Multiple Imputation and Cross-Validation for Classification of Survival Prediction

Robert Edwards

(2416963E)

MASTER THESIS

Biostatistics

University
of Glasgow

Contents

List of Figures

List of Tables

1 Introduction

2 Basic Statistical Methods

2.1 Accuracy and Kappa

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

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?)

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).

3 Statistical Methods for the Analysis

4   Results

5   Discussion

6 Conclusion

7   Bibliography

8   Appendices

8.1   Additional Material

8.2   R Code