Cox Model in Keras

Byron Smith & Camden Lopez

February 19, 2019

Cox Model Background

Survival analysis is used commonly in many fields including finance and the biomedical sciences. The focus of a survival analysis is the time to an event. Importantly, any observational unit can be 'right-censored' where that unit has been followed for a certain period of time, but whether or not the event happened was not recorded.

Although parametric survival models can be constructed to model the time to event, a Cox model provides a non-parametric method to model a time to event response with a set of covariates. The Cox model assumes that hazards are proportional:

$$\frac{h(t|\vec{x}_1)}{h(t|\vec{x}_2)} = Const.$$

Also, that the log of the hazard function is linearly related to the M covariates as:

$$log[h(t|\vec{x})] = \sum_{j=1}^{M} \beta_j x_j$$

The overall hazard function will be a combination of a 'baseline hazard' similar to an intercept and the hazard function based on the covariates. Here we focus only on the hazard function's relation to the covariates and leave the estimation of the baseline hazard to previously written code. The baseline hazard can be avoided by making use of the first equation and optimizing the partial likelihood function:

$$L(\beta|\vec{x}) = \prod_{i \in C} \frac{\exp(\vec{\beta} \cdot \vec{x_i})}{\sum_{j:t_j \geq t_i} \exp(\vec{\beta} \cdot \vec{x_j})}$$

$$\Rightarrow log(L(\beta|\vec{x})) = l(\beta|\vec{x}) = \sum_{i \in C} \beta \cdot \vec{x_i} - log[\sum_{j:t_j \geq t_i} \exp(\vec{\beta} \cdot \vec{x_j})]$$

Here C is the set of all points that are considered events. The sum in the denominator is taken over all observations with longer than or equal to the amount of follow-up that an observational unit has.

Data Inputs

The standard way that the response is recorded in time-to-event models is to represent each observation by a duple: One value for the follow up time for that observation and a binary value for whether or not the observation hand an event at that time. For example, an observation with the duple (15, 1) may mean that the observation had 15 days of follow up and had an event on the final day. A duple (6, 0) may mean that the observation had 6 days of follow up and then no more information was collected. Note that the unit of time is not important, but every observation must have the same unit of time.

The covariates should be represented by an array. In the code below we use the first index of the array to represent the observation and the other indices to represent covariates of the data.

Dealing with ties

A major problem in Cox modeling is how to deal with observations that have a tied time. Consider the following data:

```
##   time event x
## 1    1     0 2
## 2    2     1 1
## 3    3     1 4
## 4    3     0 5
## 5    4     1 2
```

If the log-partial likelihood is calculated as is based on the formula above, this is known as Breslow's method. Note that one computational difficulty is the fact that the second sum over j will be carried out over observations 3 and 4 twice, because their follow up time is tied. If no ties exist in the data, the computations can be greatly simplified using a cumulative sum.

In most packages, the default adjustment is provided by Efron which modifies the full likelihood term rather than the second sum in order to account for these ties. This method is not currently included in the following code.

Note that dealing with ties makes the use of R more difficult because of the looping required to account for the ties in the current implementation. This is likely due to the fact that python objects can work seamlessly with keras whereas R objects cannot. For this reason, the loss function in R is built through python using the reticulate package.

Difficulties in running Cox models in keras

Keras is a powerful tool for performing optimizations of deep learning models. As it stands, we have not created a modeling method that works with mini-batches of data. Alternative methods for modeling survival data may present a better solution if batch size is an issue for your data set.

The major issue with the Cox partial likelihood is that each observation contributes to the loss function only relative to the other observations (with equal to or longer follow-up) rather than absolutely as in the sum of squares error or cross-entropy. This will result in a biased estimate of the gradient of the loss function when batches are taken over uniformly random samples of the observations.

Future work

Note that many people have published on deep learning survival analysis and that these examples are relatively simple to demonstrate how one might use keras to run a Cox model.

Also note that many, many adaptations to Cox models exist including ones where covariates and/or coefficients can be time-varying. Alternatively one could incorporate competings risks such as death versus tumor growth versus progression-free survival. Finally, regularization penalties can be applied fairly easily.