Machine Learning Lesson of the Day – Overfitting and Underfitting

March 20, 2014

By Eric Cai - The Chemical Statistician

(This article was originally published at The Chemical Statistician » Statistics, and syndicated at StatsBlogs.)

Overfitting occurs when a statistical model or machine learning algorithm **captures the noise** of the data. Intuitively, overfitting occurs when the model or the algorithm fits the data too well. Specifically, overfitting occurs if the model or algorithm shows **low bias** but**high variance**. Overfitting is often a result of an excessively complicated model, and it can be prevented by fitting multiple models and using validation or cross-validation to compare their predictive accuracies on test data.

Underfitting occurs when a statistical model or machine learning algorithm **cannot capture the underlying trend** of the data. Intuitively, underfitting occurs when the model or the algorithm does not fit the data well enough. Specifically, underfitting occurs if the model or algorithm shows **low variance** but **high bias**. Underfitting is often a result of an excessively simple model.

Both overfitting and underfitting lead to **poor predictions** on new data sets.

In my experience with statistics and machine learning, I don't encounter underfitting very often. Data sets that are used for predictive modelling nowadays often come with too many predictors, not too few. Nonetheless, when building any model in machine learning for predictive modelling, use validation or cross-validation to assess predictive accuracy – whether you are trying to avoid overfitting or underfitting.