Loss functions and cross-validation with censored survival data

Anders Munch
joint work with Thomas Gerds

PhD Student, Section of Biostatistics
University of Copenhagen

September 7, 2022 – JICI

[Intro?]

Data structure and target of inference

Survival setting

$O = (\tilde{T}, \Delta, X) \sim P \in \mathcal{P}$ Oberved data with $\mathcal{O} = \mathbb{R}_+ \times \{0, 1\} \times \mathbb{R}^p$.

$Z = (T, X) \sim Q \in \mathcal{Q}$ The distribution Q (or a feature of it) is of interest.

Parameters of interest

- Low-dimensional feature of Q, e.g., the marginal survival probability $Q(T > t)$ for a fixed time horizon $t \in \mathbb{R}_+$.
- The conditional survival probability at a fixed time horizon, $x \mapsto S(t \mid x)$ for $x \in \mathbb{R}^p$, with $S(t \mid x) = Q(T > t \mid X = x)$.

The distribution Q is identifiable from the observed data distribution P under coarsening at random. Without further assumptions we would typically need to estimate the conditional survival function S for both problems.

Cross-validation and Super Learning for S

Most machine learning methods depends on one or more hyperparameters which is typically chosen using **cross-validation**.

More generally, to build robust estimators we can use **stacked regression** or **Super Learning** [Breiman, 1996, van der Laan et al., 2007] to select from or combine a collection candidate estimators.

A central component for both cross-validation and Super Learning is the partitioning of data into training and test folds. A suitable loss function is then used to evaluate the performance of an estimator in hold-out samples.

Evaluate performance in hold-out samples

Let $\mathcal{E}$ be a collection of estimators of $S \in \mathcal{S}$. Each $\nu \in \mathcal{E}$ is a mapping $\mathcal{D} \mapsto \nu(\mathcal{D}) = \hat{S} \in \mathcal{S}$, where $\mathcal{D} = (O_1, \ldots, O_n)$ is a data set and $\hat{S}$ is an estimate of the survival function S. Let $L \colon \mathcal{S} \times \mathcal{O} \to \mathbb{R}_+$ be a loss function.

Let $\mathcal{D}_1, \ldots, \mathcal{D}_K$ be a (random) partition of the data set D and let $\mathcal{D}_{-k} := \mathcal{D} \setminus \mathcal{D}_k$, for $k = 1, \ldots, K$. To evaluate the performance of an estimator $\nu \in \mathcal{E}$ we calculate for all $k = 1, \ldots, K$,

$$L(\nu(\mathcal{D}_{-k}), O_i), \quad \text{for all} \quad O_i \in \mathcal{D}_k.$$

Averaging these values across all observations O_i and folds $\mathcal{D}_n$ gives us an estimate of the average loss (risk) of the estimator. We repeat this for all $\nu \in \mathcal{E}$ and pick the estimator with lowest risk. Alternatively, we can use these value as inputs for a meta learner and combine all the estimators into a Super Learner.

The partial likelihood and hold-out samples

A popular choice of loss function for training survival models is the negative partial log-likelihood. Under coarsening at random and non-informative censoring the likelihood for the observed data factorizes as

$$\ell(P, O) = \ell_t(S, O) \cdot \ell_c(G, O) \cdot \ell_0(\mu, O),$$

where $G \in \mathcal{G}$ denotes the censoring mechanism and μ the marginal distribution of the baseline covariates. The negative partial log-likelihood for the component S is

$$-\log \ell_t(S, O) = -\left\{(1 - \Delta) \log S(\tilde{T} \mid X) + \Delta \log f_S(\tilde{T} \mid X)\right\},$$

where f_S is the conditional density or pmf corresponding to S.

However, for many common survival estimators this loss function is unsuitable for evaluating performance in hold-out samples as (a.s.)

$$f_{\hat{S}}(\tilde{T}_i \mid X_i) = 0 \quad \text{when} \quad \hat{S} = \nu(\mathcal{D}_{-k}) \quad \text{and} \quad (\tilde{T}_i, \Delta_i, X_i) \in \mathcal{D}_k.$$

[Hold-out sample illustration]

The Kullback-Leibler divergence and the partial likelihood

Inverse probability of censoring weighted loss functions

A conceptually more attractive (and necessary) strategy is to

(i) use a loss function better suited for evaluating the performance of an estimator of the *survival function* (and not its density), and

(ii) use a loss function defined for in terms of the distribution Q of interest and not P.

One example could be the Brier score

$$L_{\mathrm{Brier}}(S, Z) = \left(S(t \mid X) - \mathbb{1}\{T > t\}\right)^2, \quad Z = (T, X) \sim Q.$$

We can identify the risk of such a loss function using inverse probability of censoring weights (IPCW) [Graf et al., 1999, Gerds and Schumacher, 2006, van der Laan and Dudoit, 2003], as

$$\mathbb{E}_Q\left[L_{\mathrm{Brier}}(S, Z)\right] = \mathbb{E}_P\left[W_G \cdot L_{\mathrm{Brier}}(S, Z)\right],$$

with

$$W_G = \frac{\mathbb{1}\{\tilde{T} > t\} + \mathbb{1}\{\tilde{T} \leq t\}\Delta}{G(\tilde{T} \wedge t \mid X)},$$

where G is the conditional “survivor” function for the censoring distribution.

