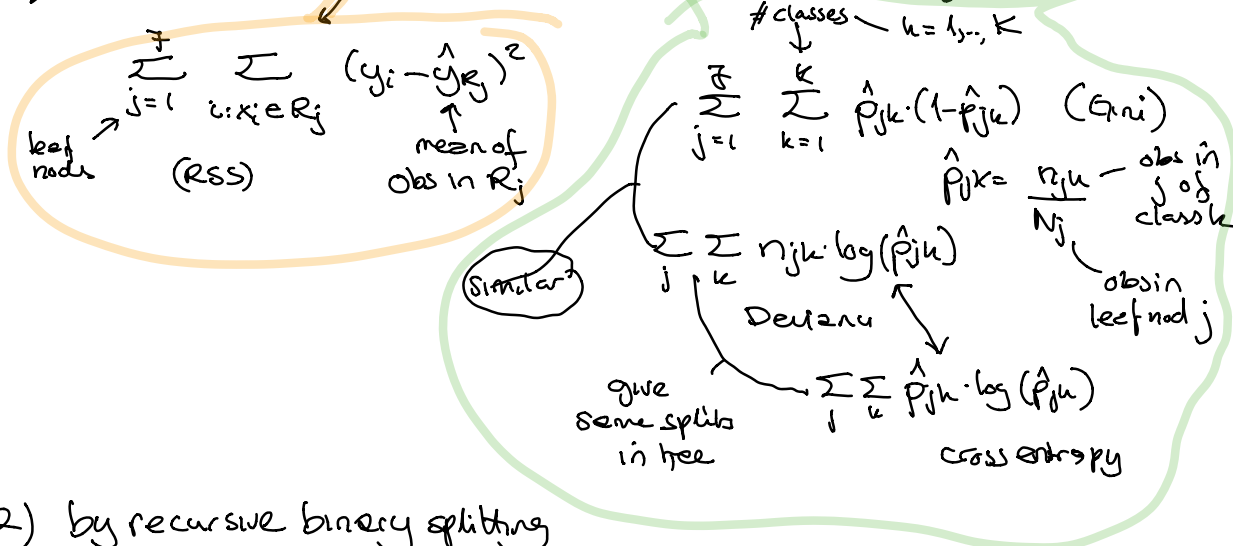


→ both for regression and classification

1) Tree constructed by minimizing a criterion



2) by recursive binary splitting
- greedy approach

$$R_1(j, s) = \{x \mid x_j < s\} \text{ and } R_2(j, s) = \{x \mid x_j \geq s\}$$

until some stopping criterion is reached.

3) (full) tree used for prediction by dropping (new) observations down the tree and predict as

$$\hat{y} = \frac{1}{N_j} \sum_{i: x_i \in R_j} y_i$$

regression

$$[x_i, y_i \quad i=1, \dots, n \text{ training data}]$$

\hat{y} = majority vote among y 's in R_j

or

$$\hat{p}_{jk} = \frac{1}{N_j} \sum_{i: x_i \in R_j} I(y_i = k)$$

n_{jk}

Observe: - linear boundary: hard
- interaction: easy

Pruning

- overfitting
 - unnecessary splits
 - interpretability
- } reduce the number of leaf nodes

Many possible pruned trees \rightarrow we use cost complexity pruning

$$C_{\alpha} = Q(T) + \alpha |T|$$

Annotations for the equation above:

- T : subtree
- T_0 : full tree
- $Q(T)$: cost function (RSS, Gini, miscl. rate)
- α : penalty
- $|T|$: size of tree (number of leaf nodes)

Given $\alpha \Rightarrow$ get a pruned tree \Rightarrow See ALG on Module page

\rightarrow Bias-variance tradeoff:

BAGGING

$f^{*b}(x)$ = decision tree from bootstrap sample b

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^B f^{*b}(x)$$

Make bushy trees. B need to be large enough (500-2000)
Out of bag OOB estimation of error.

It is hard to examine B tree for interpretation, instead we look at "variable importance plots".

RANDOM FOREST

Bagging may not help so much when we have a strong predictor and all our B trees are very similar
→ need to decorrelate trees.

only consider $m < p$ predictors at each split
in the tree

\sqrt{p} classification } tuning parameter
 $p/3$ regression }

B needs to be large enough (not a tuning parameter)

BOOSTING

(only regression trees here)

$$\hat{f}(x) = 0, \quad r_i = y_i$$

$$b = 1, \dots, B$$

$\left\{ \begin{array}{l} d = \# \text{ splits} \\ B = \# \text{ trees} \\ \lambda = \text{learning rate} \end{array} \right.$
tuning parameter

$\hat{f}^b(x)$ fitted with d splits to r_i -data

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \cdot \hat{f}^b(x)$$

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

$$\hat{f}(x) = \sum_{b=1}^B \lambda \cdot \hat{f}^b(x)$$

NB: no bootstrapping here \rightarrow the full dataset is used sequentially (so B is number of sequentially grown trees)