IMPLEMENTATION OF COBRA IN INTEGRATED BRIER SCORE

A Project Report Submitted

for the Course

MA691

by

Aditi Bihade (180123001)

Rashi Mohta (180123036)

Shashank Thool (180123043)

Vaarshik Reddy C (180123052)

Karan Gupta (180123064)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

November 2021

DISCLAIMER

This work is for learning purposes only. The work can not be used for publications or commercial products etc. without mentor's consent.

ABSTRACT

The main aim of our project is to predict the survival function for patients diagnosed with breast cancer. We have used the German Breast Cancer Study Group 2 (gbcsg) dataset for our analysis. We explore the dataset by understanding all the features and plotting the Kaplan-Meier estimates. We then estimate the survival function for all the patients using simple survival trees and a modified version of cobra which gives the combined outcome of 4 survival trees. Finally, we calculate the integrated brier score for both the models and compare their results.

Contents

Chapter 1

Survival Analysis

There are certain situations where we are interested in the time it takes for certain events to happen. The objective in survival analysis is to establish a connection between covariates and the time of an event. Any analysis that has to do with time-to-event is a form of survival analysis. Now, time can take on different meanings. Time before fault, time of retainment, time until insurance claim, etc. Event too can take various meanings, depending on the interest of the study. It may mean death, failure, relapse into a state, etc. In this report, we will be exploring the case of breast cancer using certain machine learning techniques and try to predict the survival time of a patient diagnosed with breast cancer. Let us start by understanding some fundamental concepts.

1.1 Survival function and Hazard function

Survival function helps us know whether or not the object of interest is going to survive beyond a specified time. This object of interest can be a patient, a machine or anything else.

Mathematically a survival function is denoted by S which is a function of time. Survival function is represented as :

$$S(t) = P(T > t) = 1 - F(t) \tag{1.1}$$

where T denotes the positive random variable representing time to event of interest, and $F(t)$ is the cumulative distribution function. The value of the function lies between 0 and 1 and is a non-increasing function.

Along with the survival function, we are also interested in the rate at which event is taking place, out of the surviving population at any given time t. The hazard function is defined as the instantaneous risk that the event of interest happens, within a very narrow time frame.

The hazard function is not a density or a probability. However, we can think of it as the probability of failure in an infinitesimally small time period between (t) and $(t + dt)$ given that the subject has survived up till time t. The hazard function is represented as :

$$\lim_{dt \to 0} h(t) = \left(\frac{S(t) - S(t + dt)}{dt} \right) / S(t) \tag{1.2}$$

Since we know that, $\frac{S(t)-S(t+dt)}{dt} = f(t)$, we have,

$$h(t) = \frac{f(t)}{S(t)} \tag{1.3}$$

where $f(t)$ is the probability distribution function.

1.2 Censorship

Survival analysis is a type of regression problem as we want to predict a continuous value, but with a twist. It differs from traditional regression by the fact that parts of the training data can only be partially observed – they are censored.

Censoring occurs when the event of interest is not observed for some subjects before the study is terminated. It occurs when the researcher has partial information about the subjects' survival times but is not privy to the exact survival times.

There are three general types of censoring, right-censoring, left-censoring, and interval-censoring. The most common type of censoring encountered in survival analysis data is right censored (Survival). It is called right censoring because the true unobserved event is to the right of the censoring time. Left-censoring occurs when we cannot observe the time when the event occurred. Interval-censoring occurs in survival analysis when the time until an event of interest is not known precisely (and instead, only is known to fall into a particular interval).

1.3 Kaplan-Meier Estimate

Kaplan-Meier is a statistical method used in the analysis of time to event data. This method is very useful in survival analysis as it is used by the researchers to analyze the patients or participants who lost to follow up or dropped out of the study, those who developed the disease of interest or survived it.

Kaplan Meier estimate is best statistical method used in survival analysis to analyze the data and to make comparison between two groups of partici-

pants such as treatment group and control group using the log-rank test for hypothesis testing.

In real-life cases, we never know the true survival function. That is why with the Kaplan-Meier estimator, we approximate the true survival function from the collected data. The estimator is defined as the fraction of observations who survived for a certain amount of time under the same circumstances and is given by the following formula:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1.4}$$

where d_i is defined as number of events occurred at time t_i, and n_i is defined as the population at risk at time just prior to time t_i

Chapter 2

Dataset

2.1 About the dataset

We have used the German Breast Cancer Study Group 2 (gbcsg) dataset for our analysis. This is an inbuilt dataset in the scikit-survival Python module. The dataset has 686 samples and the following 8 features/covariables:

- **Age:** age (in years),
- **Estrec:** estrogen receptor (in fmol),
- **HorTh:** hormonal therapy (yes or no),
- **Menostat:** menopausal status (premenopausal or postmenopausal),
- **Pnodes:** number of positive nodes,
- **Progrec:** progesterone receptor (in fmol),
- **Tgrade:** tumor grade (I ¡ II ¡ III),
- **Tsize:** tumor size (in mm).

and the two outputs:

- **Recurrence free time** (in days),

- **Censoring indicator** (0 - censored, 1 - event).

