Tutorial proposal

Title: Important relationships in data: magnitude and causality as flags for what to focus on

Organizer: Joseph E. Beck
Computer Science Department
Worcester Polytechnic Institute

Preferred length: half day

Theme and goals

Advances in storage and networking, and scale up in deployment of intelligent tutors, have led to an explosion of data available for analysis. This expansion of raw material for analyses has taken place both in terms of columns and rows of data. Columns of data refer to the number of variables available, and has expanded as more data are logged by tutors, and as more external sources of data are applied. For example, if a researcher builds a series of student models for various cognitive (e.g. knowledge tracing) or behavioral (e.g. off-task detector) properties, that expands the number of columns available for analyses. Another common method for greatly expanding the number of columns is to join two databases together; for example, one could relate to semantic properties words the student used in response to a question. Increases in the number of rows are a result of recording datasets involving more students, or recording data at a finer-grain size. For example, recording every student response rather than just summary data for the student.

In general, both of these of trends are useful, and have greatly extended the scope and quality of analyses performed from data collected by ITS. An increase in the number of columns means that researchers are now capable of testing many more hypotheses than they were capable of previously. In addition, an increase in the number of rows results in greater statistical power, enabling greater sensitivity for detect small effects that may exist. Although these advances have brought great benefits, there is also a definite cost in terms of an increase in the number of analyses that are reportable, but are of marginal utility and may even be false.

The reason for concern is due to arithmetic. First, as the number of columns grows, the number of testable relationships increases as columns^2, since each new variable can be tested against all of the existing variables in the database. Second, the ability to detect statistically "significant" effects increases with the number of rows, and increases according to sqrt(rows). This two effects are additive, and result in a vast increase in the number of significant relationships one can discover from the collected data. The problem arises when one considers the number of *useful* relationships in the data. Many, many variables will correlate with each other just due to random chance, or due to being associated merely by sharing a common cause. Discovering all of these chance associations is not exciting from a research standpoint, but by community standards, such results would be publishable, and it is not always immediately obvious from statistical hypothesis testing which results are of interest and which are not. Simply put, we do not want to be in a community where researchers are reporting every effect they discover that has a small p-value.

Description of content to be covered

This tutorial will provide two methods of orienting researchers towards discovering more useful relationships in the data. The first approach is conceptually the simplest, and will show researchers how to estimate magnitude of effects from the data. There are several metrics for doing so, such as R2, correlation coefficients, η^2, or ω^2. What they all have in common is that they are easily obtainable from statistics software, and provide an estimate of how large a relationship is. By focusing on relationships that are larger, researchers focus their, and their readers', time on relationships that will have a relatively large impact. Furthermore, most relationships are not of large magnitude, and acquiring more rows of data does not increase the magnitude of a relationship; so the set of large magnitude relationships is a relatively stable target for a research domain to target.

The second approach is a bit more exotic: how to infer causal relationships from observational data. As researchers, we are generally not interested in relationships that happen to co-occur, but are not causally related. For example, suppose students who ask for fewer hints tend to do better on a post test. Such a result would not entail that a successful path to improving test scores would be to remove help facilities from our tutors. A more plausible explanation is that students with less knowledge need more help and also have lower test scores. There have been advances in how to automatically recover such causal information from datasets. One tool, TETRAD, is freely available and provides functionality for uncovering causal relationships in data. This tutorial will explain how to use this tool, and some of the assumptions it makes.

Intended audience

The audience would be those who have collected large data sets, either in terms of number of variables or number of rows. This community is an increasing part of ITS, and due to the colocation with EDM there should be a reasonable number of interested attendees. Note that the issues raised in this tutorial apply to both classic statistical as well as data mining approaches, but will be presented in a statistical framework for conceptual simplicity and broader appeal. The plan is for attendees will perform some sample analyses on data, so it will be helpful if enough people are able to bring laptops such that there are two people per computer, as more than that around a laptop gets a bit unwieldy.

Expected background

Attendees should have performed statistical analyses in the past as this tutorial does not have sufficient time to present a primer on statistics. However, the exact models or tests used is not important, as the ideas presented have analogs in most approaches.

Activities

The main mode of presentation will be lecture, but it will be stronger if attendees are able to practice at least once how to apply the skills.

Expected outcomes

Attendees will learn some of the problems with doing statistical testing on large data sets, and how to focus their effort on discovering relationships that other researchers will find meaningful. Specifically, the tutorial will show researchers how to locate such information while using SPSS (R may be possible

instead, and there are tradeoffs for both options that are still being considered), and attendees will learn how to uncover relationships that are likely to be causal in their data with the TETRAD tool.

Organizer experience

The organizer has presented a well-received tutorial on common pitfalls with statistics at ITS 2010, and teaches graduate courses on empirical methods, and graphical models (covers causal discovery from data).

Facilities needed

An overhead projector is essential. Beyond that, wireless access for attendees for downloading software and datasets. These will be sent out ahead of time, but realistically some people there will need to do it during the tutorial.