Data Science Toolkit

Matt Goodwin

October 8, 2018

Contents

1 Statistical Modeling

1.1 Overview and Theory

When talking about modeling it is always important to keep in mind that “All models are wrong but some are useful” (attributed to George Box). There are different approaches to modeling depending on the discipline you come from, but personally I like the idea of the function approximation approach suggested by applied math and statistics. Taking this approach allows is to use probability theory combined with decision theory and to be able to visualize these concepts in a euclidean geometric space.

Bishop has a really nice overview of some of these concepts. The starting point I think for modeling starts with independent variable X and dependent variable Y. We want to know:

1. The nature of the relationship between the variables (inference)

2. Given an independent variable, determine the dependent variable (prediction)

Using probability we start with the joint distribution $P(X, Y)$ which completely summarizes the uncertainty between the two variables. For example, imagine that we had one independent variable and one dependent variable, then $P(X, Y)$ would be a three-dimensional distribution (the z-axis in this case would be the probability density function).

Maybe one of the most simplest models to start with is the additive error model. This is given simply as:

$$Y = f(X) + \epsilon \tag{1.1}$$

1.2 Bias-Variance Tradeoff

The bias-variance tradeoff refers to two sources of error when evaluating models - the bias and the variance. There is also a third source of error which we call the “irreducible error”.

As explained in this article, there is a slight confusion in data science between decomposing the error for an estimator, and decomposing the error for a model or a predictor. The decomposition is really about the same but there are some key insights to be aware of. The decomposition below is for a predictor. The decomposition for an estimator can be found in various books and other resources.

Bias is defined as:

$$y \tag{1.2}$$

1.3 Boosting

The concept of boosting has lead to some of the most powerful algorithms in machine learning. Boosting falls under a general class of algorithms known as ensembles (bagging would be another example of ensemble algorithms where we run separate models and then aggregate at the end by averaging for example). The general concept is that we run a weak learner on the original data, calculate the errors, run a new model on the errors, combine with the first weak learner, and repeat until some stopping criteria (that avoids overfitting).

1.3.1 Adaboost

2 Terms and Notation

2.1 Variable notiation

Below explains notation used commonly when setting-up machine learning models. To help understand the notation I use the example of predicting the sales of ice cream cones.

- X - represents an input variable. Even though input variable implies a single variable this could also be a vector. If we wanted to access a single variable from the input vector then we use notation X_j. So for example X could include variables that describe the temperature (X_j), time or year (X_{j+1}), etc.
- Y - represents a *quantitative* output variable. This could be the sales of ice cream cones in dollars.
- G - represents a *qualitative* output variable. This could be if we sale over 50 ice cream cones for example (yes or no).
- x_i - represents an observed value of the variable X. Again this could be a vector. So to get the observed scalar value of the temperature for example we would write x_{ij}.
- $\boldsymbol{X}$ - matrix typically with dimensions Nxp.
- $\boldsymbol{x_j}$ - in general vectors are not bold unless the distinction is being made that this is the vector of all observation on X_j. So $\boldsymbol{x_j}$ is of length N and x_i is of length p.

Glossary

dummy variable A vector where each element is either 0 or 1 and is used to represent a specific class. For example, if we have K classes then a dummy variable would be of length K and if we wanted to represent class 1, we would have a "1" in the first position in the vector and everywhere else would be 0.. 1

estimator A point estimator as defined by Cassella/Berger is any function $W(X_1, X_2, ..., X_n)$ of a sample. Any statistic is an estimator.. 1, 2

test A categorical variable that has ordering such as low, medium, and high, but no notion of a metric.. 1