DM STAT-1 CONSULTING **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

The Paradox of Overfitting

Bruce Ratner, Ph.D.

Overfitting, a problem akin to model inaccuracy, is as old as model building itself, as it is part and parcel of the modeling process. An overfitted model is one that *approaches reproducing* the training data on which the model is built – by "capitalizing on the idiosyncrasies" of the training data. The model brings about the complexity of the idiosyncrasies by including in the model extra unnecessary variables, interactions, and variable construction(s), none that are part of the sought-after predominant pattern in the data. Resultantly, a major characteristic of an overfitted model: The model has too many variables; it is too complex. Ergo, the overfitted model can be thought of a *too perfect picture* of the predominant pattern, essentially memorizing the training data instead of capturing the desired pattern. As such, individuals of holdout/validation data (drawn from the population of the training data) are strangers who are unacquainted with the training data, and cannot expect to "fit into" the model's perfect picture of the predominant pattern to produce good predictions. When a model's accuracy based on the validation data is "out of the neighborhood" of the model's accuracy based on the training data, the problem is one of overfitting, and the model is said to be an overfitted model. As the fit of the model increases by *including more information*, (seemingly to be a good thing), the model's predictive performance on the validation data decreases. This is the paradox of overfitting.

My Idiomatic Definition of Overfitting to Help Remember the Concept

A model is built to *represent* training data, not to *reproduce* training data. Otherwise, a visitor from validation data will not *feel at home*. The visitor encounters an uncomfortable fit in the model because s/he probabilistically *does not* look like a typical data-point from the training data. The misfit visitor takes a poor prediction. The model is overfitted.

Related to model fitting are the concepts of prediction variance and bias. Variance is a measure of spread/range of a variable. In the present situation, the range is in terms of a confidence interval about the *prediction error*. Overfitted models have a large error variance; the confidence interval about the prediction error is large.

The underfitted model, a non-frequenter model, has too few variables; it is too simple. An underfitted model can be considered as a *poorly rendered picture* of the predominant pattern, essentially without recollection of the training data to capture the desired pattern. As such, individuals of validation data are strangers who have no familiarity with the training data, and cannot expect to fit into the model's portraiture of the predominant pattern to produce good predictions. As overfitted models affect error variance, underfitted models affect error bias. Bias is the difference between the model's prediction and the correct prediction. Underfitted models

have a large error bias; individuals' predictions are wildly far from the correct value. A graphical depiction of this discussion is in Figure 1, below. [1]

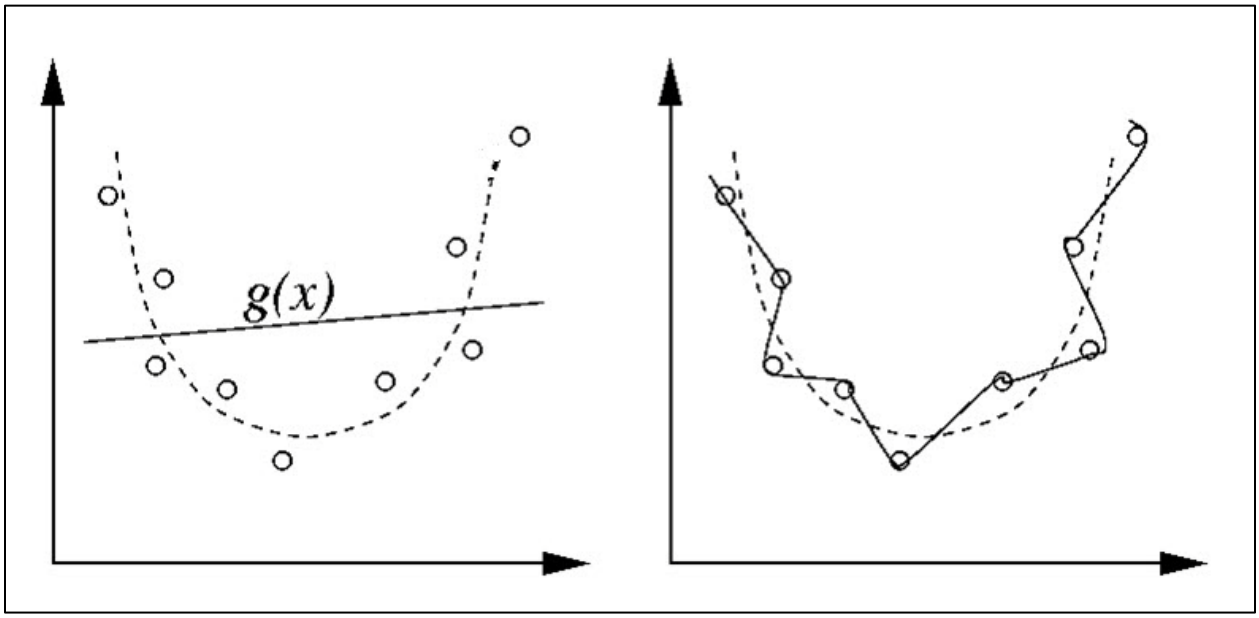

Figure 1: Over-and Under-fitted Models

Consider the two models, the simple model *g(x)* in the left-hand graph, and the *zigzag* model in right-hand graph, in Figure 1. Clearly, I want a model that best represents the predominant-pattern of the parabola as depicted by the data points indicated by circles. I fit the points with straight-line model *g(x)*, using only one variable (too few variables). The model is visibly too simple. It does not do a good job of fitting the data, and would not do well in predicting new data points. This model is underfitted, and has a large error bias.

For the *rough* zigzag model, I fit the data to "hit" every data point by using too many variables. The model does a perfect job a reproducing the data points. The prediction error at each data points is zero. This model would not do a good job of predicting for new data points. The model is utterly overfitted, and has a large error variance. The model does not reflect the obvious *smooth* parabolic pattern. As is plainly evident, I want a model in between *g(x)* and the zigzag models, a model that is powerful enough to represent the apparent pattern of a parabola. The building of the desired model is given to the reader.

It is “fitting” to digress here for guidelines of model building. A well-fitted model is one that *faithfully represents* the sought-after predominant pattern within the data, ignoring the idiosyncrasies in the training data. A well-fitted model is typically defined by a handful of variables because it does not include “idiosyncrasy’ variables. Individuals of validation data, the everyman and everywoman incognizant with the training data, can expect to fit into the model’s faithfully rendered picture of the predominant pattern to produce good predictions. The accuracy of the well-fitted model on validation data will be “within the neighborhood” of the model’s accuracy based on the training data.

Reference: 1 - http://www.willamette.edu/gorr/classes/cs449.html
Related Reading: Predicting the Quality of Your Statistical Regression Models