3-5-2025

Μέθοδοι ταξινόμησης σε δέντρα (Decision Trees)

$\hat{y} = f(x)$: συνάρτηση πρόβλεψης

$$f(x) = \sum_{m=1}^{M} c_m \cdot 1(x \in R_m)$$

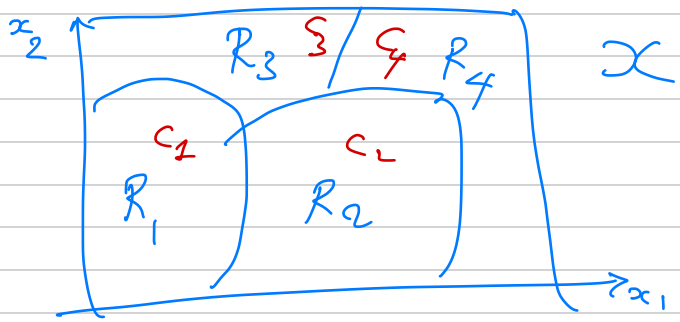
c_m : σταθερά , $m = 1, \dots, M$

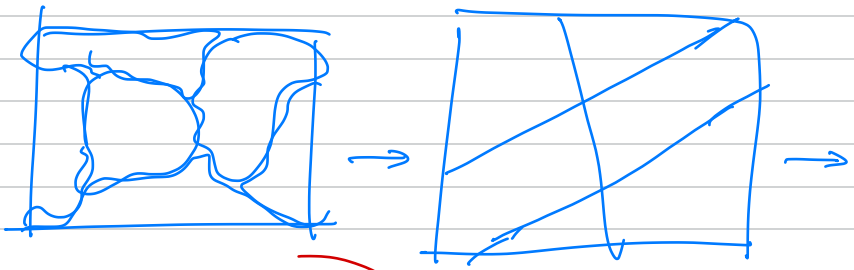
$R_m \subseteq X$ πεδίο τιμών των $(X_1, \dots, X_p)$

$R_{m_1} \cap R_{m_2} = \emptyset \quad m_1 \neq m_2$

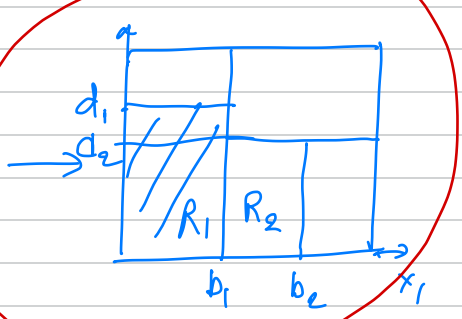
$\bigcup_{m=1}^{M} R_m = X$

$\{R_1, \dots, R_M\}$ διαμέριση του X.





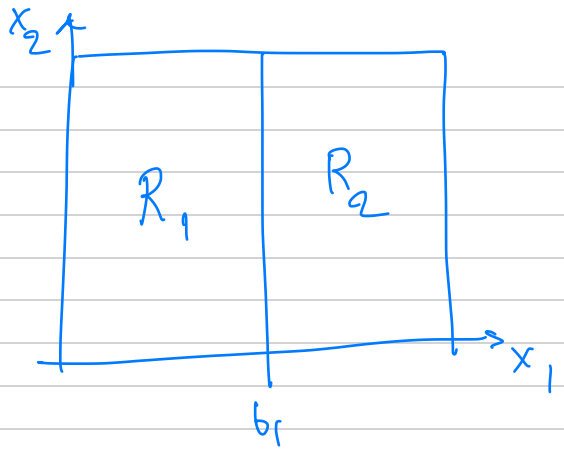
decision tree ⟹



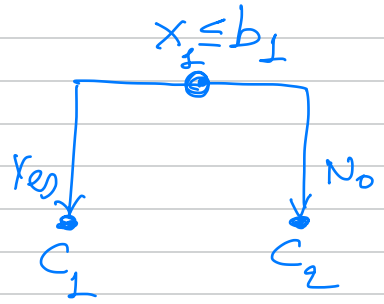
R_m : ορθογώνια.

