Model Evaluation

Considerations for Time-to-Event Studies

Daniel Saggau

11/12/2020

Overview

1. Time to Event Studies
2. Classical Model Evaluation: Brier Score and AUC
3. TTS Model Evaluation: IBS and c-index
4. Discussion
   - What measure is most useful for machine learning in TTS?
5. Further Considerations
   - What other methods are coming?

Time-to Event Studies

- Analysis working with (right) censored data
- Right censored data (event after follow up) vs. left censored data (event was not recorded when it occurred initially)
- Highly relevant for clinicians in the field of medical statistics e.g. looking at when a patient dies or when he gets a disease (clinical/epidemiological studies)
- In Economics/Finance e.g. to examine when a subject/borrower will default or when a subject will find/lose a job
- Operations research to predict the time a machine will break

Basic Notations & Concepts

- Time T and Survival S
- From hazard to cumulative hazard to survival
- Hazard h(t,x) is the eminent probability of death a specific point in time
- Capital H is the cumulative hazard
- non-parametric hazard models (KM) vs.semi-parametric proportional hazard model

KM model:

$h(t) = \frac{d}{dt}[logS(t)]$
$H(t) = -log(S(t))$
$S(t) = exp(-H(t))$
Cox model:

$h(t|x\beta) = h_0(t)exp(\beta^T x)$

Classical Model Evaluation Tools for Classification Tasks

1. **Diagnostic vs. Prognostic Study**
2. **What elements do we consider?**
   - Discrimination: Are we able to correctly discriminate between e.g. sick and healthy patients ?
   - Calibration: How concise is our prediction accuracy ?
   - Clinical Usefulness: Will our model create more benefits than harm?
3. **Label vs. Probability**
   - Brier Score (probability from true class label)
   - AUC (label based error measure via specificity and sensitivity)

Brier Score

The score is based on loss function. Other loss measures are the log loss or the integrated log loss. The general version of the brier score looks at a specific point in time. We can plot this brier score via prediction error curves (pec).

Derivation

MSE for Regression (L2 Loss):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)}) - \hat{y}^{(i)})^2$$

Where: the $MSE \in [0; \infty)$
The Brier Score is the MSE for Classification:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}(x^{(i)}) - y^{(i)})^2$$

Confusion Matrix

Components of the ROC

Sensitivity or: true positive rate

- deals with values above the threshold among the subject group which do endure an event

$$TPF = \frac{TP}{TP + FN}$$

Specificity or: true negative rate

- deals with false negatives, hence patients with a disease we classify as not having any diseases

$$TNR = \frac{TN}{TN + FP}$$

Why cant we use traditional model evaluation tools for time to event studies?

- Working with censored data
- Account for time dependent covariates

Early approaches: - excluding subjects with right censored data and only evaluate on the complete data

From AUC to Harell's C-index to time dependent C-index

- Rank correlation measure but still have to deal with censoring
- studying concordance (∼consistency) and discordance (∼inconsistency) pairs

AUC: *"... is individual A likely to have a stroke within the next 5 years?"*
C: *"... is individual A or individual B more likely to have a stroke?"*

Differentiation AUC and C

Intuitively speaking the difference between AUC and c-index is as follows:

$$\mathrm{AUC} = \Pr(\mathrm{Risk_t(i)} > \mathrm{Risk_t(j)} | \mathrm{i\ has\ event\ before\ t\ and\ j\ has\ event\ after\ t})$$

$$\mathrm{C} = \Pr(\mathrm{Risk_t(i)} > \mathrm{Risk_t(j)} | \mathrm{i\ has\ event\ before\ t})$$

c-index

- addressing right censored data via inverse of the probability of censoring weighted estimate (of concordance probability)
- Kendall rank correlation coefficient test as inspiration

Definitions of c-index

Further, we could relabel those terms for the C as:

$$\frac{\#\text{Concordant Pairs}}{\#\text{Concordant Pairs} + \#\text{Discordant Pairs}}$$

Mathematically, we can define e.g. the C for time dependent covariates as:

$$C^{td} = \frac{\Pr(z(X_i) > z(X_j) \& T_i < T_j \& D_i = 1}{\Pr(Ti < Tj | D_i = 1)}$$

IBS

- Called cumulative predictive error curves == continuous ranked probability score (crps)
- Area under the prediction error curve
- Where L is a loss function of the S(the probability that the event of interest has not taken place yet) and time

Mathematics of the IBS

Mean population

Where:

- N = Number of observations
- S_i is the predicted survival function
- t is the time of the event (death) and t* the time before death

(integrated==F) at specific time point

$$L(S, t|t^*) = \frac{1}{N} \sum_{i=1}^{N} L(S_i, t_i|t^*)$$

(integrated == T) over time

$$L(S, t|t^*) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} L(S_i, t_i|t^*)$$

