H2O.ai

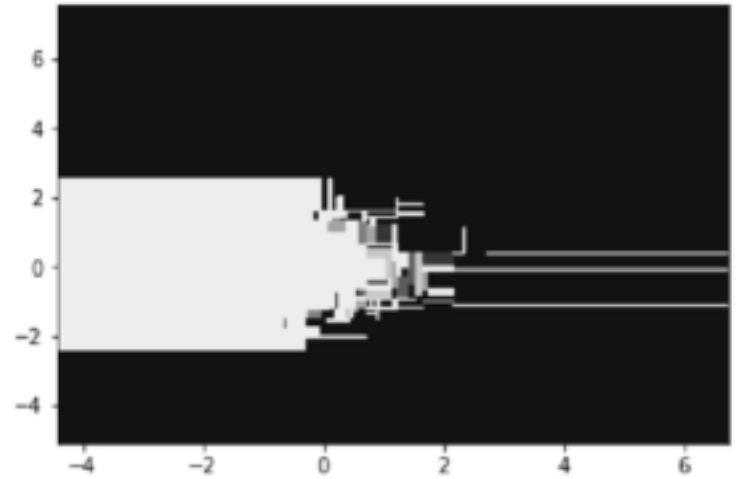
Missing and Unseen Values

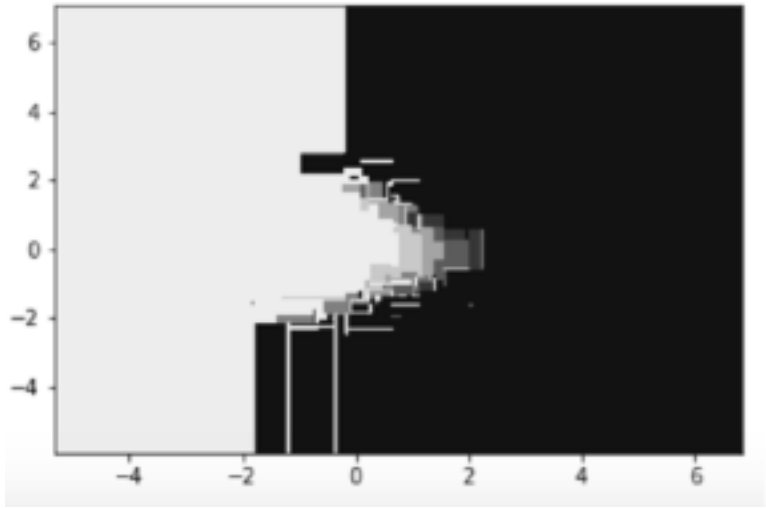
- Missing Data in Training Predictors: Treated as a “level”
- Missing Value in Training Target: Ignored
- Unseen Categorical Levels During Scoring: Treated as an “NA” value and classified with outliers

Pros/Cons of Decision Trees

- ✔ simple to understand/interpret
- ✔ little data prep
  - natural handling of "mixed" data types
  - handling of missing values
  - handing of multi-class outputs
- ✔ robustness to outliers in input space
- ✔ insensitive to monotonic transformations of inputs
- ✔ computational scalability (large N) $\mathcal{O}(n \log n)$
- ✔ automatically ignores irrelevant inputs

- ✗ weak predictors
- ✗ can be unstable to small variations in the data
- ✗ poor ability to extract linear relationships
- ✗ can create biased trees if classes unbalanced





Exercise 1: Creating Decision Tree in sklearn

Missing and Unseen Values

- Missing Data in Training Predictors: Treated as a “level”
- Missing Value in Training Target: Ignored
- Unseen Categorical Levels During Scoring: Treated as an “NA” value and classified with outliers

H2O.ai