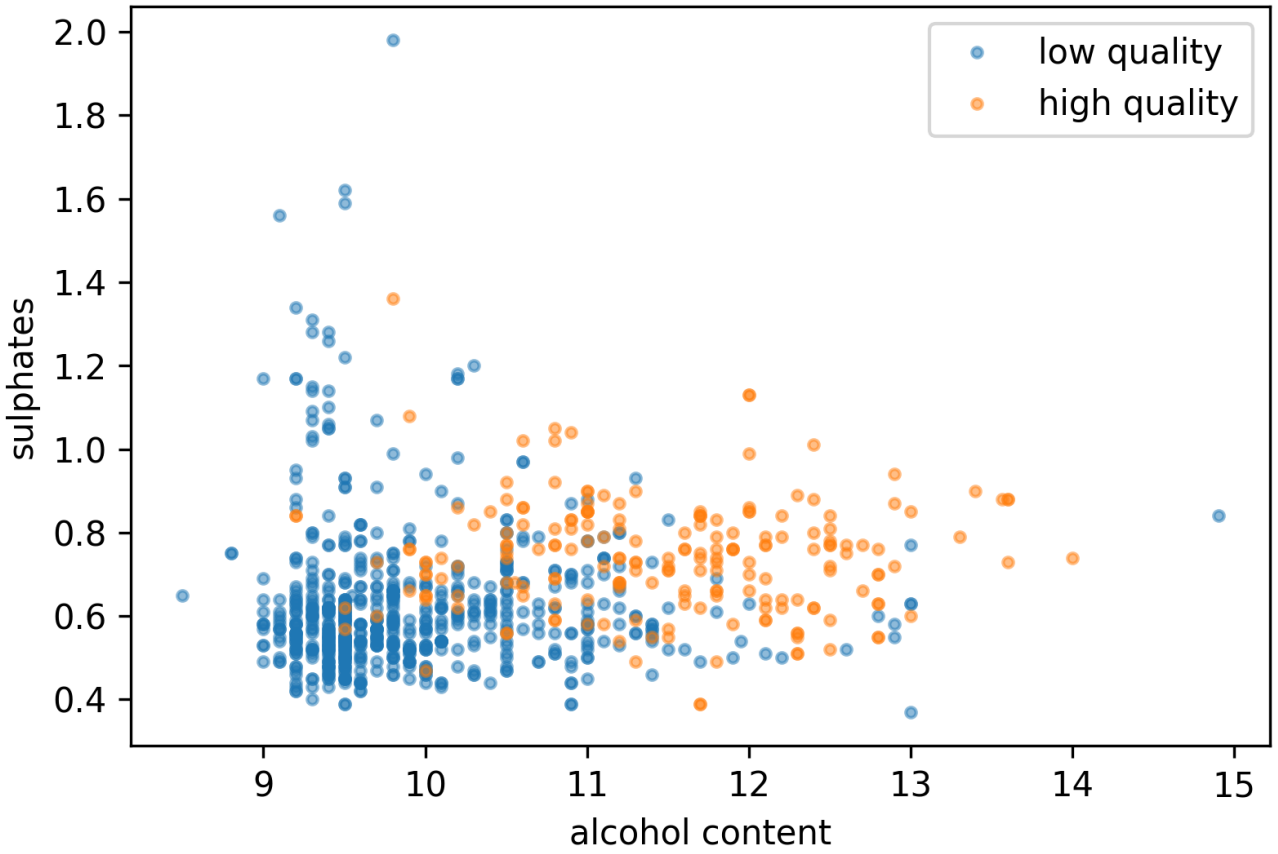
HiOA Big Data Course

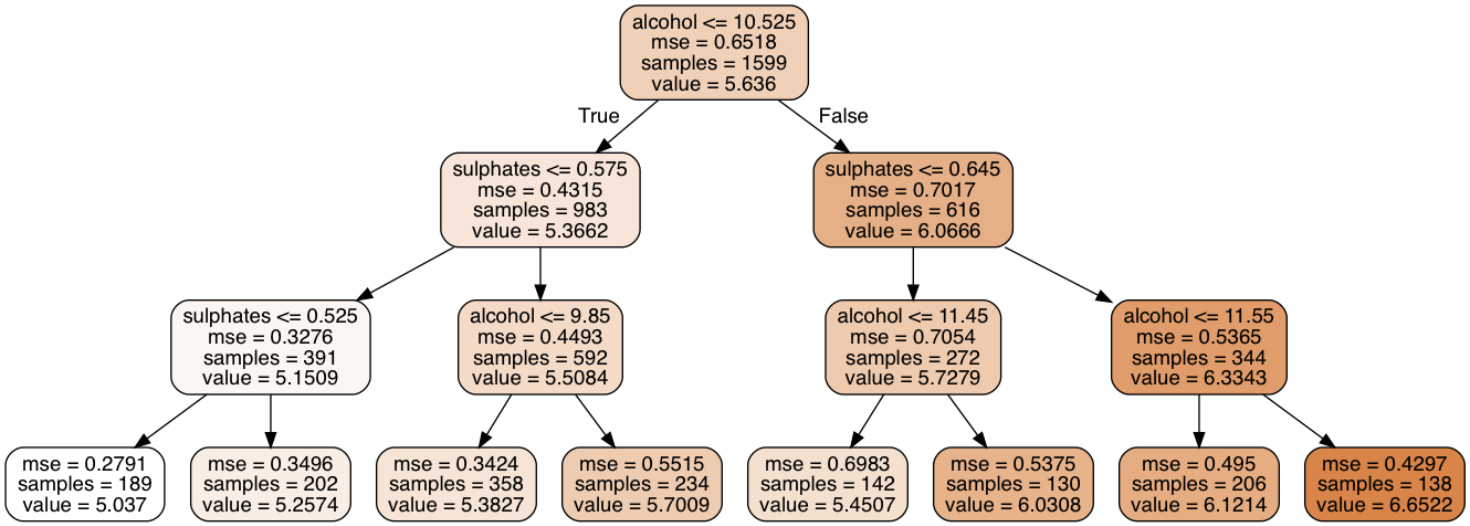
Session 6 - Trees

Dirk Hesse

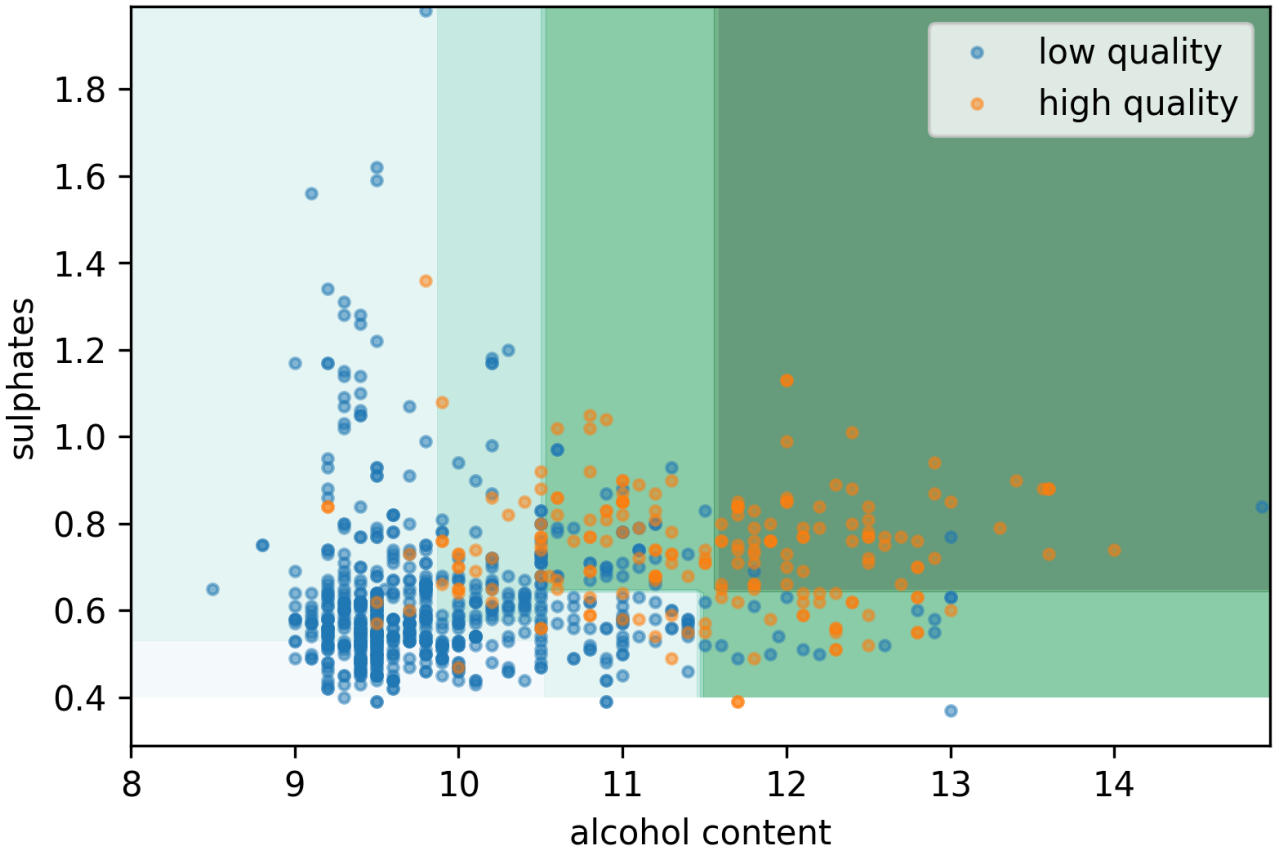
Wine Quality



Wine Quality Tree



Wine Quality Regression Tree



Why trees?

- Simple.
- Easy to explain.
  - Especially to non-experts.
- Powerful.

Calculating Trees

- Divide your data $R_L(j, s) = \{X | X^{(j)} \leq s\}$, $R_R(j, s) = \{X | X^{(j)} > s\}$.
- Find the best a_R, a_L, j, s to minimize

$$\sum_{i, x_i \in R_L(j,s)} (a_L - y_i)^2 + \sum_{i, x_i \in R_R(j,s)} (a_R - y_i)^2$$

- For given j, s, we find that $a_{R,L} = \underset{i, x_i \in R_{R,L}}{\text{avg}} \; y_i$.
- Repeat on the sub-sets.
  - Until a maximum depth is reached.
  - Until a minimum number of samples is reached.

