Measuring the Effects of Data Mining on Inference

D

Tom Burr
Statistical Sciences, Los Alamos National Laboratory, USA

S. Tobin
Nuclear Nonproliferation and Security, Los Alamos National Laboratory, USA

INTRODUCTION

Many types of data are increasingly available in large quantities. Information science and technology (IS&T) aims to find valuable information from data and corresponding data models. Example databases include data from "frequent shopper" programs, data on which movies are watched by Netflicks customers, and data on credit card usage patterns. Frequent shopper programs develop marketing strategies such as customized food coupons. Netflicks sponsors a "Netflix prize" aimed at predicting a customer's movie preferences on the basis of previous movies watched and possibly also on the basis of previous movie rankings by some of the customers. Credit card companies have many different fraud-detection algorithms, each of which aim to describe patterns of legitimate customer behavior so that fraud behaviors can be rapidly detected.

One key IS&T tool is data mining to discover patterns. Data mining is the more modern term for what was called "exploratory data analysis" in the 1970s. The explosion of available data types and huge data quantities is opening unprecedented data mining opportunities. Two of the most common goals in data mining are regression and classification. Regression is also known as response fitting or function fitting. Classification is also known as pattern recognition or discriminant analysis. To select one or multiple good models, data mining typically examines many possible models to fit training data and reserves testing data and/or uses resampling to mitigate the effects of overfitting to training data. Classical statistical inference ignores the model selection stage, focusing on model parameter estimation as if the model were chosen in advance of data collection. This article examines the effects of data mining on statistical inference, with attention given to resampling strategies, such as the bootstrap, which addresses the model selection stage of analysis.

A common data mining goal is to predict a response y (such as patient outcome) given a collection of p predictor variables (such as patient blood chemistry), as shown in Eq. (1):

$$y = f(x_1, x_2, \ldots, x_p) + \text{error} = f(x) + \text{error}, \quad (1)$$

where x denotes $x_1, x_2, \ldots, x_p$ and error is "what is left unexplained" after accounting for the effects of x on y via the function f. Because such analysis is useful to many fields of research, many approaches exist for fitting $f(x)$. For historical reasons, if $f(x)$ and y are continuous-valued functions, this goal is often called "regression." And if $f(x)$ and y are discrete-valued functions, this goal is often called "classification." This article considers the impact of choosing a good functional form for f and then fitting parameters of the chosen f for the regression application.

The objectives of this article are to include additional background to review several regression function options, describe effects of measurement errors in y and/or x, give example results, describe challenges, and then end with discussion and summary.

BACKGROUND

As a simple example that we use throughout to illustrate and refer to as "Example 1," suppose that a data analyst plans to evaluate three linear models for a set of n observations of $\{y, x_1, x_2\}$. Models M1, M2, and M3 fit the response y as a linear function of predictor x_1, of predictor x_2, or of both predictor x_1 and x_2, respectively. The expected value of y in models M1–M3 are

DOI: 10.4018/978-1-4666-5888-2.ch176

$E(y) = \beta_1 x_1$, $E(y) = \beta_2 x_2$, and $E(y) = \beta_1 x_1 + \beta_2 x_2$, respectively, and each y is assumed to be generated as $y = E(y) + e$, where e is a random error term with mean 0 and variance σ^2. Some type of model selection criterion, such as the adjusted residual sum of squares, can be used to select a model from M1, M2, and M3. The adjusted residual sum of squares is defined as

$$R^2_{\text{adjusted}} = 1 - \frac{n-1}{n-p}(1-R^2),$$

where

$$R^2 = 1 - \frac{SS_{\text{error}}}{SS_y}$$

is the unadjusted

$$R^2,\ SS_{\text{error}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is the error sum of squares,

$$SS_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

is the sum of squared y values around the mean $\overline{y}$ of y, and p is the number of model predictors ($p = 1$ for M1 and M2 and $p = 2$ for M3).

Suppose model M1 is chosen on the basis of the R^2_{adjusted}. To what extent should the analyst accept standard least squares regression results as a way to assign an uncertainty statement to the coefficient β_1? It might surprise readers to know that nearly all statistical texts devote much space to the least squares regression result

$$\text{var}(\hat{\beta}_1) = \frac{x_1^T y \sigma^2}{x_1^T x_1},$$

but a large portion of the variance in the estimate $\hat{\beta}_1$ of β_1 is due to choosing among models M1–M3—which is typically ignored. That is, standard text books condition on the chosen model as if it were chosen in advance of data analysis rather than as part of the data analysis. If instead some type of data mining is applied to choose a model, then the variance in the estimate $\hat{\beta}_1$ depends on what other predictors are in the chosen model, which varies randomly as the value of the model selection criterion changes across realizations of the predictor y.

MAIN FOCUS

Issues

Many options exist for creating the *f(x)* function; eight are described in this section. The options make different assumptions about the form of *f(x)*, such as whether *f(x)* is a linear combination of terms that are functions of *x* or whether *f(x)* can have nonlinear parametric forms with parameters that can be estimated or are smoothly varying but without specifying a particular parametric form. The following seven options are described in Burr et al. (2013), Ripley and Venables (1994), and Hastie et al. (2001). In Burr et al. (2013) the main data mining application supports new measurement options within the field known as NonDestructive assay as described below.

1. **Weighted Least Squares (WLS):** Regression uses a linear combination of predictors only; linear predictors and quadratic predictors; or linear, quadratic, and linear interaction predictors. WLS can be used for the simple Example 1 that we use throughout.
2. **Projection Pursuit Regression (PPR):** Uses a linear combination of nonlinear transformations of linear combinations of explanatory variables, or

$$y = f(x) + e = \beta_0 + \sum_{j=1}^{r} f_j(\beta_j x) + e,$$

where x is the original p predictor variables for a given observation; $\{\beta_j\}$ is a collection of r parameters to be estimated (each a unit vector of length p), e is random error; and r is the number