Owkin Data Challenge

E. Levecque (Labaro)

February 2020

1 Approach

This is the first time I dealt with a Survival Analysis so I wasn't comfortable with the data set. However after playing with the data I noticed several issues. First there is some missing important data such as the age and the histology. Second the data set was composed of two different analysis and we could see some differences depending on the source. Third the data set was composed of only 300 sample and 53 features so it is small and some features are correlated.

After this identification I had to choose between different solutions. For the difference between the sources and the fact that features are correlated, the best would have been to build a model based on the images and not using the mask on them. However the small size of the data set was an issue to train a CNN. For example, it could have been very interesting to determine the histology of all samples and then using 4 different Cox-PH models trained on each histology separately. An other solution would have been to separate the data set depending on the source of the samples. Then using 2 different models trained separately. For the big number of features, one solution is to perform a dimension reduction.

I chose to use the two last solutions: separating the data set into 2 smaller data set and using an algorithm of dimension reduction based on the singular value decomposition.

2 Model

First I transformed the data set to have no categorical features such as histology and to separate the two different centers of study. For the dimension reduction I used a deterministic column sampling algorithm based on leverage score [2] with parameters $k = 7$ and $\theta = 6.9$. After that I have obtained 8 features only:

- original_glcm_ClusterShade;
- original_glcm_ClusterProminence;
- original_firstorder_Range;

- original_firstorder_Minimum;
- original_glcm_Autocorrelation;
- original_firstorder_StandardDeviation;
- original_firstorder_Uniformity;
- original_glrlm_HighGrayLevelRunEmphasis.

Then I used pysurvival package to build two similar Cox-PH model which were fitted with the data separately [1]. The parameters used were: $lr = 5.10^{-2}$, $l2\_reg = 1.10^{-3}$, $init\_method = zeros$, $max\_iter = 1000$. The parameters were tuned by cross validation using 4 folds. I have obtained a mean score of 0.6874 with this model over 20 different shuffles of training and testing sets. And it performs with 0.7055 on the testing data set.

3 Future work

Others models are available on pysurvival. However having only one week to finish this project I decided to limit my work at cox-PH which was the baseline model. But it could be interesting to try and compare different models. For example I rapidly try SVM survival model and obtain very good cindex values. The difficulty with a SVM model is to find the expectation lifetime from the risk because there is no survival function.

Features like age should have a lot of importance in this kind of model. Therefore it could be useful to recover the missing value (using inference for example). In the same idea, by reading some literature on the survival analysis, I also discovered that sex is a very important feature that can be used to separate the data set. Here it wasn't available but I'm quite sure that separating the data set using the sex would have increase the performances of the model.

References

[1] Stephane Fotso et al. *PySurvival: Open source package for Survival Analysis modeling*. 2019–.

[2] Dimitris Papailiopoulos, Anastasios Kyrillidis, and Christos Boutsidis. "Provable Deterministic Leverage Score Sampling". In: (June 2, 2014). arXiv: `1404.1530 [cs, math, stat]`.