Data Science Toolbox Portfolio Questions

05 Decision Trees and Random Forests

Daniel Lawson — University of Bristol

Block 6

Portfolio 06

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 R06.1: Read Section 2 of XGBoost: A Scalable Tree Boosting System. Explain what the second-order gradient statistics for each leaf are, and how and why they are used in boosting.

Question R06.2: Extend the workshop to make a specific comparison between the choice of splitting measure - i.e. compare the **Gini index** to **ID3** and potentially other measures, both for **Decision Trees** and inside a tree ensemble, either bagged decision trees or Random Forest. You may find Section 2.1 of Decision trees: a recent overview helpful. Your code and figures would be an appendix, you should focus on any conceptual issues and conclusions in your portfolio.

Question R06.3: Decision trees align decision boundaries with **Features**. Either empirically or theoretically, discuss the use of using **PCA to construct features** for use in decision trees. You can do this either by referencing the literature as a mini-review, or by extending the workshop as in Q6.2.