Function approximation (Regression)

Authors: Rehan Daya, Manan Desai, Samir Epili. (PDF)

Parametric Models

A parametric model is a particular class of statistical models such as linear regression. A parametric model is considered linear if it can be written like this:

$y(x, w) = \sum_{i=0}^{M} w_i \phi_i(x) = w^T \phi(x)$

where $y = T$=Target, $x = \phi$=Transformed input, $\hat{y} = y$=Answer, $\beta = w$=Weight.Specifically, a parametric model is a family of probability distributions that has a finite number of parameters.

Non-parametric Models are statistical models that do not often conform to a normal distribution, as they rely upon continuous data, rather than discrete values. A good example of this is K-Nearest Neighbors.

When making the parametric model we have certain assumptions:

1. Expected value of T given the basis function vector phi is linear in phi.
2. All distributions around the expected values are assumed to be i.i.d. zero mean Gaussian with constant variance.
3. Minimizing Mean Squared Error (MSE) on the training data yields the Maximum Likelihood Estimate (MLE) solution of the assumed generative model.

To ensure the parametric model is accurate we try to minimize the SSE (Sum-of-Squares Error) by utilizing the RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$

This is essentially the difference between the target variable and your prediction. As this value shrinks your R-Squared will increase.

The takeaway here is:

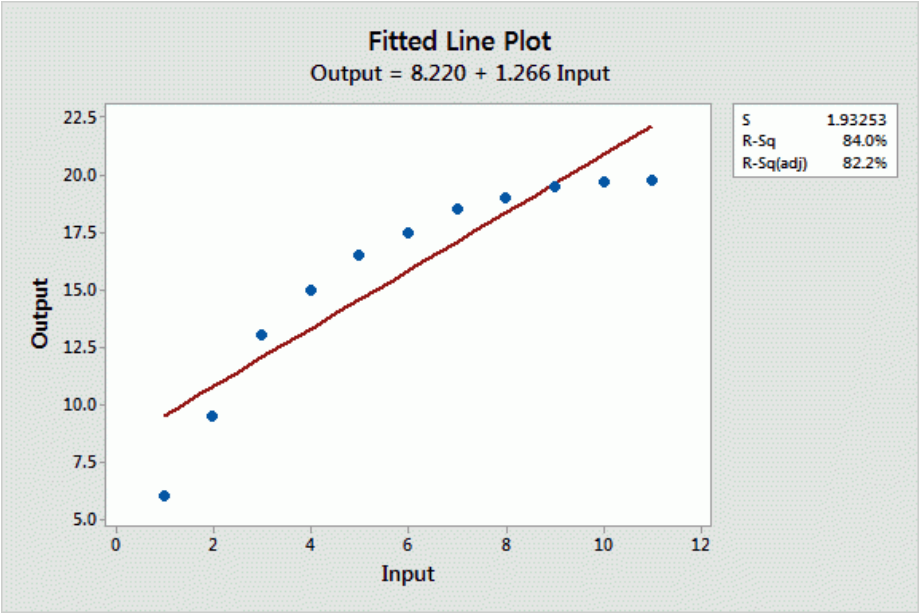
There is an exact closed form solution for the optimal weights for a linear model. Assuming the inverse of a matrix exists, so the determinant cant be 0.

Collinearity

In statistics, multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. We tend to drop a variable that is collinear with another variable and test the RMSE change. Then repeat the step with all other collinear variables to see which combination of variables provides the least amount of collinearity and complexity.

Fitting

Least Squares which uses euclidean distance is utilized to fit the model to the data. A good example is seen below:



You can try higher polynomial functions which will cause odd looking models to form versus this straight line. But the only way to know if it fits well is to look at the out-of-sample RMSE. Many times we look at in-sample RMSE which leads to high bias since the model is trained on the training dataset and tested on it too. Be wary though, that the higher order polynomial your function is, the more complex your model becomes.

A Deeper Dive into Parametric Models

There are several different ways to map input variables to output variables. However, there is a lot standing in the way of making an accurate prediction. In order to simplify the learning process, algorithms can make certain assumptions. These assumptions come in the form of simplifying the function to a known form. This known form has a set of parameters, or weights, of *fixed* size. The number of weights does not change with respect to the size of the data.

Models that make these assumptions are called parametric models. Parametric algorithms take two big steps: 1. The algorithm will pick a form for the function 2. The algorithm will attempt to learn the coefficients for the weights from the training set of data by finding the Maximum Likelihood Estimate solution.

We have seen Maximum Likelihood Estimation already so I won't rehash it.

