Decision trees

Performing classification

Herman Kamper

http://www.kamperh.com/

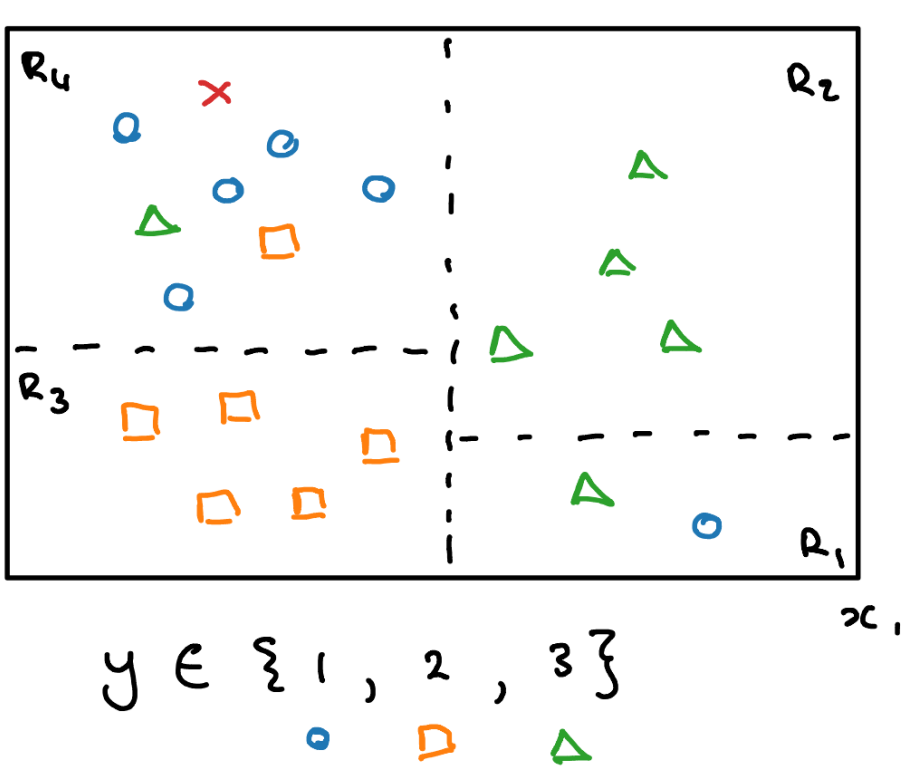
Using decision trees for classification:

Fraction of points in region R_m from class k:

$$\hat{p}_{m,k} = \frac{\#\text{points in } R_m \text{ with label } k}{\#\text{points in } R_m}$$

$$= \frac{1}{N_m} \sum_{i:\, \underline{x}^{(i)} \in R_m} I\{y^{(i)} = k\}$$

$$\hat{y}_m = \arg\max_k \hat{p}_{m,k}$$



Decision trees

Tree building algorithm

Herman Kamper

http://www.kamperh.com/

Tree growing algorithm:

