Course Structure

Statistical Modeling

Sample vs. Population

- Sample: a subset of all the data
  - E.g., Results of polling only a few individuals per state
  - Problem: incomplete picture
- Population: all the data
  - E.g. US census
  - Problem: potentially inaccurate picture
    - Measurement error (systematic, random)
    - Fluctuations: If we could “rewind the tape of history, ...” [SJ Gould]

Statistical Modeling

- Statistical model: assumed, idealized description of the data
    - Selects a few key variables of interest
    - Might make assumptions about how they are distributed
    - Might describe how they relate to one another
- Desiderata
    - Plausible
    - Interpretable
    - Simple (“the simplest explanation is best”)
    - Generalizable (i.e., applicable well beyond the sample)

Course Structure

1. Descriptive Statistics: summarizing data
    - Histograms
    - Measures of central tendency
    - Measures of dispersion

No underlying statistical model. No learning. No inference.

Just Exploratory Data Analysis.

Course Structure (cont’d)

2. (Classical) Statistical Inference
   - Assume an underlying statistical model of a population
   - Learning: Estimate true parameters of the model, using in-sample data
     - sample mean, sample variance, etc.
     - confidence intervals, hypothesis testing, etc.
   - Inference: Generalize to the entire population (i.e., out-of-sample)

Model checking is key!

“All models are wrong, but some are useful.” -- George Box

Course Structure (cont’d)

3. (Statistical) Machine Learning
    - Assume an underlying statistical model of a population
    - Learning: Estimate true parameters of the model, using in-sample data
        - sample mean, sample variance, etc.
        - confidence intervals, hypothesis testing, etc.
    - Inference: Generalize to the entire population (i.e., out-of-sample)

Model checking is key!

“All models are wrong, but some are useful.” -- George Box

Learning vs. Inference

Learning vs. Inference

- Learning/Train: given sample data, build a model
    - Supervised: model from input variables to output variables
        - E.g., from symptoms to likelihood of disease
        - E.g., from indicators to school performance
        - Relies on some notion of ground truth
    - Unsupervised: estimate model parameters
- Inference/Test: given a single new datum, apply model

Learning vs. Inference: Cholera

- Learning/Train: given sample data, build a model
  - Unsupervised: estimate model parameters
    - Average number of deaths per 10,000 houses in London
- Inference/Test: given a single new datum, apply model
    - Number of deaths per 10,000 houses in Lambeth
    - Number of deaths per 10,000 houses in S&V

Learning vs. Inference: Baseball

- Learning/Train: given sample data, build a model
    - Supervised: model from input variables to output variables
        - E.g., from various indicators to player performance
- Inference/Test: given a single new datum, apply model
        - Predict (infer) performance of, say, Carlos Correa (rookie)

Learning vs. Inference: Elections

- Learning/Train: given sample data, build a model
  - Supervised: model from input variables to output variables
    - Use regression analysis to predict future state polls
  - Unsupervised: estimate model parameters
    - Aggregate state polls into national model
    - 538: tweaks polls based on past election results
- Inference/Test: given a single new datum, apply model
    - Predict future state polls and future national model
    - Infer winning probabilities using future national model

Learning vs. Inference: Netflix

- Learning/Train: given sample data, build a model
    - Unsupervised: estimate model parameters
        - Define clusters of users based on movie preferences
- Inference/Test: given a single new datum, apply model
        - Infer movie recommendations for users