Data Science Toolbox Question Sheet

06.1 Trees, Forests, Decisions

Daniel Lawson

Block 6

1. Describe a decision tree, and how a Random Forest is constructed from it. What is the difference between a boosted decision tree and a random forest?
2. Why would choosing a feature i to split in a decision tree by minimising $G_i = \sum_{j=1}^{J} p_{ij}(1-p_{ij})$ be a good idea?
3. How is choosing features by minimizing $H_i = \sum_{j=1}^{J} p_{ij} log(p_{ij})$ different?
4. Explain bagging, and how it would be used to reduce overfitting of a decision tree?
5. A colleague uses a random forest and obtains features importances (0.6,0.3,0.2). They conclude that feature 1 is twice as important as feature 2, and that all three features add value to the classifier. Critique these conclusions.