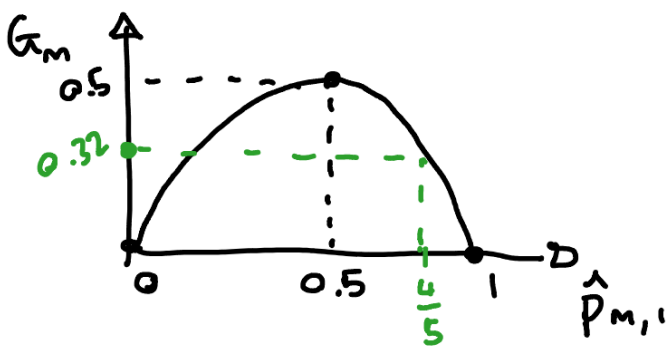
top-down

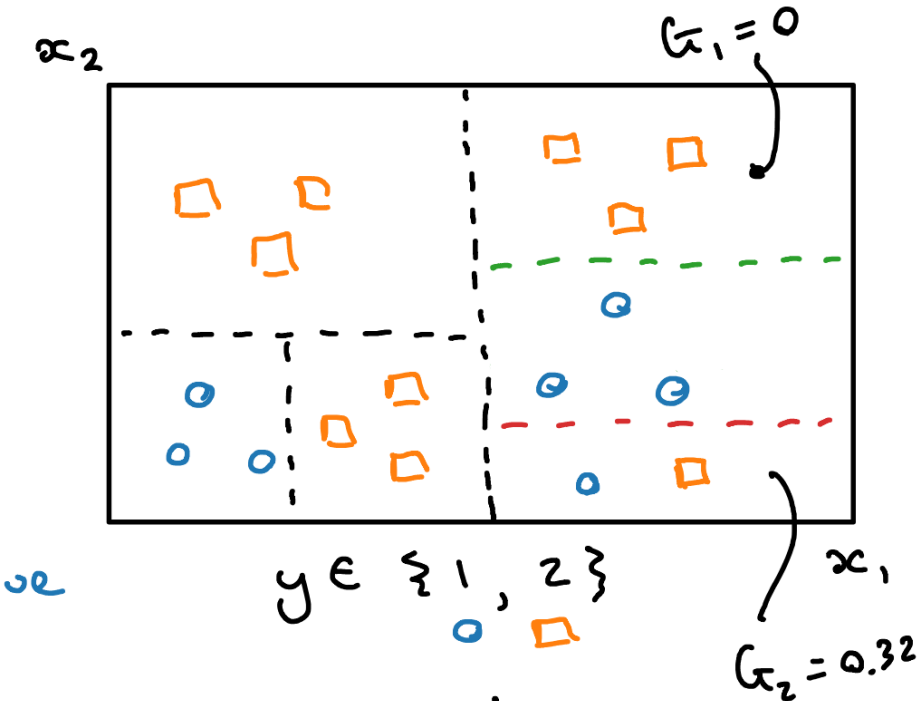
1. Start at top of tree (everything in one region)
2. for each leaf node (region):
   for each feature x_j and split point s:
   Calculate reduction in leaf impurity
3. Choose best (j, s) combination to split; create new child nodes (regions) ← greedy
4. Repeat from (2) until stop condition is met ← recursive

Gini index: $G_m = \sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$

For two classes: $\hat{p}_{m,1}$ & $\hat{p}_{m,2}$

$\hat{p}_{m,2} = 1 - \hat{p}_{m,1}$





$\hat{p}_{2,1} = \frac{4}{5}$

Classification error rate:

