yes/no, apple/banana/pear)
  - Regression: predict parameters (e.g. given weight, predict height)

- **Unsupervised**: have input data but no given labels, want to cluster the data

- **Semi-supervised**: e.g. cluster handwritten digits, ask a human to label a few from each cluster (perhaps those at boundaries), continue learning
Clustering

J. Leskovec, A. Rajaraman, J. Ullman
Scaling Clustering Algorithms to Large Databases

P.S. Bradley, U. Fayyad, C. Reina
Next week

• No class on Tuesday, October 13 (Monday schedule)
• Thursday: Databases
• Assignment 2
• Project Proposal
Clustering

- Major categories:
  - Hierarchical: start with individual points and join neighbors together
  - Point Assignment: start with known cluster "centers" and put points into the "nearest" cluster