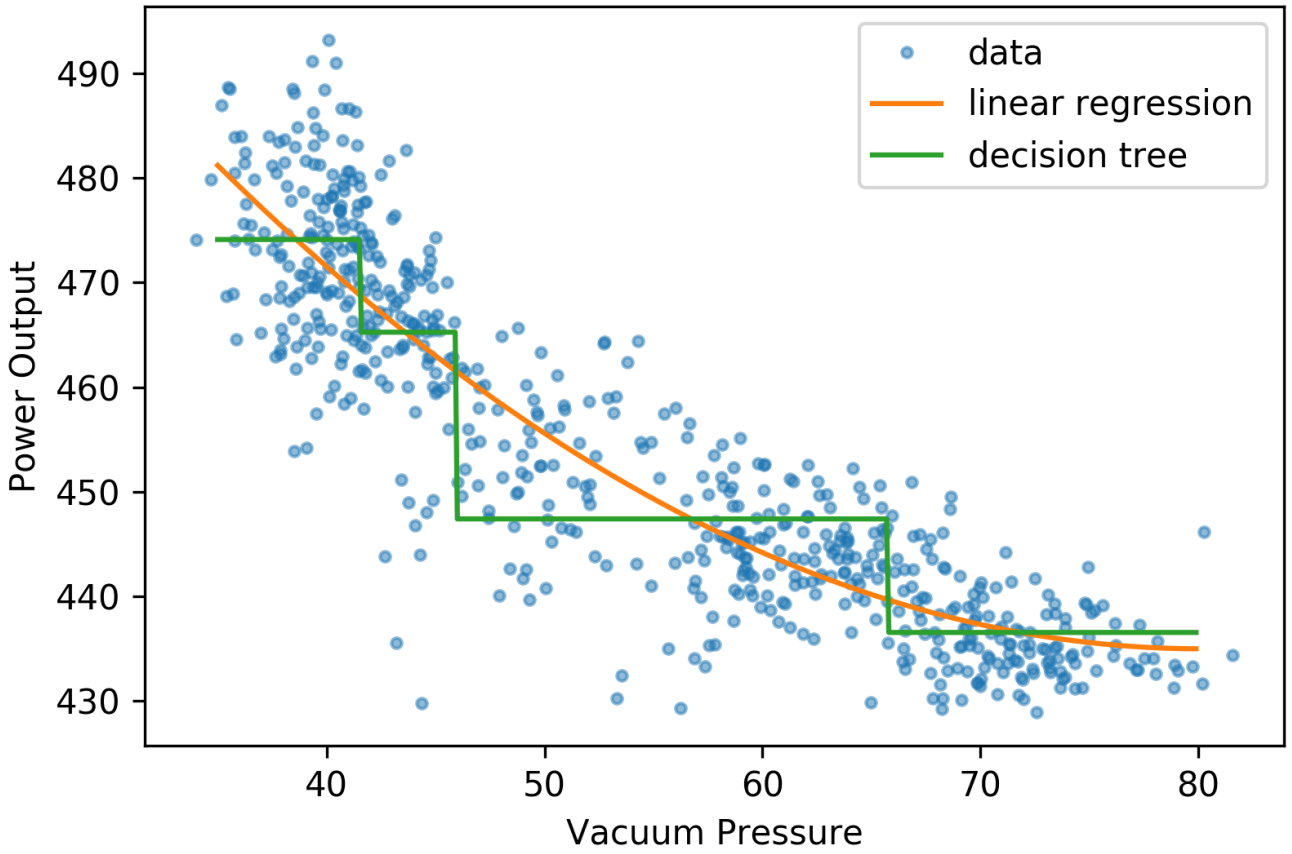
Regression Tree

Our resulting model reads

$$\hat{f}(X) = \sum_m c_m I\{X \in R_m\}.$$

Hence trees are an example of a general class of *additive models*.

