The negative log-likelihood loss and cross-validation with censored data

Anders Munch

PhD Student, Section of Biostatistics
University of Copenhagen

June 13, 2022

Outline

Problem setting: Model and hyperparamter selection for survival model

The least false model in the presence of censoring

Hold-out samples and survival model estimators

Selecting a model from a collection of candidate models

Let $\mathcal{P}$ denote a collection of probability measures on the sample space $\mathcal{O}$. Let $\mathcal{V}$ denote a parameter space and $L\colon \mathcal{V} \times \mathcal{O} \to \mathbb{R}_+$ a loss function. Consider

$$\nu(P) := \operatorname*{argmin}_{\tilde{\nu}\in\mathcal{V}} P[L(\tilde{\nu}, \cdot)], \quad \text{where} \quad P[f] := \int_{\mathcal{O}} f(o)P(\mathrm{d}o).$$

We approximate P with the empirical measure $\hat{\mathbb{P}}_n$, as $\hat{\mathbb{P}}_n[L(\tilde{\nu}, \cdot)] \approx P[L(\tilde{\nu}, \cdot)]$.

Maximum likelihood estimator (MLE)

For $\mathcal{V}$ a collection of densities and $L(\nu, O) := -\log(\nu(O))$, $\nu(\hat{\mathbb{P}}_n)$ is the MLE.

Hyper-parameter selection

For estimation in high-dimensional settings we often introduce a regularization parameter ν (e.g., LASSO, kernel smoothing). Each choice of ν gives us an estimator, say $\hat{f}_\nu$, and we select the optimal choice of ν using cross-validation,

$$\operatorname*{argmin}_{\nu\in\mathcal{V}} \hat{\mathbb{P}}_n[L(\hat{f}_\nu, \cdot)], \quad \text{where} \quad \hat{\mathbb{P}}_n \perp\!\!\!\perp \hat{f}_\nu.$$

Also useful for combining models [Breiman, 1996, van der Laan et al., 2007].

Survival data

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

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

Assuming coarsening at random [Gill et al., 1997] we can write

$$\mathcal{P} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\},$$

where $\mathcal{G}$ denotes a collection of conditional distributions for the censoring mechanism. Assuming also non-informative censoring the likelihood factorises as $\ell(P_{Q,G}, O) = \ell_F(Q, O) \cdot \ell_C(G, O)$, with

$$\ell_F(Q, O) := q(\tilde{T} \mid X)^{\Delta} \bar{Q}(\tilde{T} \mid X)^{1-\Delta} m(X),$$

where q and $\bar{Q}$ are the conditional density and survivor function, respectively, and m the marginal distribution of X.

Natural to use $-\log \ell_F$ as loss function, or only the first part

$$-\left\{\Delta \log q(\tilde{T} \mid X) - (1 - \Delta) \log \bar{Q}(\tilde{T} \mid X)\right\}.$$

Kullback-Leibler divergence for factorizing likelihoods

Maximum likelihood estimation is closely connected to minimizing the Kullback-Leibler divergence,

$$D_{\mathrm{KL}}(P_1 \,||\, P_2) := P_1\left[\log \frac{p_1}{p_2}\right], \quad \text{where} \quad P_1 = p_1 \cdot \mu, P_2 = p_2 \cdot \mu.$$

By Jensen's inequality $D_{\mathrm{KL}} \geq 0$ and equals 0 when $P_1 = P_2$. Under regulartity condtions, the limit of the MLE under the model $\mathcal{P}_* \subset \mathcal{P}$, when $O \sim P_0$, is the minimizer of

$$P \longmapsto D_{\mathrm{KL}}(P_0 \,||\, P), \quad \text{with} \quad P \in \mathcal{P}_*.$$

If $P_0 \notin \mathcal{P}_*$ the minimizer is referred to as the *least false model*.

If the likelihood for the model $P_{\nu,\gamma}$ factorises with respect to the parameters ν and γ and we do MLE for the *partial* likelihood for ν, when $O \sim P_{\nu_0,\gamma_0}$, the limit is the minimizer of

$$\nu \longmapsto D_{\mathrm{KL}}(P_{\nu_0,\gamma_0} \,||\, P_{\nu,\gamma_0}), \quad \text{with} \quad \nu \in \mathcal{V}.$$

For any value $\gamma \in \Gamma$ we have that $D_{\mathrm{KL}}(P_{\nu_0,\gamma} \,||\, P_{\nu_0,\gamma}) = 0$, so ν_0 is optimal for any γ. However, if $\nu_0 \notin \mathcal{V}$ the minimizer might depend on the value of γ.

