Homework 3-0

Please answer the following questions based on your reading in Chapter 3 of the “Doing Data Science” text, pages 51-70. Brief, but complete answers are best (26 points total).

1. (6pts) Consider the two practices of statistical modeling and machine learning. They share many similarities but do have their differences. State three similarities they share and three differences between them.

Statistical modeling practices involve mathematics to find relationships between variables unlike machine learning which uses computer science and programming languages. Another difference is statistical modeling has been around for a long time whereas machine learning practices have been implemented recently. Statistical modeling is generally used for smaller data sets with less attributes and machine learning is used with a variety of online tools to predict data billions and trillions of observations. They are both alike in the sense that both involve very similar techniques such as linear regressions, vector machines and more. Both these methods are also used to come to the same conclusion or objective, how do we get more insights from our preexisting data sets. Each step in both processes mirror each other for example machine learning uses learning whereas statistics is estimation, or a hypothesis vs a classifier.

1. (5pts) Explain the difference between a model and an algorithm and give an example of each.

A model is given inputs and returns an output whereas an algorithm is used to train a model and will tell one about what to do. An example of a model could be an ATM since its given an input and returns an output. An example of an algorithm could be a sorting algorithm that sorts numbers such as insertion sort.

1. (5pts) Explain in your own words why the residuals that result from fitting a linear regression model should be evenly distributed around zero. What could you do if they aren’t?

They should be distributed evenly around the linear regression since it allows us to see the validity of our assumptions that a linear relationship exists. Any points deviating from the line would be considered lurking variables, in which a time series plot is useful.

1. (5pts) A model that overfits data has low fitting error, but lack the ability to generalize trends from the data. Describe the structure (in terms of its parameters) of a linear regression model that overfits a data set.

The structure of this linear regression would have estimated and actual sampling variances that would be excessively large, the estimator’s accuracy becomes poor with overfit data. These are also free of bias in the parameter estimators.

1. (5pts) Given models A with 5 parameters and model B with 10 parameters, briefly explain why a model with fewer parameters would be preferable given that they both perform equally well predicting new data.

This is simply because it is easier to work with fewer parameters when you can achieve the same level of performance when predicting data.