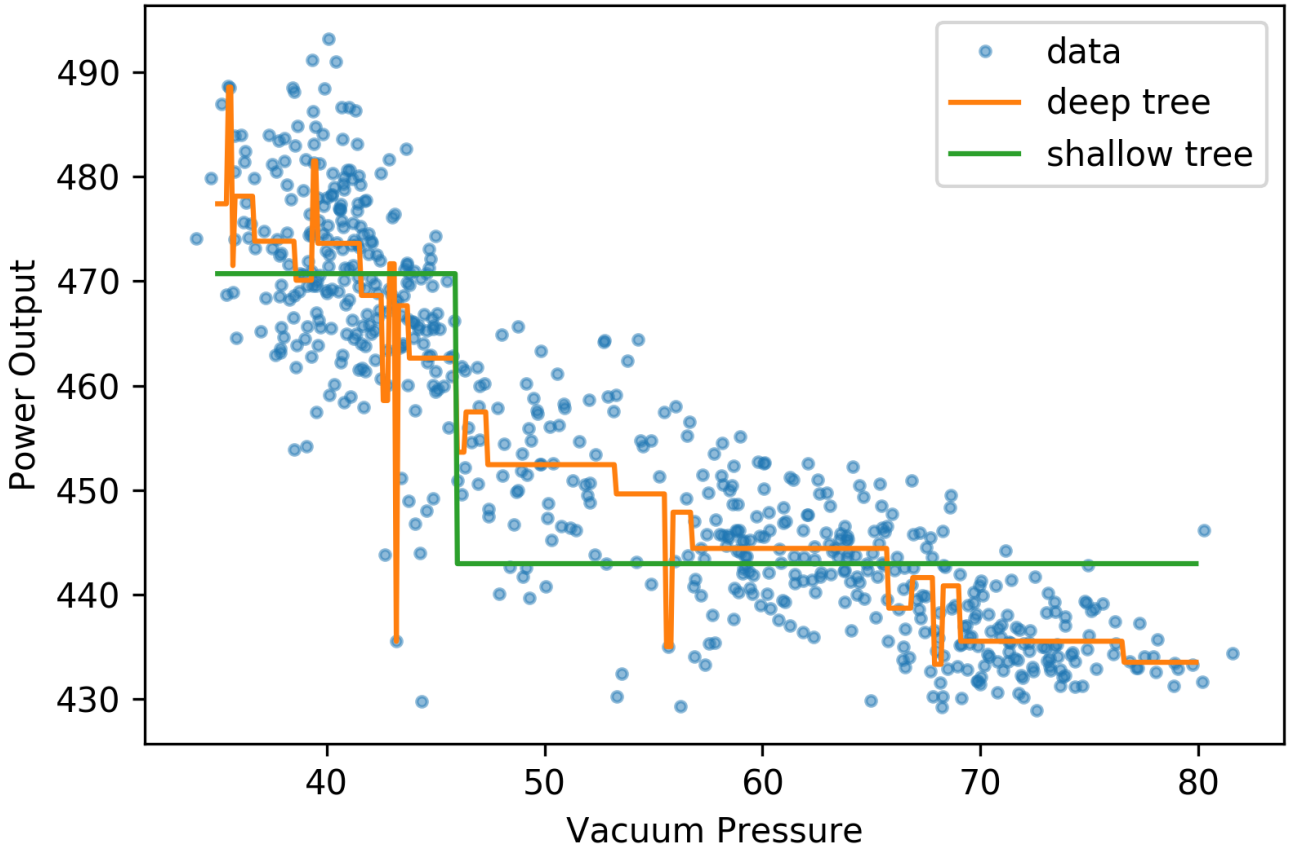
Tree Vs Linear Regression



