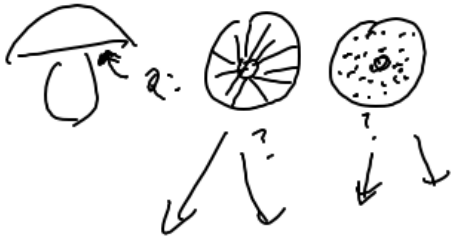
Decision trees

- ask several yes-no question and then make a prediction based on the answers

recognizing berries, mushrooms



Constructing the decision tree

$(x_1, x_2, \dots, x_p) \in \mathbb{R}^p$ – predictors ($p=2$ in the pictures)

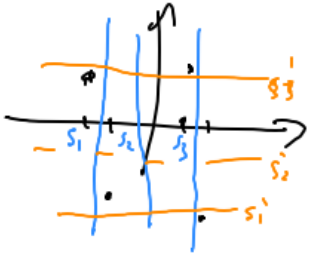
1) We divide the predictor space into J distinct non-overlapping regions $R_1, \dots, R_J$ (At the beginning $J=1$).

2) Consider all $R = R_j$ and all $i = 1, 2, \dots, p$ and all $s \in \mathbb{R}$ and:

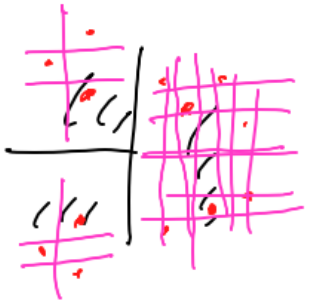
$$R_1(i,s) = R \cap \{(x_1, \dots, x_p): x_i \le s\}$$

$$R_2(i,s) = R \cap \{(x_1, \dots, x_p): x_i > s\}$$

$R_1(i,s) \cup R_2(i,s) = R$, $R_1^{(i,s)} \cap R_2^{(i,s)} = \emptyset$. It is enough to consider a finite number of thresholds s.



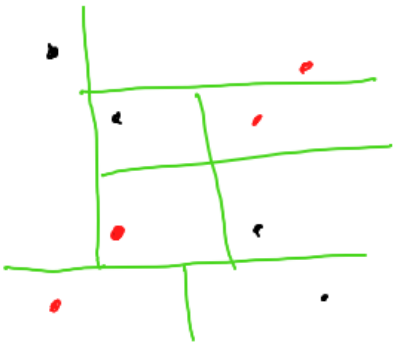
We take the „best" split and so we have $J+1$ regions.



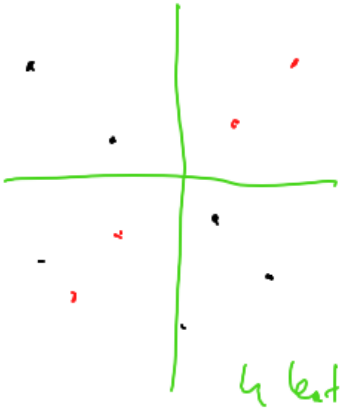
Repeat until: • classification: stop when in the leafs (there observations of just one class e.g. or when the number of obs. in leafs < threshold)

• regression: number of leafs < threshold of obs.

Greedy algorithm: optimises locally.



...but we find a tree with 8 leafs

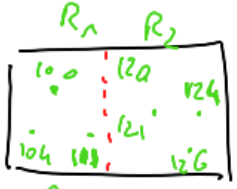


4 leafs are enough

For regression trees

we want to minimise (over all R_1, R_j, $i=1,\dots,p$, s)

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 + \sum_{h \neq j} \sum_{i:\, x_i \in R_h} (y_i - \hat{y}_{R_h})^2$$



$$\hat{y}_{R_1} = \frac{100+104+101}{3} = 105$$

For classification trees

$$\hat{p}_{mk} = \frac{\#\text{ samples in } R_m \text{ of class } k}{\#\text{ samples } R_m}$$

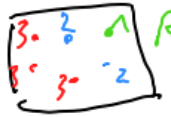
$$\hat{p}_{m1} + \dots + \hat{p}_{mK} = 1 \qquad (1, 2, \dots, K - \text{classes})$$

$$E = 1 - \max_k (\hat{p}_{mk}) \qquad \text{error of classification}$$

Gini: $$G = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) = \sum_k \hat{p}_{mk} - \sum_k \hat{p}_{mk}^2 = 1 - \sum_k \hat{p}_{mk}^2$$

entropy: $$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

$\log(x) \approx x-1$
$-\log x \approx 1-x$ for $x \approx 0$



$$\hat{p}_{m1} = \frac{1}{6}$$

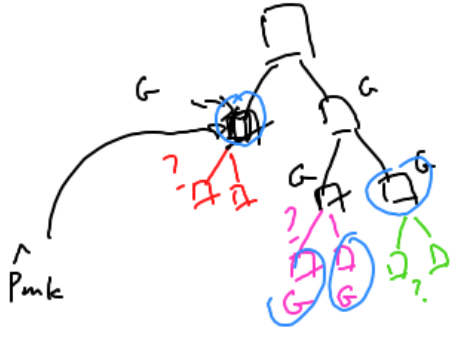
$$\hat{p}_{m2} = \frac{2}{6} = \frac{1}{3}$$

$$\hat{p}_{m3} = \frac{3}{6} = \frac{1}{2}$$

Say we decide to use G.

Best split: the one that minimises the weighted average value of G in leaves

Weights are given by number of samples.



After growing the tree we prune it.

Cost complexity pruning : (regression trees)

$$\text{loss function } L_\alpha(T) = \sum_{m=1}^{|T|} \underbrace{\sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2}_{\text{for classification: replace by e.g. } G \text{ for the leaf } m \; G(m) \text{ with the weight given by the number of samples in the leaf, } G(m) \cdot \frac{\#\{i : x_i \in R_m\}}{\#\{i\}} \text{ (all samples)}} + \underbrace{\alpha |T|}_{\text{penalty for the size of the tree } (\alpha \geq 0)}$$

$|T|$ = number of leaves in the tree T

For each inner node N we calculate the smallest α for which the loss function :

$$L_\alpha(T) = L_\alpha(T \text{ with the subtree under the node } N)$$

we take the smallest and prune the tree at the corresponding node $\rightsquigarrow$ we obtain a new tree T_2

We repeat with the new tree.

That way we obtain a sequence of trees $T = T_1, T_2, \ldots, T_N$
$\|$
root