Challenges with model and tuning parameter selection when using cross-validation with censored data

Anders Munch

June 11, 2022

1 Abstract

Most machine learning algorithms depend on one or more tuning parameters to control the trade-off between bias and variance. To select the optimal value for a tuning parameter, a popular approach is to use cross-validation to determine which value minimizes a given loss function. We consider the problem of selecting a regression model from a collection of candidate models using cross-validation on an external data set where the observations might be right-censored in continuous time. In this situation it is common to use the component of the negative log-likelihood corresponding to the outcome as the loss function. This ignores the contribution to the likelihood made by the censoring distribution, and we argue that this approach is problematic in at least two ways: Firstly, we show that the least false parameter according to this loss function is in general not well-defined as it depends on the censoring distribution. Secondly, for many commonly used survival models the likelihood will a.s. be zero for any hold-out sample. This means that the negative log-likelihood loss cannot be used to compare general survival models and hence does not provide a general approach for model selection in the survival setting. We discuss how these problems can be alleviated by modeling the censoring distribution, which, on the other hand, comes at the cost of introducing a new nuisance parameter to be estimated.

2 Parts of old version

We show how modeling of the censoring distribution can alleviate these problems. On the other hand, this introduces the problem of estimating the

censoring distribution, which without further assumptions is as equally complicated as estimating the outcome model. We make some comments on this and consider the possibility of using other loss functions. Finally, we argue that it is worth considering whether the final goal of the analysis is the selection of a well-performing risk predicting model, or if the model selected is part of the a nuisance parameter estimation step. In the latter case, different strategies for model selection could be considered.