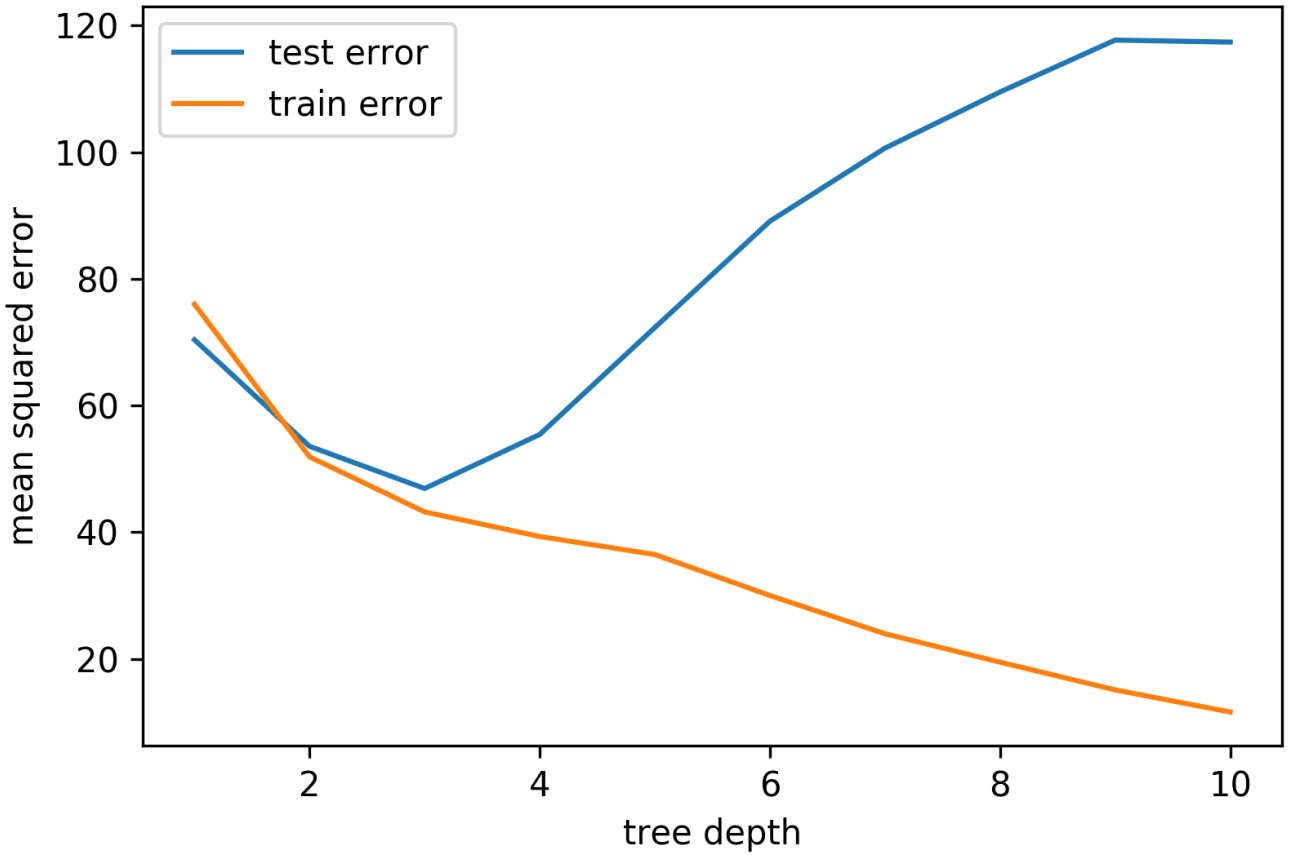
How Deep Should You Go?