Least false model depends on the censoring distribution

A special case of this is the survival setting. Consider the simple case with no covariates so and loss function is $-\log \ell_F$.

Ranking models according to their average loss with respect to this loss function is equivalent to ranking them according to $D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q,G})$ when $O \sim P_{Q_0,G}$.

Let $Q_0, Q \in \mathcal{Q}$ with $Q_0 \neq Q$ and $G \in \mathcal{G}$ be given. Then (under regularity conditions) we can find $\tilde{Q} \in \mathcal{Q}$ and $\tilde{G} \in \mathcal{G}$ such that

$$D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q,G}) < D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{\tilde{Q},G}),$$

and

$$D_{\mathrm{KL}}(P_{Q_0,\tilde{G}} \,||\, P_{Q,\tilde{G}}) > D_{\mathrm{KL}}(P_{Q_0,\tilde{G}} \,||\, P_{\tilde{Q},\tilde{G}}).$$

Proof.

Construct $\tilde{Q}$ such that it performs better than Q_1 on $[0, t]$ but worse on (t, ∞). Construct $\tilde{G}$ such that observations on (t, ∞) are less likely than under G. $\square$

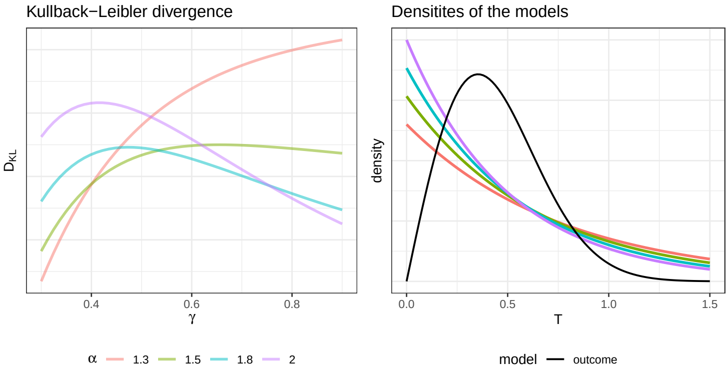
A simple example

Consider four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3, 1.5, 1.8, 2\},$$

and let

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma).$$



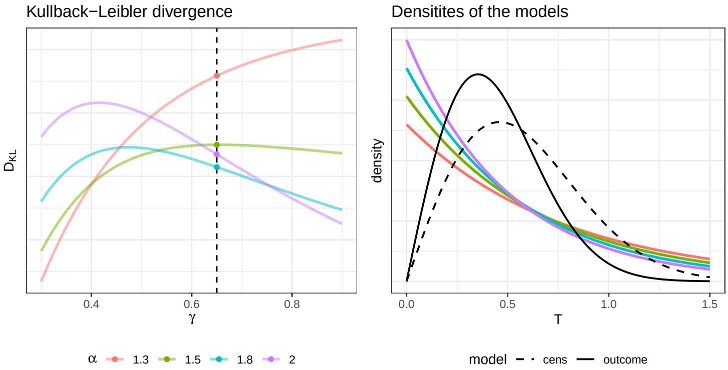
A simple example

Consider four candidate models indexed by α,

$$Q_\alpha = \mathsf{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3, 1.5, 1.8, 2\},$$

and let

$$Q_0 = \mathsf{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \mathsf{Weibull}(2, \gamma).$$



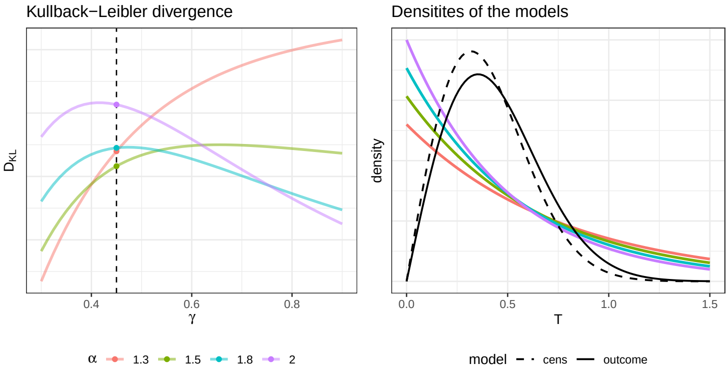
A simple example

Consider four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3, 1.5, 1.8, 2\},$$

and let

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma).$$



Breaker

Survival curve estimators evaluated on hold-out samples

[sort of: ignores this and proceeding anyway. . .]

Modeling the censoring

An (infinite?) loop

References

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

R. D. Gill, M. J. Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

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