$E_m = 1 - \max_k \hat{p}_{m,k}$

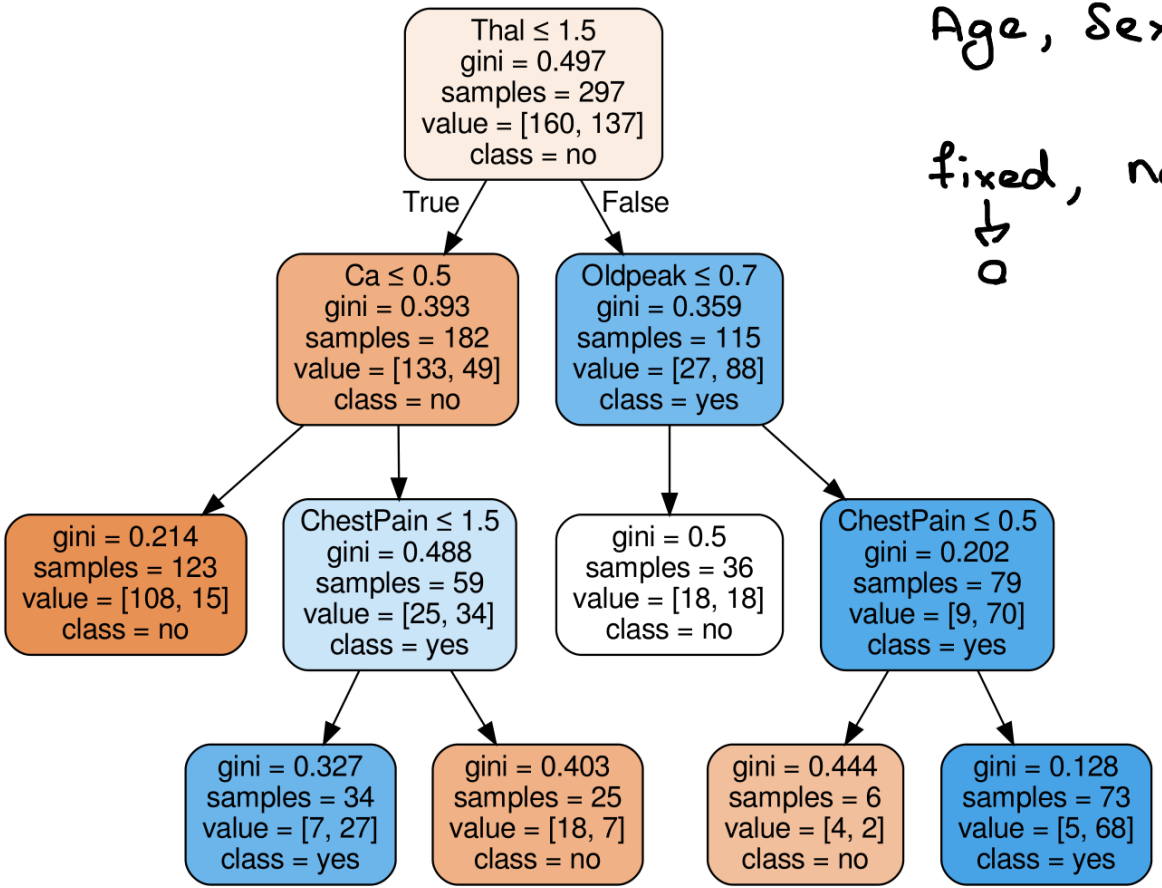
Decision trees

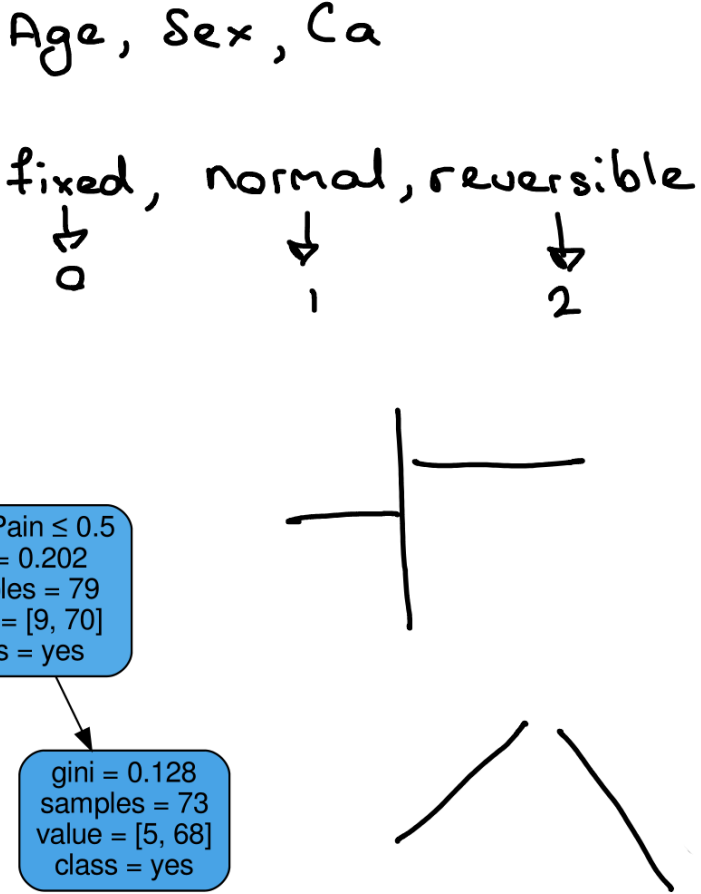
Decision trees in practice

Herman Kamper

http://www.kamperh.com/

Decision tree on heart data





Source: http://faculty.marshall.usc.edu/gareth-james/ISL/data.html