6.S897/HST.956 Machine Learning for Healthcare Feb 22, 2019

Recitation 2: Survival Analysis

Instructors: David Sontag, Peter Szolovits *Notes: Irene Chen*

1 Why Survival Analysis?

In the context of risk stratification, survival analysis asks one way to analyze patients: **how long until an adverse event for each patient?** Unlike binary prediction problems, we are predicting a continuous time to event. Survival analysis is more akin to regression tasks then. Unlike canonical regression problems, we would like to include two types of training data: patients with observed outcomes and censored data.

Put another way, we may have data for when a patient died (e.g. in the hospital or cross-referencing with other records) and we may not know if the patient has died but we know the last time we have data on file. If we only restrict to patients for whom we observe death, we may exclude a large portion of the dataset. How can we then incorporate this **censored data**, aka data where we do not observe the adverse event but know if it did happen, the adverse event happened *after* a certain time?

Survival analysis has many potential applications:

- Consider patients with **heart failure**. We know as time t goes to infinity, probability of death is 1, but we are unsure of when. Furthermore, patients may leave the healthcare system for which we have records. For a given patient, we would like to estimate the time until death using as much available data as possible.
- As mentioned in recitation, **laptop malfunction** is a cause for concern for students. We would like to estimate then how long until my new laptop will malfunction, given a dataset of other new laptops and time until malfunction. If we only used data from laptops that had recorded malfunction times, we greatly reduce our dataset. Instead, we want to use the fact that my friend Rebecca has had the same laptop for four years without a malfunction. Intuitively, we would expect that information to push up the estimated survival time of my laptop.

2 Setup

In this recitation, we assume that data is *right-censored* as in we don't necessarily observe the event in question (e.g. mortality). Although researchers are often interested in *left-censored* data as well (e.g. some early input features not observed), we restrict our analysis to right-censorship.

Similar to [WLR17], we define the following terms:

- For each patient i, we have data (X_i, y_i, δ_i) where X_i is the feature vector for patient i, δ_i is a binary variable representing whether patient i was censored or the adverse event was observed, and depending on δ_i, y_i represents the observed event time T_i or the censoring time C_i (e.g. last time saw alive), and C_i represents whether the patient had an observed adverse event or was censored.
- **Survival function** $S(t)$ is the probability that the random variable T, aka time to the adverse event, is greater than a specific time t: $S(t) = P(T \geq t)$
- **Cumulative death distribution** $F(t)$ represents the opposite of the survival function, or the probability that the time of event T occurs before some specified time t: $F(t) = P(T < t) = 1 - S(t)$
- The **death density function** is then the probability of death, or the derivative of the cumulative rate of death at time t: $f(t) = \frac{d}{t}F(t)$ for continuous cases

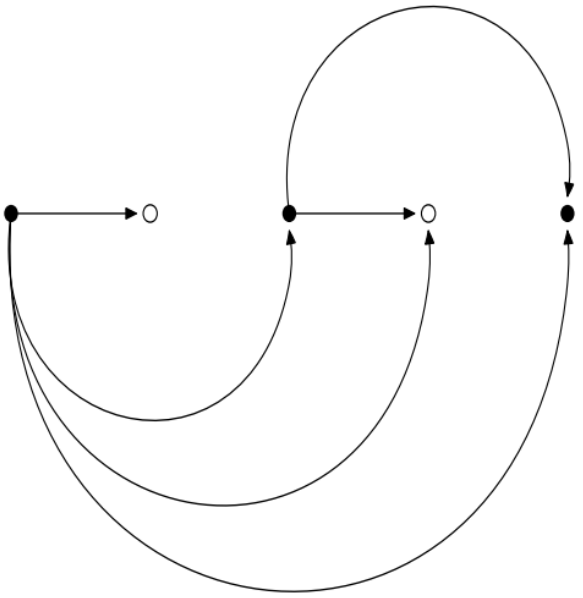

Figure 1: Example of graph to illustrate concordance index, Figure 1b from [SKDo$^+$08]. Open dot refers to censored data (adverse event not observed) whereas filled dot refers to observed adverse events. Dots are sorted left to right such that $y_1 < y_2 < y_3 < y_4 < y_5$ where y_i is the survival time of individual i. Edges are draw from y_i to y_j if $i < j$ and i is a filled dot (observed adverse event).

