Data Science Toolbox Portfolio Questions

05 Supervised Learning and Ensembles

Daniel Lawson — University of Bristol

Block 5

Portfolio 05

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R05.1: The Kaggle competition MAP - Charting Student Math Misunderstandings is a classification problem based on free text information. Examine the code for a small variety of the top entries, and describe broadly how they have framed the problem. How does this practical classification differ from how it was presented in lectures, and how does it differ from the practical use case?

Question R05.2: Take a look at the Scikit-learn documentation on Histogram Boosting. Briefly summarise how it works, and “why it’s faster”. Giving a minimal code example as an appendix, compare its performance to “GradientBoostingClassifier” for a range of sample sizes. What is the trade-off between the two methods and what hyperparameter choices can impact it?

Question R05.3: Our Receiver-Operator Curve (ROC) is defined for binary classification problems. How can we extend it to multi-class classification problems? Using the package pROC in R or Multiclass Receiver-Operator curve in python, apply at least two multiclass scoring functions to the prediction of the `service` variable from which we derived the (binary) `http` variable for the data from Block 5 Workshop. Discuss how they differ both conceptually and in practice.