2.2 Exploring the data

One of the main challenges of survival analysis is right censoring, i.e., by the end of the study, the event of interest has only occurred for a subset of the observations.

Taking a look at the number of right-censored samples :

- **Number of samples:** 686,
- **Number of right censored samples:** 387
- **Percentage of right censored samples:** 56.4%

Thus, there are 387 patients (56.4%) who were right censored (recurrence free) at the end of the study. The below graphs represents the estimated survival function using Kaplan Meier estimate.

Figure 2.2 compares the approximated survival function of the two cohorts, one which has adopted the method of harmonal therapy and one who hasen't. Figure 2.3 compares the estimated survival functions of patients with Stage I, Stage II and Stage III cancer

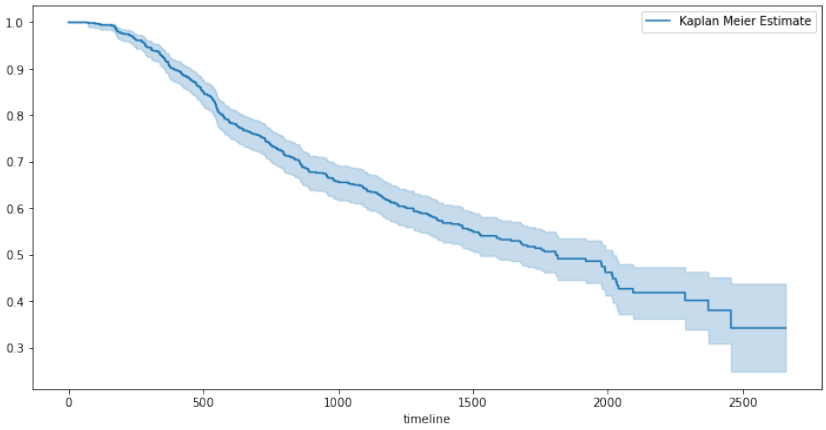

Figure 2.1: Approximate survival function using Kalpan-Meier estimate

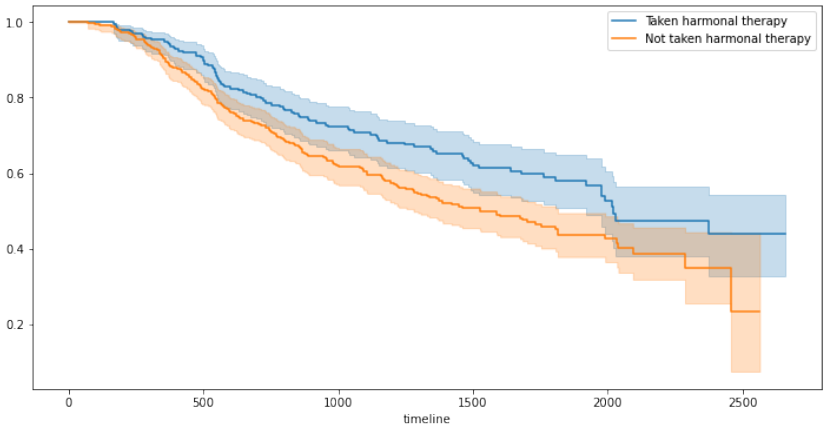

Figure 2.2: Comparison between approximated survival function of the two cohorts

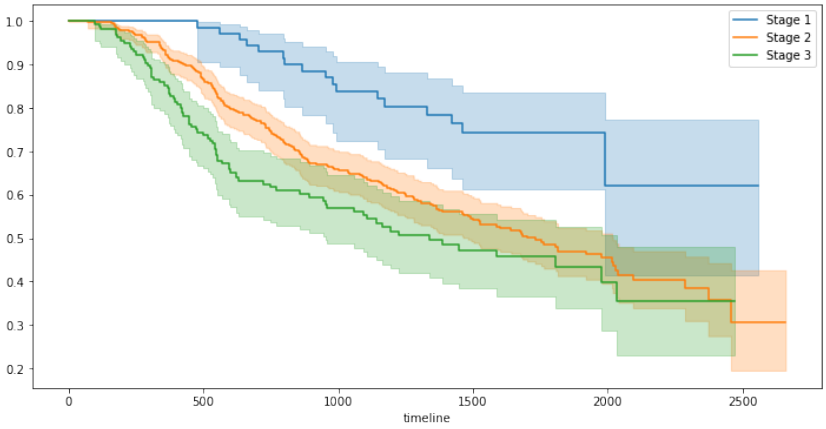

Figure 2.3: Comparison between the estimated survival functions of patients

Chapter 3

Survival Trees

Survival trees are composed of leaves and nodes. Leaves define a partition for the data and are responsible for making predictions. Nodes guide examples towards appropriate leaves using binary splits based on boolean-valued rules.