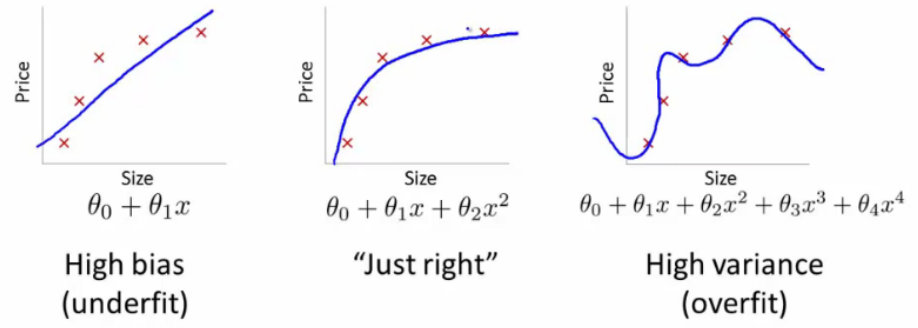
Parametric models come with benefits and drawbacks. The main benefit of parametric models is how simple they are. They are easy to understand, and the existence of weights allows for easier interpretability. Consider, for example, a multiple linear regression model. Once the model is fit, the weights for each input variable clearly show the effect of that variable on the target. Another

benefit that stems from simplicity is speed. These models can be fit very quickly. Additionally, they do not require a large amount of data. These models can work well in simple situations even with limited training data.

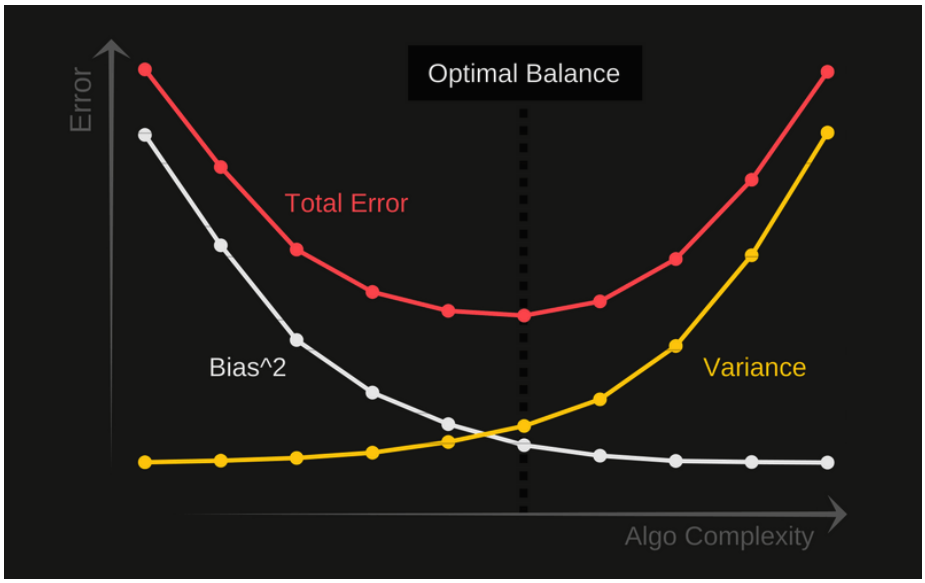
However, there are many downsides to parametric models. Parametric models are, by definition, limited to its parameters and its form. Simply by choosing a form and a number of parameters, the model is already handicapped if the true underlying form of the data is different. This can lead to poor fit in many scenarios. Additionally, parametric models are inherently limited in complexity. There are only so many model forms from which to choose, and once they are exhausted, parametric algorithms have nothing else to try.

A Deeper Dive into Model Complexity and Overfitting

At this point in the course, we have observed that there exists a tradeoff between a more complex model with several features and parameters, and the overfitting of test data. Simply finding the optimal parameters for a model such that every point in a training set is fit perfectly can provide diminishing returns for multiple reasons. The issue for simply fitting the most complex model is the reduction in signal-to-noise ratio. Noise is the randomness present in all raw collected data, while the signal consists of the points whose pattern the model is being used to discern. Because noise is random in nature it is important to select a model that is flexible enough to adapt to different test data, but not to the point that it follows noise and reduces interpretability.



For instance, the image above shows parametric models of increase complexity from left to right. By simply increasing the number of variables, or features, the models can fit data points more accurately. However, the model on the right becomes too sensitive and suffers from high variance due to weighting of the additional variables. Conversely, the model on the left is too simple and suffers from high bias, such as assuming non-linear data is linear. This introduces the bias-variance tradeoff argument. Because we will be discussing this tradeoff more in detail in later lectures, only a brief overview will be provided here.



Since squared bias and variance are inversely related, simpler and more biased models will likely underperform the same as very complex and highly flexible models on test data. Additionally, the total error is the sum of squared bias, variance, and irreducible error. Thus, an optimal model will be at the point where the total error is minimized. Methods for obtaining this optimal regression model include using more data to train more complex models, regularization, and cross validation.