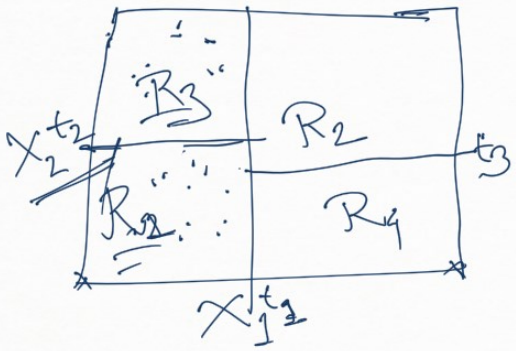
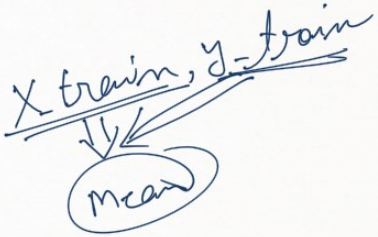
CART ⇒ Classification & Regression Trees.

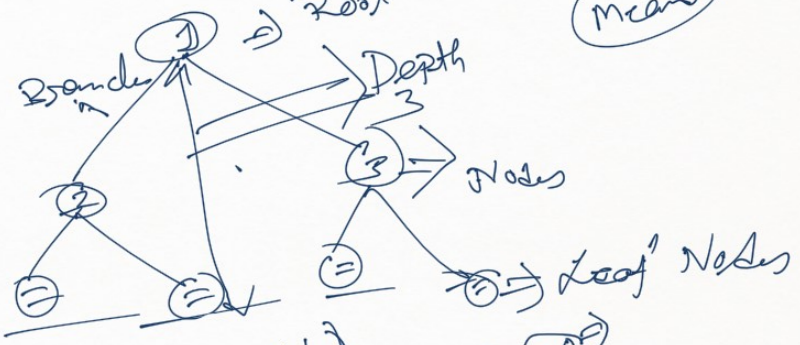
Regression ⇒

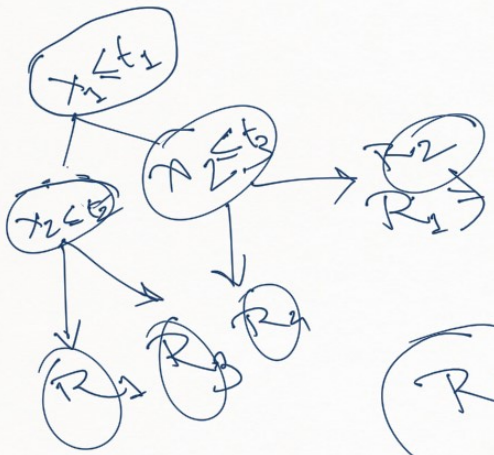


⇒ Recursively Split

Binary Tree



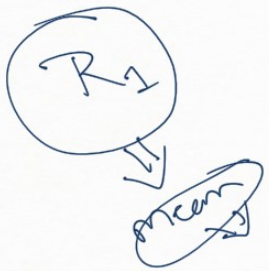




$R_2 \Rightarrow X_1 > t_1$

$R_1 \Rightarrow X_1 \le t_1$

R_2



$R_1 \in \{X \mid X_1 \le t_1 \& X_2 \le t_2\}$

$\bar{y}_{R_1} \Rightarrow$

Output = J Regions

$\sum_{i=1}^{J} \sum_{x \in R_i} (\bar{y}_{R_i}) \Rightarrow$ Output.

