STATISTICAL HORIZONS

Machine Learning

Kevin Grimm, Ph.D.

Upcoming Seminar:
January 31-February 1, 2020, Fort Myers, Florida

Machine Learning

Kevin J. Grimm

Department of Psychology
Arizona State University

Introductory Comments

R. A. Berk (2009)

"As I was writing my recent book on regression analysis (Berk, 2003), I was struck by how few alternatives to conventional regression there were. In the social sciences, for example, one either did casual modeling econometric style, or largely gave up quantitative work. The life sciences did not seem quite as driven by causal modeling, but causal modeling was a popular tool. As I argued at length in my book, *causal modeling as commonly undertaken is a loser*."

R. A. Berk (2009)

"There also seemed to be a more general problem. Across a range of scientific disciplines there was often too little interest in statistical tools emphasizing induction and description. With the primary goal of getting the 'right' model and its associated p-values, the older and more interesting tradition of exploratory data analysis had largely become an under-the-table activity: the approach was in fact commonly used, but rarely discussed in polite company. How could one be a real scientist, guided by 'theory' and engaged in deductive model testing, while at the same time snooping around in the data to determine which models to test? In the battle for prestige, model testing had won."

R. A. Berk (2009)

"At the same time, I became aware of some new developments in applied mathematics, computer sciences, and statistics making data exploration a virtue. And with this virtue came a variety of new ideas and concepts, coupled with the very latest in statistical computing. These new approaches, variously identified as 'data mining', 'statistical learning', 'machine learning', and other names, were being tried in a number of natural and biomedical sciences, and the initial experience looked promising."

Morgan & Sonquist (1963)

"The essence of research strategy consists of putting some restrictions on the process in order to make it manageable. One possibility is to cut the number of explanatory factors utilized, and another is to restrict the freedom with which we allow them to operate. One might assume away most or all interaction effects"

Morgan & Sonquist (1963)

"Clearly, the more theoretical or statistical assumptions one is willing to impose on the data, the more she/he can reduce the complexity of the analysis. A difficulty is the restrictions imposed in advance cannot be tested."

"There seems some reason to argue that it would be better to use an approach which developed its restrictions as it went along."

Most people only have partially confirmatory models

1. People who are very good at research only have "partially confirmatory models."
2. That is, they seem to know what they are planning to do, and hence can use *a priori* distributions and probabilities.
3. But they are also willing to adjust to the data (including outlier, transformations, and the addition of other model variables and parameters). This is OK!
4. The only problem is we do not have a priori ideas about this, so the use of regular distributions is just a way to guarantee a result will be found. This is not OK!
5. So we probably need to change our research strategies and use a more optimal method of analysis for these problems.

The Confirmatory-Exploratory Dimension of Research

Historical notes on statistics in the social sciences

- The history of psychological statistics shows great respect for *a priori* hypothesis testing (Fisher, 1928), and has led to many organized and successful research programs (see Box, 1976).

- Unfortunately, this also led to unwarranted skepticism and disdain for the good parts of "Exploratory Data Analysis" (e.g., Tukey, 1963, 1977).

Historical Notes on Statistics in the Social Sciences

- As a result, many data analysts started to search for clear and *a priori* hypotheses. Unfortunately, these hypotheses seemed to be lacking.
- In response, many researchers often trimmed their datasets, relied on post hoc adjustments, and/or searched their data for 'significant' associations
- In sum, researchers said they used confirmatory methods when, in fact, their work was, out of necessity, largely exploratory in nature. At the most it was restrictive.

Machine Learning Topics

Machine learning techniques

- There are a large set of techniques that come under the general rubric of *machine learning*.
  - One common feature of these techniques is that they are *exploratory* and rely on *computer assisted* analysis.
- One large sub-division of these techniques uses a single outcome and tries to make a optimal prediction of this outcome from multiple predictor variables
  - Termed *supervised learning* techniques.
- Second subdivision does not require an outcome and merely classifies persons into subgroups based on similarities among a set of variables
  - Termed *unsupervised learning* techniques.

Regression

- Often not used as a statistical learning technique, but it is more flexible than you think

Smoothers

- Goal is to determine form of association between two variables
  - Lots of different ways to create smooth curves (e.g., B-splines, local regression)
  - Some are parametric, semi-parametric, or non-parametric

Variable selection methods in regression

- Goal is create a regression model that is sparse
  - Model only includes variables that are necessary for accurate prediction
- Some methods enable the analysis of data where the number of variables (p) is larger than the sample size (N)
- Some methods can include or search for interactive effects and some methods allow for nonlinear associations

Classification and regression trees

- Goal is to create a sparse prediction model
- Involves variable selection and determining optimal cut-off values on the chosen variable
- Creates a series of decision rules that are easy to follow and interpret

- Expansions include Bagging, Boosting, and Random Forests

Principles

Be honest (Rogers, 1929)

- The first principle of ethical data analysis is that we do not need PURITY of statistical rules and assumptions -- but we do need HONESTY.
- Try to tell us exactly (as briefly as possible), how you selected the people, variables, and occasions, even if it is complicated. Consider missing data, outliers, and transformations, but tell us their impacts on the results.
- Try to tell us exactly how you found the statistical models you used, especially if they were not *a priori* and if they emerged as part of the analysis.
- Tell us *all* relevant results, not just the *best* ones.

Goal is replication (Lykken, 1968)

- Key criterion is any experiment/analysis is replication
- How much do we actually study replication?
  - Confirmatory analyses are confirming an *a priori* theory, so is replication necessary?
- Data mining is inherently exploratory, so evaluating replicability is key
  - Training data
  - Test data

Confirm then explore (Tukey, 1962)

Phase C: Confirm –

- Try to come into an analysis with a plan about the set of ideas you are going to examine and the data you are doing to use to do so – this will permit a full and appropriate use of statistical probability tests and other indices of fit.
- Remember that we do not want the "Best" model, we want the "Set" of models that fit well separated from those that are "Average" and "Poor"

Phase E: Explore –

- Whether or not your favorite model fits the data on hand, try to improve the fit using the data on hand.

Introduction to the R Statistical Environment

R

- Free and *open source* software environment for statistical computing and graphics.
- Open source indicates the original source code is freely available, may be redistributed, and modified.
  - Allows & encouragers researchers to modify, extend, and develop 'additions' to the base program
  - Additions are referred to as *packages*

RStudio

- User friendly interface for R
- Makes programming easier, particularly if you are unfamiliar with key terms

Object-oriented programming

- Programming paradigm based on the concept of "objects", which may contain data, in the form of fields, often known as attributes; and code, in the form of procedures, often known as methods.
- A feature of objects is that an object's procedures can access and often modify the data fields of the object with which they are associated

Data structures

- Data structures are a form of organizing and storing data
- There are five basic types of data structures
  - Vector, Matrix, List, & Data frame