- The **hazard function** is the instantaneous death rate. Note that it is *not* a probability and instead a rate of event at time t that no event occurred before time t: $h(t) = \frac{f(t)}{S(t)}$

Concordance Index One traditional way of measuring the performance of a survival analysis function S is the concordance index. Similar to the area under the receiver operator curve (AUC), the concordance index (CI) assess the calibration. Put another way, it analyzes how accurate the given scores for a model are.

Concordance index can be measured with the following algorithm

1. Create graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ containing one node for each patient i.
2. Sort patients such that $y_1 < ... < y_N$. Draw a directed edge from patient i to patient j if $\delta_i = 1$ (patient i has an observed adverse event) and $i < j$.

Intuitively we are drawing edges for relations about which we are certain. Patients with $\delta_i = 0$ are censored (no observed adverse event), so we know the patient lived at least as long as y_i but potentially longer. Patients with $\delta_i = 1$ (observed adverse event) have a definitive time, so we know that the survival time is less than any patient j with $y_i < y_j$.

We can then define concordance index $c(\mathcal{G}, f)$ for graph $\mathcal{G}$ formed from data (X_i, y_i, δ_i) and survival function f which predicts the survival time for each patient.

$$c(\mathcal{G}, f) = \frac{1}{|\mathcal{E}|} \sum_{\mathcal{E}_{ij}} \mathbf{1}_{f(x_i) < f(x_j)}$$

That is, of the edges that we are sure about, how many of them have the predicted survival time in the right pairwise relation $f(x_i) < f(x_j)$? Let's sum over all of the edges and compute the average.

3 Kaplan-Meier

The Kaplan-Meier (KM) Curve is one of the most popular methods of graphing survival analysis. Note that the method is non-parametric and assumes no underlying distribution for the event times. Given our data (X_i, y_i, δ_i), we ignore X_i because the KM Curve only uses the recorded times y_i and whether or not a time indicates an observed adverse event δ_i.

We can plot the KM Curve by:

1. Sort the times y_i such that $y_1 < \ldots < y_N$. Note that we assume no two times are the same.
2. For each specific time $(j = 1, 2, \ldots, K)$, assess the number of patients who are at risk r_j and the total number of observed patients n_j. Note then that the number of dead patients is $d_j = r_j - n_j$.
3. The conditional probability of survival beyond time y_j is then the number of individuals surviving longer than t divided by the total number of individuals studied

$$p(y_j) = \frac{d_j}{r_j}$$

4. The product limit is then used to estimate survival time S_t

$$S_t = \prod_{t_i \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

5. Confidence intervals for KM curves come from variance [Gre26] calculated from an asymptotic maximum likelihood solution [KP11].

$$\text{Var}(S_t) = S_t^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

4 Cox Proportional Hazards

Semi-parametric models allow for a more consistent estimator while still not requiring knowledge about the underlying distribution of time to event. Recall that we are given data (X_i, y_i, δ_i). The Cox model hazard function $h(t, X_i)$ follows the proportional hazards assumption given by

$$h(t, X_i) = h_0(t) \exp(X_i \beta)$$

for patient i. The proportional hazards assumption says that any increase in the covariate vector X_i results in a proportional increase in hazard. The model is semi-parameteric because the baseline hazard function $h_0(t)$ is unspecified.

We can see this proportional hazards assumption by comparing two instances X_1 and X_2.

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t) \exp(X_1 \beta)}{h_0(t) \exp(X_2 \beta)} = \exp((X_1 - X_2)\beta)$$

As we will see in the coding section, we can test out this proporitonal hazards assumption to ensure we are using the right model for the data.

References

[Gre26] Major Greenwood. The natural duration of cancer (report on public health and medical subjects no 33). *London: Stationery Office*, 1926.

[KP11] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[SKDo+08] Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216, 2008.

[WLR17] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*, 2017.