Dividing input space into K-parts ⇒ Cross-validation

Based on feature values divide input space

$\Downarrow$

$(y_{R_i}) \Rightarrow (y_{act} - y_{pred})^2$

J

$MSE \Rightarrow \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$

$R_1, R_2 \Rightarrow MSE \Rightarrow \sum_{i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i \in R_2} (y_i - y_{R_2})^2$

Cost Complexity Pruning ⇒ Prune a decision tree ⇒

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 + \alpha |T|$$

⇒ Depth of a tree

$S \Rightarrow T_0$ —

α ⇒ Observed parameter

$i \in R_j$ $\hat{y}_{R_j}$ ⇒

K-class

p_K = Probability of $x_i \in$ Class-K

$$\text{Entropy} = -\sum_{i=1}^{K} p_K \log(p_K)$$

$$\text{Gini Index} = \sum_{i=1}^{K} p_K (1 - p_K)$$

$P(A) + P(B) = 1 \Rightarrow P(B) = 1 - P(A)$

$P(A) \log P(A) + (1 - P(A)) \log (1 - P(A))$

$0,1 \Rightarrow P(0) + P(1) = 1 \;=\; P(0) = 1 - P(1)$

Entropy $\Rightarrow -P_0 \log P_0 - P_1 \log P_1 = -P_0 \log P_0 - (1-P_0)\log(1-P_0)$

$P(0) = 1 = x_i \in$ Class 0

Class 0 $\Rightarrow$ all values

Class 1 $\Rightarrow$ all values.

$P_0 = P_1 = 1/2 = 2^{-1} \Rightarrow$

$E \Rightarrow -1/2\left(\log_2 2^{-1}\right) - (1/2)\left(\log_2 (2)^{-1}\right)$

$= 1/2 + 1/2 = 1$

$-1 \log 1 = 0$

$-0.6(\log_2 0.6) - 0.4(\log_2 0.4)$

$= 0.6 \times 0.737 + 0.4 \times 1.322$

$= 0.97$

$P_0 = 1, P_1 = 0$

Purest

Messiest / Impurest

Better than 0.5

0.5

0.6

0.4

0.6

0.4

Gini Index $= \sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k)$

$\Rightarrow p_0(1-p_0) + p_1(1-p_1) = p_0(1-p_0) + (1-p_0)(1-1+p_0)$

$= 2p_0(1-p_0)$

$p_0 = 0.5$ $\quad 2 \times 0.9 \times 0.1 = 0.18$ $\quad = 2 \times 0.5 \times 0.5 = 0.5$

$p_0 = 0.7$ $\quad = 2 \times 0.7 \times 0.3 = 0.42$

$p_0 = 1$ $\quad = 2 \times 1 \times (1-1) = 0$

$p_0 = 0.9$

K = 3 $\quad P(A), P(B), P(C)$

Entropy : $-P(A)\log P(A) - P(B)\log P(B) - P(C)\log P(C)$

Gini $\Rightarrow P(A)(1-P(A)) + P(B)(1-P(B)) + P(C)(1-P(C))$

$P(A) = P(B) = P(C) = 0.33$ $\Big\}$ $\quad 3 \times 0.33 \times 0.67$

$P(A) = 0, P(B) = 0, P(C) = 1$ $\quad \simeq 0.67$

$P(A) + P(B) + P(C) = 1$

$P(C) = 1 - P(A) - P(B)$

Ensemble methods =>

Combining multiple trees.

Bagging => Taking diff. bags of input data

10 => 10 => 10

(A) (15)

Out of Bag Error =>

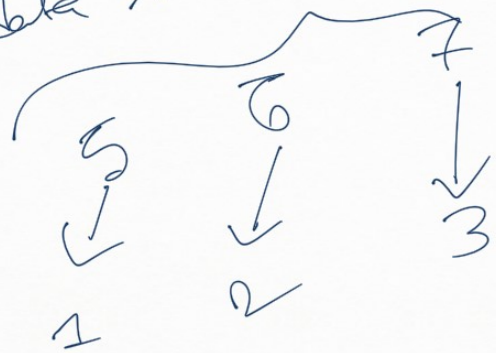
OOB-score

(10) => Bag 1

(B) => Out of Bag 1

Random Forest =)

Input data bootstraps.



Estimators.