Contents

Draft

1 Overfitting

Draft

Overfitting describes the error which occurs when a fitted model is too closely fit to a limited set of observations. Overfitting the model generally takes the form of making an overly complex model (i.e. using an excessive amount of independent variables) to explain the behaviour in the data under study.

In reality, the data being studied often has some degree of error or random noise within it. Thus attempting to make the model conform too closely to sample data can undermine the model and reduce its predictive power.

2 Dummy Variables Draft

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0,1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup.

3 Least-Angle Regression (LARS)

Draft

In statistics, least-angle regression (LARS) is an algorithm for fitting linear regression models to high-dimensional data, developed by Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani.[1] Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients. Instead of giving a vector result, the LARS solution consists of a curve denoting the solution for each value of the L1 norm of the parameter vector. The algorithm is similar to forward stepwise regression, but instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual. The advantages of the LARS method are: It is computationally just as fast as forward selection. It produces a full piecewise linear solution path, which is useful in cross-validation or similar attempts to tune the model. If two variables are almost equally correlated with the response, then their coefficients should increase at approximately the same rate. The algorithm thus behaves as intuition would expect, and also is more stable. It is easily modified to produce solutions for other estimators, like the lasso. It is effective in contexts where $p >> n$ (i.e., when the number of dimensions is significantly greater than the number of points)[citation needed]. The disadvantages of the LARS method include: With any amount of noise in the dependent variable and with high dimensional multicollinear independent variables, there is no reason to believe that the selected variables will have a high probability of being the actual underlying causal variables. This problem is not unique to LARS, as it is a general problem with variable selection approaches that seek to find underlying deterministic components. Yet, because LARS is based upon an iterative refitting of the residuals, it would appear to be especially sensitive to the effects of noise. This problem is discussed in detail by Weisberg in the discussion section of the Efron et al. (2004) Annals of Statistics article.[2] Weisberg provides an empirical example based upon re-analysis of data originally used to validate LARS that the variable selection appears to have problems with highly correlated variables. Since almost all high dimensional data in the real world will just by chance exhibit some fair degree of collinearity across at least some variables, the problem that LARS has with correlated variables may limit its application to high dimensional data.