
H2O.ai

Growing a Tree: Splitting Criteria

- If starting with criterion to minimize, then learning an optimal decision tree is an NP-complete problem

- heuristic algorithms employ greedy procedure

- splitting criteria:

    Classification: k classes

        Entropy:

        Gini:

        Squared error:

    Regression

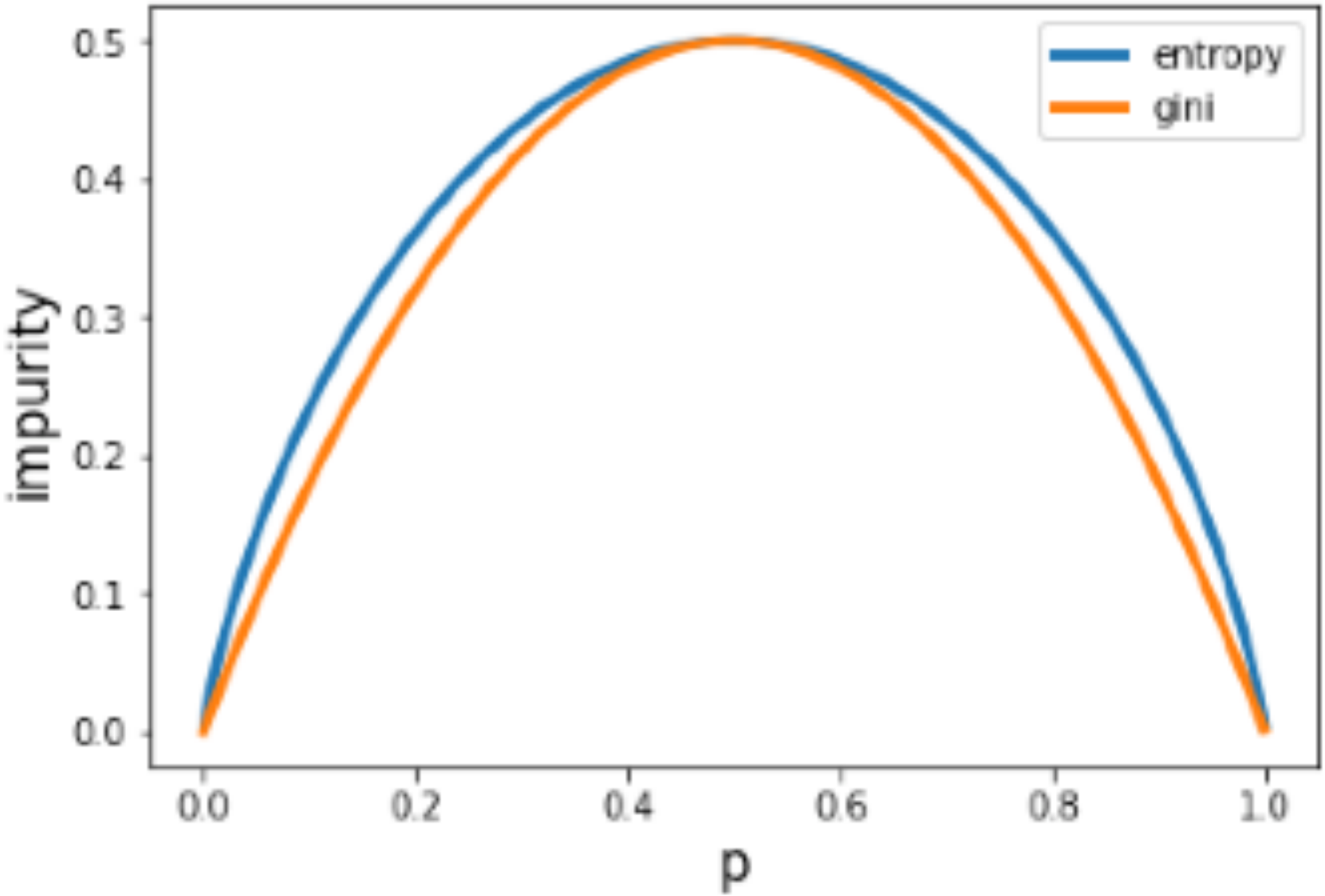
        Squared error

$$\sum_{i=1}^{k} \hat{p}_i(1 - \hat{p}_i)$$

$$\sum_{i=1}^{k} \hat{p}_i \log(\hat{p}_i)$$

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$


entropy
gini
impurity
p
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0

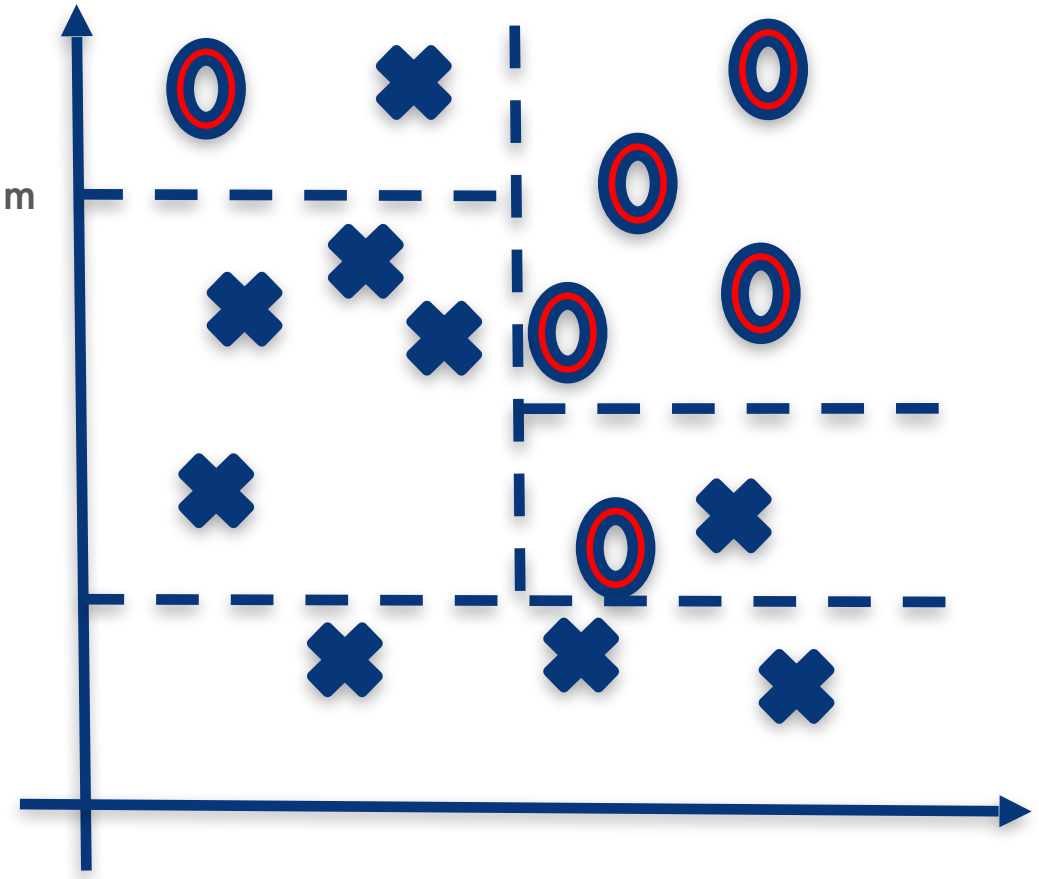
Growing a Tree: Stopping Criteria

1) All instances in leaf have the same y value

2) A maximum tree depth has been reached

3) Number of instances in leaf is below some minimum

4) Splitting criteria is below some threshold



Growing a Tree: Splitting Criteria

- If starting with criterion to minimize, then learning an optimal decision tree is an NP-complete problem
- heuristic algorithms employ greedy procedure
- splitting criteria:

Classification: k classes

Entropy:

$$\sum_{i=1}^{k} \hat{p}_i \log(\hat{p}_i)$$

Gini:

$$\sum_{i=1}^{k} \hat{p}_i (1 - \hat{p}_i)$$

Squared error:

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$

Regression

Squared error

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$