Data Science Toolbox Portfolio Questions

06 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: LightGBM Experiments show some impressive computational and accauracy results using "Best-first Decision Tree Learning", vs "Leaf-first" approaches. What empirical or theoretical evidence can you find (or create) to support or reject the claim that the critical difference is the "Best-first" approach?

Question R06.2: 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 and random forests, boosted or otherwise. You can do this either by referencing the literature as a mini-review, or by extending the Block 06 workshop.

Question R06.3: Add LightGBM to the Block 06 workshop and compare its performance to Random Forests and xgBoost on at least one dataset from this course. With references to examples in the wild where Random Forests beat GBMs and vice versa, perform testing of selected hyperparameters to see if you can replicate these phenomena. From your examinations how confident are you of the superiority of one method over the other, and why?