[Iteration / loop]

Estimation of the conditional "survivor" function for the censoring, G, is also a survival problem in the sense that the event time of interest is now observed when $\Delta = 0$ an only partly observed when $\Delta = 1$. Hence we could use any estimator in $\mathcal{E}$ and apply it to the data set with observations $(\tilde{T}_i, 1 - \Delta_i, X_i)$ to get and estimator of G.

$$
\begin{tikzcd}
\hat{S} \arrow[r] & W_{\hat{G}} \arrow[d] \\
W_{\hat{S}} \arrow[u] & \hat{G} \arrow[l]
\end{tikzcd}
$$

The conditional survivor function as nuisance parameter

Consider now the situation where we want to estimate a low dimensional feature of Q; as example we take the marginal survival at a fixed time point, $Q(T > t)$. Under coarsening at random and a positivity assumption we can write

$$Q(T > t) = \Psi(P), \quad \text{where} \quad \Psi(P) = \mathbb{E}_P \left[S(t \mid X) \right],$$

where S denotes the conditional survival function identifiable from P.

As S is not of interest in itself, we might hope to be able to side-step the issue of finding a suitable loss function by focusing directly of the target parameter instead.

Double robustness

Many estimators based on the efficient influence function has a double robustness property. For instance, the efficient influence function of Ψ is $\psi(O, P) = \varphi(O, S_P, G_P) - \Psi(P)$, with

$$\varphi(O, S, G) = S(t \mid X) \left(1 - \int_0^t \frac{N(\mathrm{d}u) - \mathbb{1}\{\tilde{T} \geq u\}\Lambda_S(\mathrm{d}u \mid X)}{G(u \mid X)S(u \mid X)} \right),$$

where $N(u) = \mathbb{1}\{\tilde{T} \leq u, \Delta = 1\}$ is the counting process and Λ_S is the conditional cumulative hazard corresponding to S. It holds that

$$\mathbb{E}_P \left[\varphi(O, S_P, G_*)\right] = \mathbb{E}_P \left[\varphi(O, S_*, G_P)\right] = \Psi(P),$$

for any S_* and G_*, where S_P and G_P are the conditional survivor functions of the data generating distribution.

This motivates estimating $\Psi(P)$ with

$$\hat{\Psi} = \frac{1}{n} \sum_{i=1}^{n} \varphi(O_i, \hat{S}, \hat{G}),$$

which is consistent if either $\hat{S}$ or $\hat{G}$ is consistent.

Fluctuation risk – exploiting double robustness

Let $\mathcal{G}$ be a (finite) collection of models for G. The double robustness property implies that $\mathbb{E}_P\left[\varphi(O, S_P, G)\right] = \mathbb{E}_P\left[\varphi(O, S_P, G')\right]$ for any $G, G' \in \mathcal{G}$. In particular,

$$\max_{G, G' \in \mathcal{G}} \left| \mathbb{E}_P\left[\varphi(O, S_P, G)\right] - \mathbb{E}_P\left[\varphi(O, S_P, G')\right] \right| = 0.$$

This motivates the "fluctuation risk",

$$R(S) = \max_{G, G' \in \mathcal{G}} \left| \mathbb{E}_P\left[\varphi(O, S, G)\right] - \mathbb{E}_P\left[\varphi(O, S, G')\right] \right|.$$

Let $\mathcal{E}_c$ be a collection of estimators of G. For any $\nu \in \mathcal{E}$, $\gamma \in \mathcal{E}_c$, and $k = 1, \dots, K$ define

$$\hat{\Psi}^k_{\nu,\gamma} = \frac{1}{|\mathcal{D}_k|} \sum_{O \in \mathcal{D}_k} \varphi(O, \nu(\mathcal{D}_{-k}), \gamma(\mathcal{D}_{-k})).$$

For any $\nu \in \mathcal{E}$ we approximate the fluctuation risk with

$$\hat{R}(\nu) = \frac{1}{K} \sum_{k=1}^{K} \max_{\gamma, \gamma' \in \mathcal{E}_c} \left| \hat{\Psi}^k_{\nu,\gamma} - \hat{\Psi}^k_{\nu,\gamma'} \right|.$$

[Illustration of the method]

[Theoretical results??]

[Compare to pre-selected estimators]

Also shows

The conditional survivor function as target parameter

Compare a finite collection of models

Training models on IPCW'ed data

References

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6): 1029–1040, 2006.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18 (17-18):2529–2545, 1999.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.