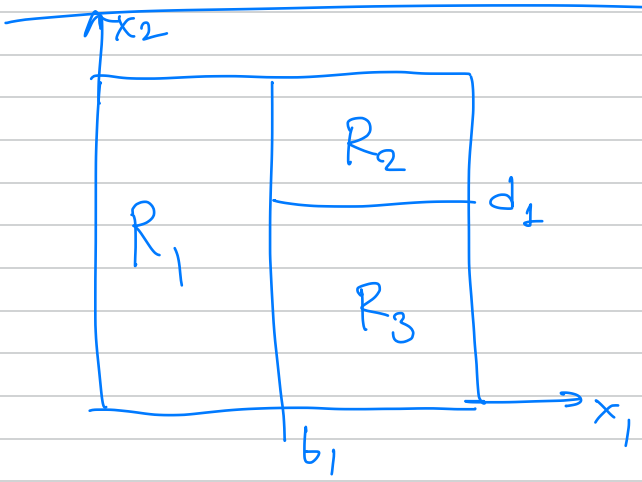
R_1 : $0 \leq x_1 \leq b_1$
$0 \leq x_2 \leq d_1$

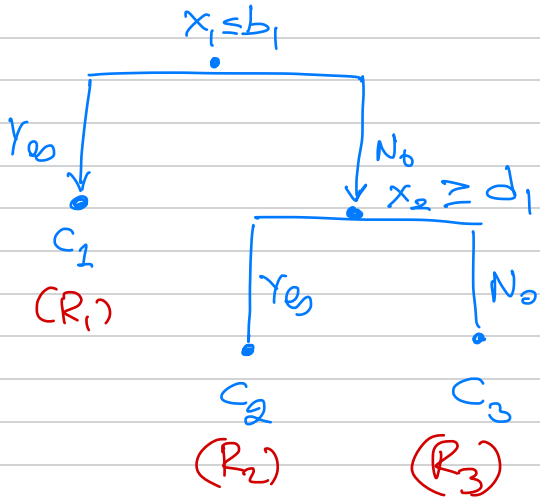
R_2 : $b_1 < x_1 \leq b_2$
$0 \leq x_2 \leq d_2$