Coding Settup

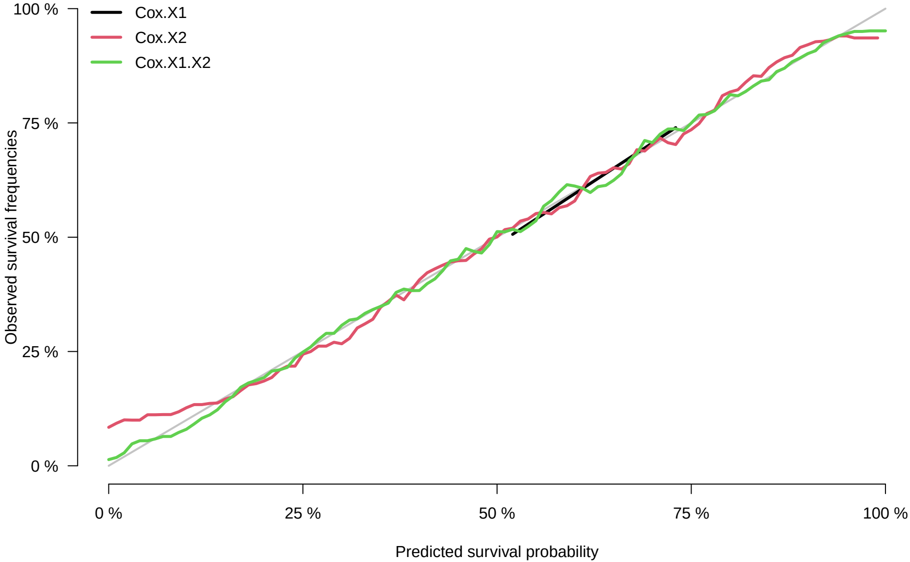
```r
set.seed(123)
dat=SimSurv(10000)
models <-  list("Cox.X1"=coxph(Surv(time,status)~X1,
                      data=dat, x=TRUE,y=TRUE),
               "Cox.X2"=coxph(Surv(time,status)~X2,
                      data=dat,x=TRUE,y=TRUE),
               "Cox.X1.X2"=coxph(Surv(time,status)~X1+X2,
                      data=dat,x=TRUE,y=TRUE))
```

Defining the prediction error based on the brier score

```
perror <- pec(object=models,
                 formula=Surv(time,status)~1,
              data=dat,
                 exact=TRUE, cens.model="marginal",
              splitMethod="none",
                 B=0,# number boostrap samples
                 verbose=TRUE)
```

Calibration Plot

```
calPlot(models)
```



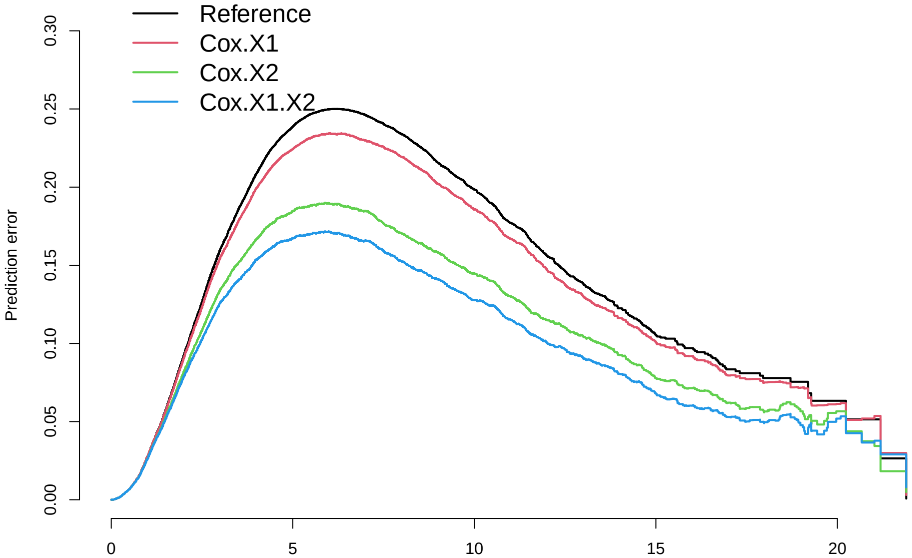
Summary Prediction Error Curve

```r
summary(perror,times= quantile(dat$time[dat$status==1], c(.25, .5, .75,1)))
```

```
##
## Prediction error curves
##
##
## No data splitting: either apparent or independent test sample performance
##
##  AppErr
##      time n.risk Reference Cox.X1 Cox.X2 Cox.X1.X2
## 1  2.568   7892     0.132  0.128  0.112     0.106
## 2  4.270   5644     0.220  0.208  0.174     0.159
## 3  6.513   3179     0.249  0.233  0.188     0.169
## 4 21.189      1     0.026  0.030  0.018     0.029
```

