CS-E5890 – Statistical Genetics and Personalised Medicine – 17.3.2017, Tomi Peltola

Survival analysis project

In this project, you apply survival analysis to study the risk factors for acute myeloid leukemia (AML) and develop a risk prediction model and evaluate its predictions. The problem is based on the AML outcome prediction DREAM challenge (https://www.synapse.org/#!Synapse:syn2455683/wiki/64007). You can focus on the overall survival (you are, of course, free to explore also the remission duration if interested).

Instructions:

1. Familiarize yourself with the challenge description at https://www.synapse.org/#!Synapse:syn2455683/wiki/64007. Specifically, in addition to the front page, it's probably good to check at least `0.5.1 Data description` and `06.5 Benchmark models`. `04.2 Hackathlon` might also be useful.

2. Get the data: At the website, register for a Synapse account (`Register` link on top-right on the page). After logging in, click `Access the Data` (below the AML banner) and download the training data file `trainingData-extendedChallenge-noChemo-release.xlsx` (note that the `trainingData-release.csv` is not the same dataset; we won't use the `scoring data` as the survival outcomes have not been made publicly available).

3. Explore the dataset.

- Find out if there is missing data and impute it using median method (or some other approach; see `06.5` for discussion related to missing data imputation). The median method is a crude imputation approach that fills in the missing values of some covariate by replacing them with the median of the observed values for that covariate. (Don't impute any outcome variables.)
- Study the distributions of the covariates and correlations between them.
- Compute Kaplan-Meier curves for the full set (and if there are possibly important categorical covariates, you should also look at the survival curves by subgroups).

4. Predict survival probabilities.

- First, fit a Cox proportional hazard model on basic covariates (age, cytogenic category, prior chemotherapy, HGB, and albumin; following `06.5`) and make predictions using it.
- Find out which covariates seem most important for predictions.
- Assess the predictive performance using 10-fold cross-validation and the C-index statistic for predictions at 5 years after diagnosis.
- Plot the risk predictions from the cross-validation the five-year time point for all the samples. Compare informally to the known outcomes.
- Fit some other model or models that use more covariates and are perhaps more advanced in other ways too (see below for some possible ideas). Compare the predictive performance to the basic Cox-PH model using cross-validation and C-index. (Don't use any information *from the future* as a covariate; these include `resp.simple`, `Relapse`, `Remission_Duration` when predicting the overall survival, to the extend I understood from the challenge webpage.)

5. Write a report with a structure: 1) Introduction, 2) Data, 3) Methods, 4) Results & Discussion, 5) Conclusions. Try to include some analysis of the results and/or methods (e.g., rather than just reporting that you got accuracies X, Y, Z for the tested prediction methods, try to also understand reasons why one approach might have been better than another; or/and, if you are more medical-application oriented, you can try to compare your results to what is known about risk factors for AML mortality).

Some possibly useful resources:

- R-package Hmisc has rcorr.cens that can be used to compute the C index. Just give the predicted risk probabilities as the predictor variable.
- Example for getting survival probability prediction after coxph regression in R (alternatively, one can use predictSurvProb-function in pec package):

```
fit <- coxph(Surv(futime, fustat) ~ age, data = ovarian)
survprob <- summary(survfit(fit, newdata=data.frame(age=60)), time=300)$surv
```

- survival R-package has basic estimtors and regression models (https://cran.r-project.org/web/packages/survival/index.html).

Some ideas for other models than the Cox PH with the basic covariates:

- Sparse models, that is models where many of the covariates will turn out to not have any effect on the prediction, can be implemented using, e.g., glmnet R-package (https://cran.r-project.org/web/packages/glmnet/index.html; predictProb.coxnet-function in c060 package might give survival probability predictions for this; haven't tested). Such models could be useful for including the proteomic measurements.
- Another idea to include the proteomic measurements would be to reduce the dimensionality of the protein data by PCA before feeding them into the prediction model. This could make sense if the proteins are correlated with each other.
- Parametric survival regression is relatively simple to implement in the probabilistic programming language Stan (http://mc-stan.org). The model could also then include sparsity-inducing priors for the protein data.
- GPstuff toolbox for Matlab can fit non-linear survival models based on Gaussian processes (http://research.cs.aalto.fi/pml/software/gpstuff/).