11.1 and 11.2 Survival and Censoring

Survival Analysis

- Concerns a special type of outcome variable: the time until an event occurs.
- Ex: suppose we have conducted a five year medial study, in which patients have been treated for cancer.
- We would like to fit a model to predict survival time, using features such as baseline health measurements or type of treatment.
- This sounds like a regression problem, but some of the patients' will have survived to the end of the study.
  - These patients' survival time is said to be *censored*.
- We do not want to discard this subset of surviving patients, since their survival itself is valuable information.

Non-Medical Example

- Consider a company that wants to model *churn*, the event when customers cancel a service subscription.
- The company might collect data on customers over some time period.
- However, not all customers will have cancelled during this time period, so their time to cancellation is censored.
- Survival analysis is well-studied in statistics, but has received relatively little attention in the machine learning community.

Survival and Censoring Times

- For each individual, suppose there is a true *failure* or *event* time T, as well as a true censoring time C.
- The survival time represents the time at which the event of interest occurs.
- The censoring time is the time at which censoring occurs: for example, the time at which the patient drops out of the study or the study ends.

Survival and Censoring Times

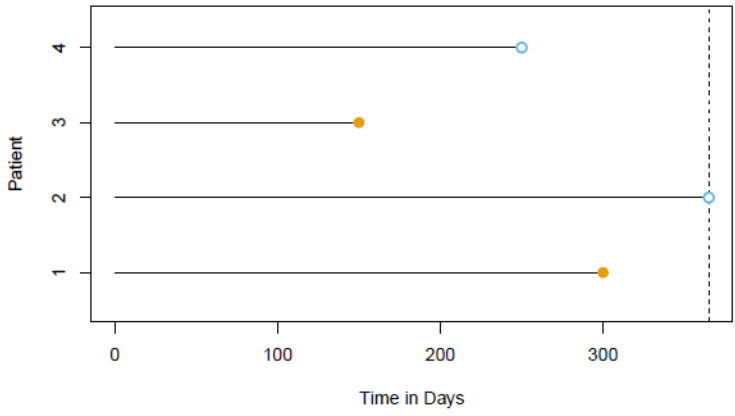
- We observe either T or C, but not both:

$$Y = \min(T, C)$$

- If the event occurs before censoring ($T < C$), we observe the true survival time T
- If censoring occurs before the event ($C < T$), we observe the censoring time C
- We also observe a status indicator

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$$

Illustration



- Patients 1 and 3 event was observed.
- Patient 2 survived to the end of the study.
- Patient 4 dropped out of the study.

A Closer Look at Censoring

- Suppose a number of patients drop out of a cancer study early because they are very sick.
- An analysis that does not take into the considering the reason *why* patients dropped out will likely overestimate the true average survival time.
  - (Very sick patients may have worse prognoses.)
- Further, suppose men who are very sick are more likely to drop out than women who are very sick.
  - A direct comparison of men/women survival times might wrongly suggest that men survive longer than women.

A Closer Look at Censoring

- In general, we assume that, conditional on the features, the event time T is *independent* of the censoring time C.
  - However, the two examples on the previous slide violate this assumption.
- In this chapter, we assume the independence assumption is reasonable.

A Closer Look at Censoring

- We focus on *right censoring*, where $T \geq Y$.
- However, other types of censoring are possible.
- *Left censoring* happens when $T \leq Y$
  - Ex: suppose we want to study pregnancy duration and survey patients 250 days after conception, some of whom have already given birth.
- *Interval censoring* is the setting where we do not know the exact event time, but we know it falls in some interval.
  - Ex: weekly survey that asks if the event happened during the previous week.