Day-by-day Objectives for Math 300R

Danny Kaplan

Lessons 1-18

As done in Fall 2020. Possible revisions to those lessons is not a topic of this proposal.

Lesson 19: Decisions with data (nti)

1. Distinguish between the two settings for decision-making:

    a. **Prediction**: predict an outcome for an individual
    b. **Relationship**: characterize a relationship with an eye toward intervention or a better understanding of how a mechanism works.

2. Given a research question, identify whether it corresponds to a prediction setting or a relationship setting.

Lesson 20: Reality versus gaming (nti)

1. Understand that gaming is a way of improving our skills and identifying potential opportunities and problems.

2. Enumerate the four stages of the games we will use and identify which ones correspond to non-gaming, real-world work with data.

The Four Stages (to be moved to NTI)

i. building the deck: Instructors provide a simulation of a mechanism that generates rows of a data frame.
ii. the deal: Some of these rows will be dealt to you, constituting the data you have to work with. [real-world]
iii. the play: Build models and extract results. [real-world]
iv. the reveal: Compare your results from (iii) either to the mechanism given in (i) or to more data generated by the simulation.

1. Distinguish between a sample, a row, and a sample of samples.

Lesson 21: DAGs, noise, and simulation (nti)

1. Determine whether a proposed graph is directed and acyclic.
2. Read notation to identify response variable, explanatory variable, covariates, and effect sizes.
3. Characterize the magnitude of random noise.
4. Generate data from simulations and summarize variables individually.

Lesson 22: Sampling variation (nti)

1. Implement on the computer a procedure to generate a sample, calculate a regression model, and produce a summary.
2. Iterate the procedure and collect the summaries across iterations.
3. Graphically display the distribution of summaries and generate a compact numerical description ("confidence interval") of the sampling distribution.

Lesson 23: Estimate sampling variation from a single sample (nti)

1. Use bootstrapping to estimate sampling variation.
2. Infer sampling variation from a regression table.

Lesson 24: Effect size (nti)

1. Estimate an effect size from a regression model of the two variables.
2. Construct a confidence interval on the effect size.
3. Evaluate whether confidence interval indicates that estimated effect size is consistent with simulation.
4. Understand and use scaling of confidence interval length as a function of n.

Lesson 25: Mechanics of prediction (nti)

1. Given a sample from a DAG simulation, construct a predictor function for a specified response variable.
2. Use the predictor function to estimate prediction error on a given DAG sample and summarize with root mean square (RMS) error.
3. Distinguish between in-sample and out-of-sample prediction estimates of prediction error.

Lesson 26: Constructing a prediction interval

1. Identify the two components that make up a prediction error, one that scales with n and the other that doesn't.

Lesson 27: Covariates

1. Show that including covariates in a prediction model always reduces in-sample mean square residual, but may not reduce residuals out-of-sample.
2. Given regression coefficients, calculate model degrees of freedom and residual degrees of freedom.
3. Calculate amount of in-sample mean square error reduction to be expected with a useless (random) covariate. (Residual sum of squares divided by residual degrees of freedom.)

Lesson 28: Covariates eat variance

1. Construct F statistic as ratio of incremental increase in model mean square due to model term(s) divided by residual mean square.
2. Use software to construct ANOVA report and correctly interpret F statistics for prediction model term selection.

Lesson 29: Confounding

1. Identify confounding in a DAG
2. Choose whether to include covariate depending on form of DAG

Lesson 30: Non-causal correlation

1. Distinguish “common cause” and “collider” forms of DAG.
2. Construct appropriate DAG to match a narrative hypothesis.

Lesson 31: Experiment and random assignment

1. Properly use nomenclature of experiment.
2. Correctly re-draw DAG for an ideal experimental intervention.
3. Use blocking to set assignment to treatment or control.

Lesson 32: Measuring and accumulating risk

1. Distinguish between absolute and relative risk and identify when a change in risk is being presented as absolute or relative.
2. Calculate and correctly interpret other presentations of differences in risk: population attributable fraction, NTT, odds ratio.
3. Interpret effect size as stated in log odds.

Lesson 33: Constructing a classifier

1. Build a classifier from case-control data.
2. Cross-tabulate classifier results versus true state. Evaluate false-positive rate, false-negative rate, accuracy.
3. Calculate different forms of conditional probability: p(A|B) versus p(B|A) and identify which form of conditional probability is useful for prediction of an individual’s outcome.

Lesson 34: Accounting for prevalence

1. Explain why case-control data may not give an proper measure of “prevalence.”
2. Convert

Lesson 35: Hypothesis testing

1. Understand and use properly hypothesis testing nomenclature: test statistic, sampling distribution under the null, Type-1 and Type-2 error, rejection threshold, p-value
2. Contrast hypothesis testing versus Bayesian framework.

Lesson 36: Calculating a p-value

1. The permutation test
2. Interpret correctly from regression/ANOVA reports
3. Traditional names for hypothesis tests in different "textbook" settings.
4. Distinguish between p-value and effect size, that is, "significance" and "substance."

Lesson 37: False discovery with hypothesis testing

1. Identify signs of false discovery in a research paper.
2. Estimate how overall p-value should change when study is replicated.

Alternative 1

Theme: Classifiers: ROC and loss function

Alternative 2

Theme: Accumulating risk: Logistic regression