3.1 Likelihood ratio statistic

Each node of our trees partitions the dataset in more homogeneous subsets such that the split results in maximum reduction of deviance. For a node C in the tree, and all the individuals i in that node, we can define the node deviance as

$$D_C = \sum_{i \in C} \delta_i log(\frac{\delta_i}{\hat{\lambda} t_i}) - (\delta_i - \hat{\lambda} t_i)$$

where. $\hat{\lambda} = \sum_{i \in C} \delta_i / \sum_{i \in C} t_i$ is the maximum likelihood statistic for the rate parameter in the exponential model. Hence while partitioning, it maximises the improvement in fit caused by the partition, $D_{parent} - (D_{left} + D_{right})$.

3.2 Survival function prediction

Terminal node predictions of survival trees are made with the Kaplan-Meier estimator. Let C_j denote the index set of patients with terminal node j, we compute survival prediction at terminal node j with the Kaplan-Meier estimator which was discussed earlier.

$$\hat{h}j(t) = \prod i \in \mathcal{C}j : ti \leq t \left(1 - \frac{N_j\left(t_i\right)}{Y_j\left(t_i\right)}\right)$$

where in terminal node j, $N_j(t_i)$ is the number of events at time t_i and $Y_j(t_i)$ is the total number of individuals at risk just before t_i. The terminal nodes partition the sample space and the survival function can be given by,

$$\hat{h}(t; x_i) = \sum_j T(i_j)\hat{h}_j(t)$$

Chapter 4

Combined Regression Alternative

Cobra is a method for combining several initial estimators of the regression function. Instead of building a linear or convex optimized combination of a collection of basic estimators, they are used as a collective indicator of the proximity between the training data and a test observation. It is a nonlinear method for combining the outcomes over some list of candidate procedures.

4.1 Overview of the model

Usually, in the experts' aggregation theory, we use a convex combination of the experts' predictions to make $\hat{y}$. But COBRA has a very different approach. It is based on a similar idea as the k-nearest neighbourhood algorithm. At each time t we have a new observation x_t, and the K experts' predictions $p_1(x_t), p_2(x_t), ...p_k(x_t))$ are computed. Then, average is calculated over the realizations of y, not used to generate the experts, that have predictions in the same neighbourhood of $p_1(x_t), p_2(x_t), ...p_k(x_t))$.

4.2 Mathematical Explanation

Formally, the COBRA estimator is made as the following. Let D_n be a sample of n independent and identically distributed observations of the pair of random variable (X, Y). The sample is divided into two independent samples, D_l and D_m. Then, D_l is used to generate a set of experts $p_1, p_2 ... p_k$ and D_m is used for the calculation of $\hat{y}_t$, the combined predicted value for a new observation X_t. We have the following formulas:

$$\hat{y}_t = \sum_{i=1}^{m} W_i\left(x_t\right) y_i \tag{4.1}$$

where the random weights W_i take the form:

$$W_i\left(x_t\right) = \frac{\mathbf{1}_{\cap_{j=1}^{k}}\left\{\left|p_j\left(x_t\right) - p_j\left(x_i\right)\right| \leq \epsilon_m\right\}}{\sum_{i=1}^{m} \mathbf{1}_{\cap_{j=1}^{k}}\left\{\left|p_j\left(x_t\right) - p_j\left(x_i\right)\right| \leq \epsilon_m\right\}} \tag{4.2}$$

The ϵ_m is the smoothing parameter. The larger ϵ_m, the more tolerant the process. Conversely, if ϵ_m is too small, many experts are discarded. Therefore, its calibration is a crucial step.

Chapter 5

Integrated Brier Score

Integrated Brier Score (IBS) is an overall measure for the prediction of a model for a given interval of time. In our study, we use IBS as a performance measure to evaluate the predicted survival distributions.

5.1 Brier Score

The Brier score is used to evaluate the accuracy of a predicted survival function at a given time t; it represents the average squared distances between the observed survival status and the predicted survival probability and is always a number between 0 and 1, with 0 being the best possible value.

Given a dataset of N samples, $\forall i \in 1, N, (\vec{x}_i, \delta_i, T_i)$ is the format of a datapoint, and the predicted survival function is $\hat{S}\left(t, \vec{x}_i\right), \forall t \in \mathbb{R}^+$.

In the absence of right censoring, the Brier score can be calculated such that:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(1_{T_i > t} - \hat{S}\left(t, \vec{x}_i\right)\right)^2 \tag{5.1}$$

However, if the dataset contains samples that are right censored, then it is necessary to adjust the score by weighting the squared distances using the

inverse probability of censoring weights method.