Plotting prediction error

```r
plot(perror)
```



Cumulative Prediction Error

Components of Cumulative Prediction Error Score (IBS)

```
crps(perror,times= quantile(dat$time[dat$status==1], c(.25, .5, .75, 1)))
```

```
##
## Integrated Brier score (crps):
##
##            IBS[0;time=2.6) IBS[0;time=4.3) IBS[0;time=6.5) IBS[0;time=21.2)
## Reference            0.051           0.102           0.150            0.142
## Cox.X1               0.050           0.099           0.143            0.134
## Cox.X2               0.046           0.086           0.120            0.108
## Cox.X1.X2            0.044           0.081           0.111            0.097
```

```
# ibs(perror,times= quantile(dat$time[dat$status==1], c(.25, .5, .75, 1)))
```

cindex Implementation

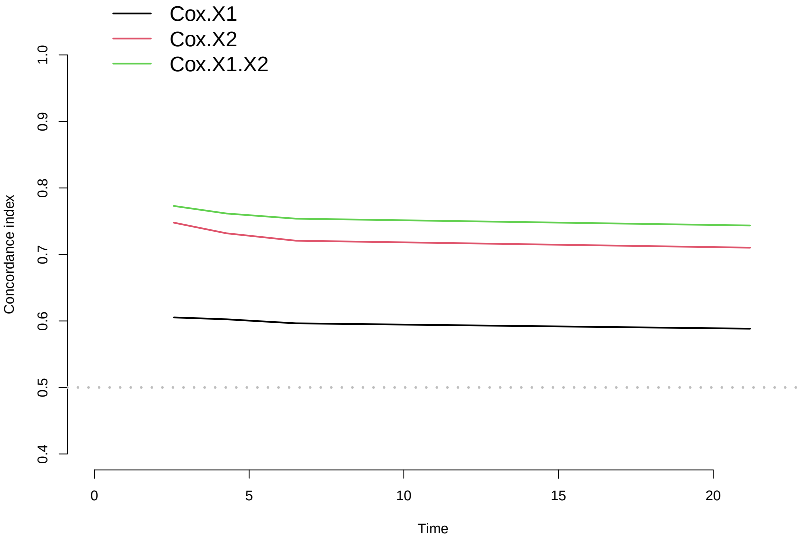
Components of the c-index function

```
cindex = cindex(models, formula = Surv(time,status) ~ 1,
     cens.model="marginal", data = dat,
      eval.times= quantile(dat$time[dat$status==1], c(.25, .5, .75,1)))
```

- **formula** is our survival formula (Surv(time,status)~x1+x2 for cens.model="cox" or Surv(time,status)~1 for cens.model ="marginal")
- **cens.model** is our method for estimating the inverse probability of censoring weights (e.g. cox, marginal, nonpar)
- **splitMethod** is the internal validation design, B the number of boostrap samples & M the size of the boostrap sample
- Extensions: **cause** used for competing risks (default is the first state of the response)

c-index plot

```r
plot(cindex)
```



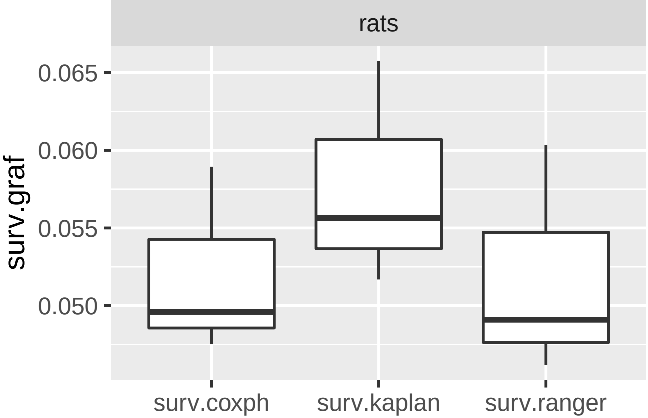
mlr3Proba

Methods based on the loss function:

- Integrated Graf Score (other Name for IBS based on Author Graf)
- Integrated Log Loss (surpress scale of variation)
- Log Loss (censored data ignored)

mlr3Proba Example

```r
##' measure = msr("surv.graf") # for c-index you can use surv.cindex
##' bmr = benchmark(benchmark_grid(task, learners, rsmp("cv", folds = 3)))
##' bmr$aggregate(measure)
autoplot(bmr, measure = measure)
```



Discussion

- c-index has gained popularity because of it's interpretability
- Integrated Brier Score accounts for both calibration and discrimination
- Irrespective, neither model accounts and leaves room for improvement
- IBS allows for differentiation of 'useless' and 'harmful'
- Estimators can be influenced by data
- Clinical consequences problematic

Novel Research

- Decision Curve Analysis (clinical consequences): plotting different exchange rates with the net benefit equation
- Net Reclassification Improvement (clinical consequences)
- Other estimators like SVM estimators for the evaluations tools for the censored data
- IPA
- Competing Risks

Conclusion

- There are various different modifications for model evaluation, neither being superior
- The Brier Score and the AUC are pivotal for many of these methods
- While there has been a lot of research on this topic, the debate is on going

Literature and Recommendations

Introduction:

- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128.
- Blanche, P., Kattan, M. W., & Gerds, T. A. (2019). The c-index is not proper for the evaluation of-year predicted risks. Biostatistics, 20(2), 347-357.

Modifications:

- Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 183-192).

Use Cases:

https://rpubs.com/kaz_yos/survival-auc https://datascienceplus.com/time-dependent-roc-for-survival-prediction-models-in-r/
https://rdrr.io/cran/pec/ https://adibender.github.io/pammtools/