- Deep trees have many degrees of freedom and hence high **variance**.
- Too shallow trees can't capture the *shape* of the data.
  - Hence have high **bias**.

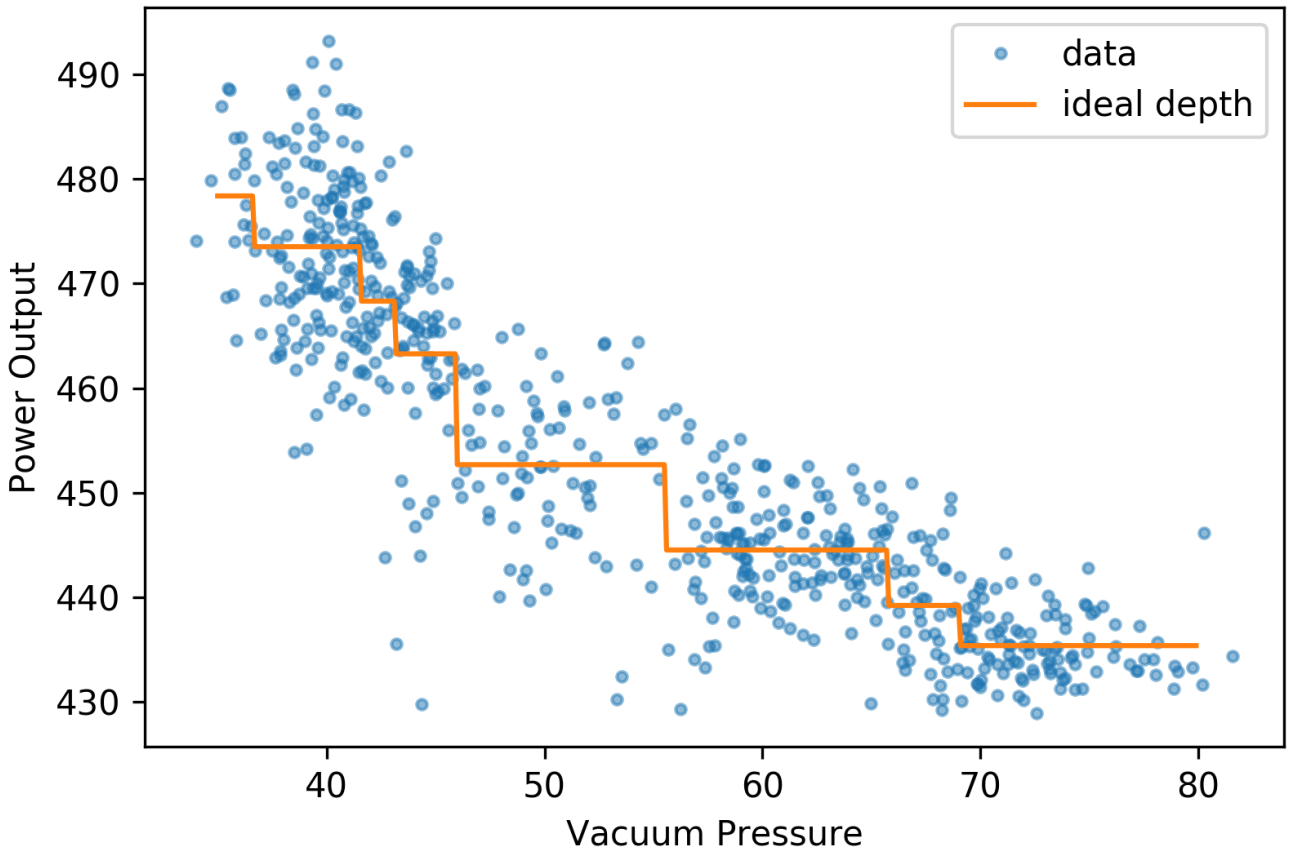
Bias-Variance Trade-off for Trees



Training and Test Error



The Best Tree



Trees for Classification

Just *modifying* our tree formulas to use the **mode**

$$a_{R,L} = \underset{i, x_i \in R_{R,L}}{\text{mode}} \; y_i$$

yields a classification algorithm.

How Find the Splits for Classification?

Define

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{i; x_i \in R_m} I(y_i = k),$$

such that $k(m) = \operatorname{argmax}_k \hat{p}_{mk}$

- Misclassification: $1 - \hat{p}_{mk(m)}$.
- Gini index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
- Cross-entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

Two-Class impurity Measures