5.2 Integrated Brier Score

The Integrated Brier Score (IBS) provides an overall calculation of the model performance at all available times. For prediction over the range of all future times we aggregate the Brier Score over time resulting in the Integrated Brier Score. A model with perfect skill has a score of 0 and the worst has a score of 1.

$$IBS\left(t_{\max}\right) = \frac{1}{t_{\max}} \int_0^{t_{\max}} BS(t)dt \tag{5.2}$$

where I stands for the indicator function.

Chapter 6

Implementation and Results

For the breast cancer survival analysis, we have used a modified version of PyCobra. Survival trees have been used as the weak learner. Our objective is to predict the survival function for every patient and calculate the integrated brier score.

6.1 Modified PyCobra

We create a class **PyCobraSurvivalTree** which predicts the survival function using survival trees and outputs an array instead of the step function. The time points corresponding to each element in the array can be extracted using the *event_times_* method.

Using the epsilon parameter, we identify the the data points close to the current data point and take the average of all the predicted survival functions corresponding to these data points and predicted by all the machines. This calculation of the mean survival function has been done in the function **pred_surv**.

6.2 Predicted survival functions

Predictions made using Survival trees and the modified PyCobra model which uses 4 weak estimators are shown in Figure 6.1 and 6.2.

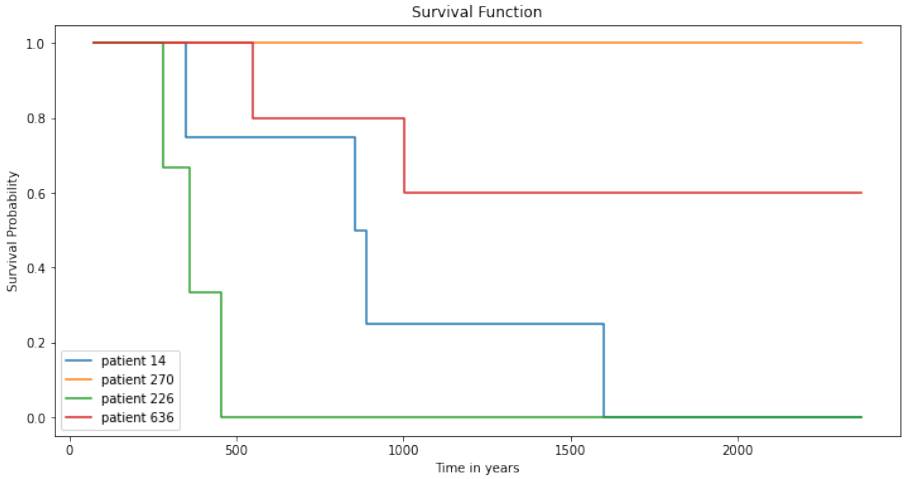

Figure 6.1: Predicted survival function using Survival Tree

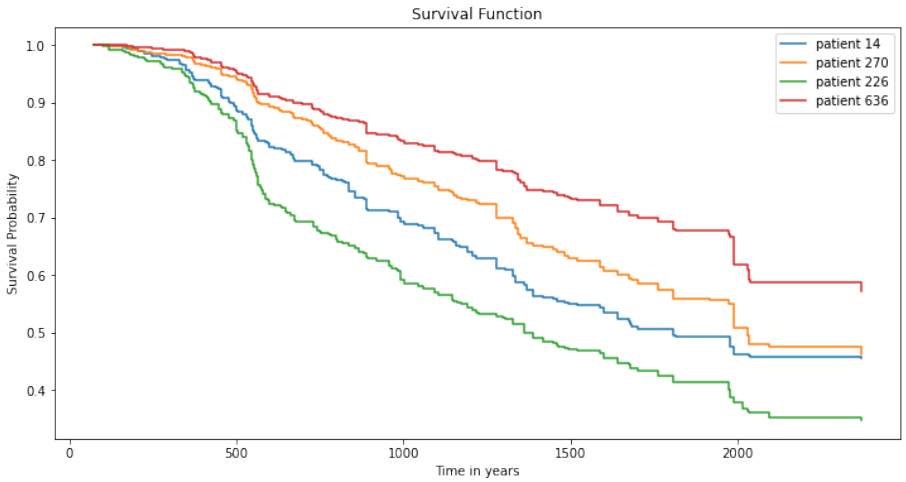

Figure 6.2: Predicted survival function using Cobra

6.3 Results and Discussion

The Integrated Brier scores for both the methods are:

- **Survival Tree:** 0.24
- **Modified Cobra:** 0.16

We observe an increase in accuracy when Cobra is used. These results are justified as Cobra is a combination of several of these estimators and is thus giving better results than any of the estimator used separately.