Overfitting

Cook JA, Ranstam J

Overfitting is a key threat to the integrity and reliability of a statistical model. It is common to explore data looking for associations between potential risk factors and the development of a disease or outcome after treatment (e.g. influence of patient characteristics such as age and stage of disease upon recurrence of colorectal cancer after a partial coloctemy). Alternatively it may be desired to seek to build a prognostic model to enable prediction of future events (e.g. recurrence of disease or complications post-surgery). In such settings there are often a multiple of potential factors which could be modelled. However, the desire to include all available factors needs to be tempered. First, the dataset may not be large enough to support a model containing every factor. Second, some of the factors may be redundant and inclusion of them unnecessary and may even adversely impact (e.g. due to multicollinearity) upon the model by obscuring the true relationship between variables. Strategies are available (e.g. pre-selecting factors, centering, cross-validation) to minimise the possibility and impact of overfitting though ultimately replication of findings in another dataset is the only sure way to address the issue. Some more complex analyses (e.g. structural equation modelling) can be used which do not derive all of the information needed to support the analysis from the data; however this can only be done by greater prescription regarding relationships at the outset and the associated increased risk of having a self-fulfilling analysis. To paraphrase Antoine de Saint-Exupéry, an analyst knows he has achieved perfection not when there is nothing left to add to the model, but when there is nothing left to take away.

Word count 273