Multiple Imputation and Cross-Validation for Classification of Survival Prediction

Robert Edwards

(2416963E)

MASTER THESIS

Biostatistics

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim of the Thesis

1.2 The Clinical Study

1.3 Study Population & Data Description

1.4 The Statistical Challenege

2 Methodology

2.1 Basic Statistical Methods

2.1.1 Logistic Regression

2.1.2 Linear Discriminant Analysis

2.1.3 Quadratic Discriminant Analysis

2.1.4 K-Nearest Neighbors

2.1.5 Random Forests

2.2 Missing Data

2.3 Multiple Imputation

2.4 Validation & Cross-validation

2.5 Accuracy Metrics

These are the default metrics used to evaluate algorithms on binary and multi-class classification datasets in caret.

2.5.1 Accuracy

Accuracy is the percentage of correctly classifies instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes (e.g. you need to go deeper with a confusion matrix). Learn more about Accuracy here.

Don't use accuracy (or error rate) to evaluate your classifier! There are two significant problems with it. Accuracy applies a naive 0.50 threshold to decide between classes, and this is usually wrong when the classes are imbalanced. Second, classification accuracy is based on a simple count of the errors, and you should know more than this. You should know which classes are being confused and where (top end of scores, bottom end, throughout?)

2.5.2 ROC

2.5.3 Kappa

Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes (e.g. 70-30 split for classes 0 and 1 and you can achieve 70% accuracy by predicting all instances are for class 0).

2.5.4 Brier Score

2.5.5 F1 Score

3 Statistical Methods for the Analysis

Describe the methods step-by-step for the analysis

3.1 Complete Case Analysis

3.2 Mean Imputation

3.3 Multiple Imputation

3.3.1 Joint-Model

3.3.2 FMC(?)

3.3.3 Predictive Mean Matching

Predictive Mean Matching (PMM) is a semi-parametric imputation approach. It is similar to the regression method except that for each missing value, it fills in a value randomly from among the a observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model (Heitjan and Little 1991; Schenker and Taylor 1996). The PMM method ensures that imputed values are plausible; it might be more appropriate than the regression method (which assumes a joint multivariate normal distribution) if the normality assumption is violated (Horton and Lipsitz 2001, p. 246).

4 Results

5 Discussion

6 Conclusion

7 Bibliography

8 Appendices

8.1 Additional Material

8.2 R Code