Machine Learning in healthcare: Survival analysis

Jorge Parreño Hernández, Riccardo Conforto Galli, Filippo Nardi

November 2023

Preprocessing

The first step in the preprocessing involved analyzing and handling missing values in the data. The data was loaded into the environment using the library provided in the documentation of the lab, which loads it as a pandas dataframe. In order to impute missing values we use an RNN, since the missing values are only in numeric columns.

The defined architecture for the RNN is arbitrary, when varying the different sizes of the RNN we observed a marginal influence on the outcome, hinting at the possibility that the model may have been adequately sized to handle the underlying patterns in the data.

Clustering and dimensionality reduction

To initiate the clustering process, we begin by conducting dimensionality reduction on the dataset. For this task, we opt for Uniform Manifold Approximation and Projection (UMAP). UMAP, an acronym for Uniform Manifold Approximation and Projection, is a dimensionality reduction technique specifically designed to emphasize the preservation of topological structures during the reduction process.

Following dimensionality reduction, our clustering approach involves the utilization of Gaussian Mixture models. We employ probabilistic clustering techniques, leveraging the Bayesian Information Criterion (BIC) to determine the optimal number of components for the Gaussian Mixture models. The BIC serves as a valuable criterion in guiding the selection of the most suitable number of clusters, contributing to the robustness and accuracy of our clustering methodology.

Kaplan-Meier vs Cox-Hazard model

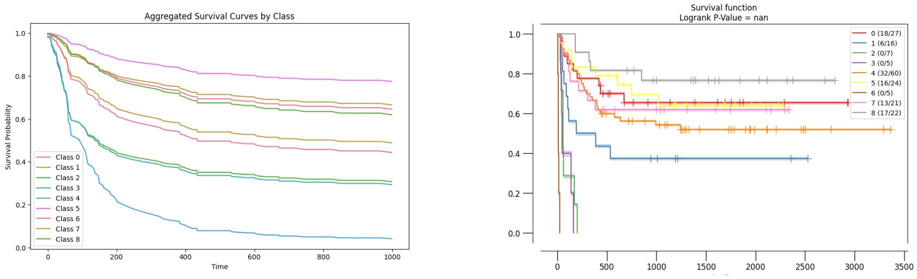

Figure 1: Survival curves for Cox-Hazard model vs Kaplan-Meier

At first glance we observe that the Cox-Hazard model and the Kaplan-Meier models predict similar patterns in the general survival, however on a class basis, even considering that we had to remove some features so that the Cox-Hazard model converged because they provided perfect separation. The Cox-Hazard model curves are generated aggregating by the median, as the output of the built-in predict method. Feature importances are computed for all features in the Cox-Hazard model, for different values we observe the decrease in feature importance through time.

Code source

The code can be found in the following github repository, where we have stored all our code regarding this course, under the subfolder 'lab2'.