MA691 : COBRA-18 Project Report

Anmol Choudhary - 180123004
Ashish Kumar Barnawal - 180123006
Jay Vikas Sable - 180123019
Shivam Kumaar Arya - 180123044

November 20, 2021

Our objective is to analyse the survival time of patients diagnosed with Primary biliary cholangitis (PBC) using COBRA [1]. Throughout the project, we will be using NCSU's PBC dataset [2] `https://www4.stat.ncsu.edu/~boos/var.select/pbc.html`

We use a modified version of `PyCobra` [3], which uses Survival Trees as weak learner [4]. Our objective is to predict the survival function $S(t) = Pr\{T \leq t\}$ which gives the probability that the patient is alive at time t.

We will modify the survival tree from scikit-survival library (see [5]) so that its `predict()` method gives the expected survival time of the patient rather than the survival function. We call named the new class `CobraSurvivalTree`. To find the expected survival time we use the following formula:

$$\begin{aligned} E[T] &= \int_0^\infty t\, dF(t) = -\int_0^\infty t\, dS(t) \\ &\approx -\int_0^{T_{\max}} t\, dS(t) + T_{\max} S(T_{\max}) \\ &\approx \sum_{i=1}^{n-1} t_i\, \Delta S_i \; + \; t_n S_n \end{aligned}$$

Here n is the number of elements in `sksurv.SurvivalTree.event_times_`
We apply COBRA using `CobraSurvivalTree` as the weak learner and take parameter $\epsilon = 3$ years. We get the following results after doing so:

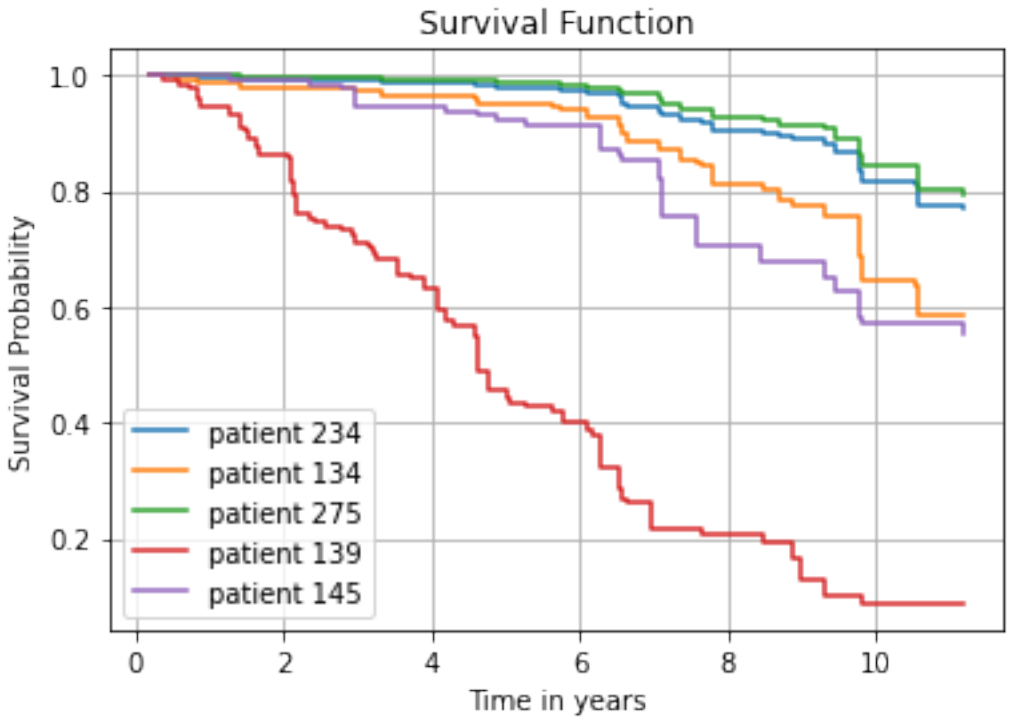
```
Mean Error : 2.31 years
C-index    : 0.68
```

To predict survival function, we will take the average of survival function over the set of selected observations $\{(x_i, y_i)\}_{i \in \mathcal{D}}$ and over the set of selected machines α_m (A datapoint (x_i, y_i) is selected if it in within ϵ range of all selected machines i.e. $i \in \mathcal{D}$ if $|r_j(\mathbf{x}_i) - r_j(\mathbf{x})| < \epsilon \; \forall \; r_j \in \alpha_m$).

If $\mathcal{D}$ denotes the set of selected points and α_m denotes the set of machines the survival function will be

$$S(t) = \frac{1}{|\alpha_m| N} \sum_{r_j \in \alpha_m} \sum_{i \in \mathcal{D}} S_{r_j}(t \,|\, x_i)$$

[5]



References

[1] Gérard Biau, Aurélie Fischer, Benjamin Guedj, and James Malley. COBRA: A Combined Regression Strategy. *Journal of Multivariate Analysis*, 2016.

[2] Fleming and Harrington. Primary Biliary Cirrhosis (PBC) Data, Appendix D, 1991.

[3] Benjamin Guedj and Bhargav Srinivasa Desikan. Pycobra: A python toolbox for ensemble learning and visualisation, 2019.

[4] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008.

[5] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.