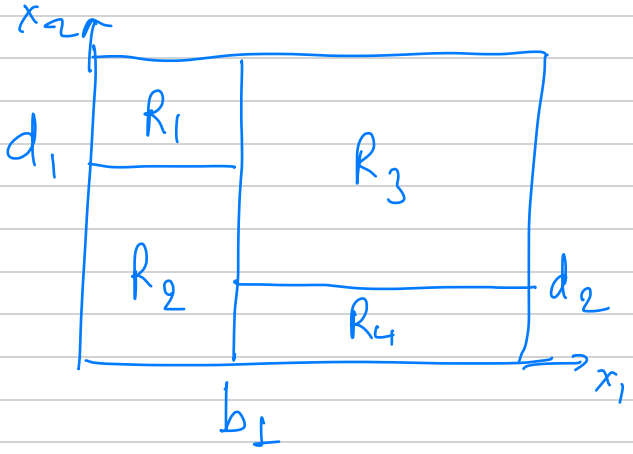


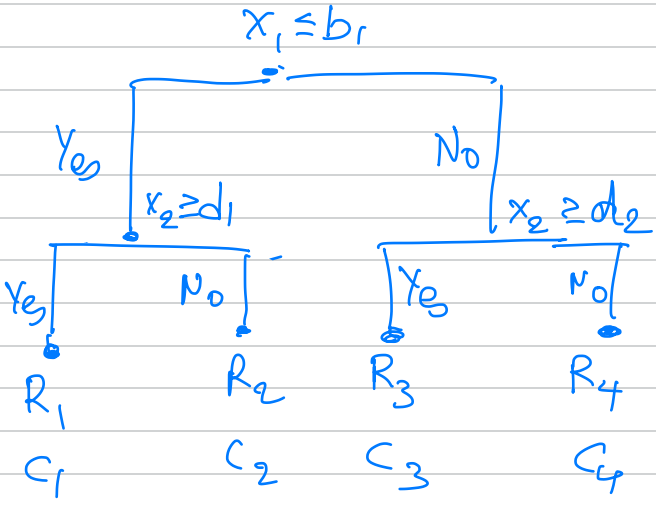
$$f(x) = \begin{cases} c_1 & \text{av } x_1 \le b_1 \\ c_2 & \text{av } x_1 > b_1 \end{cases}$$

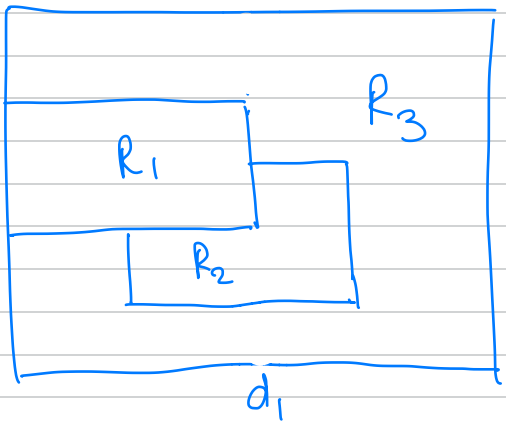




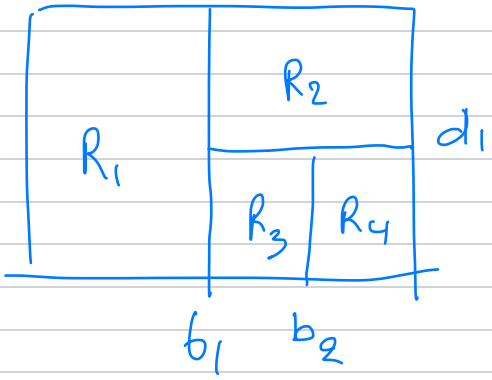


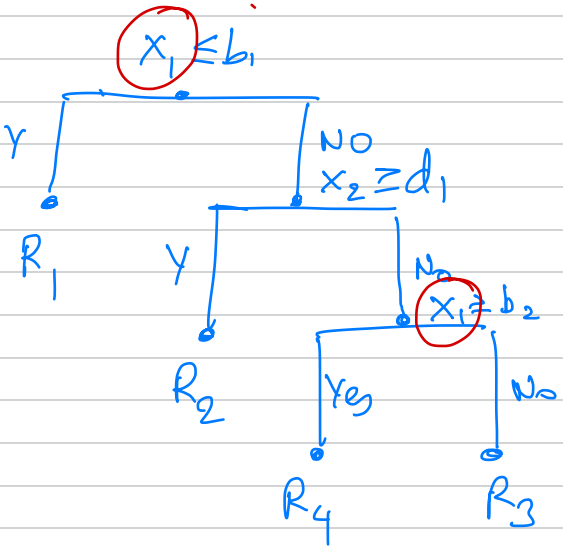


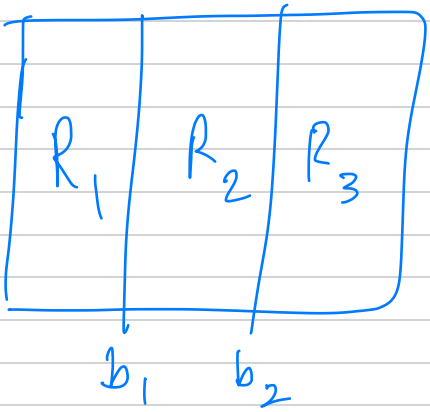


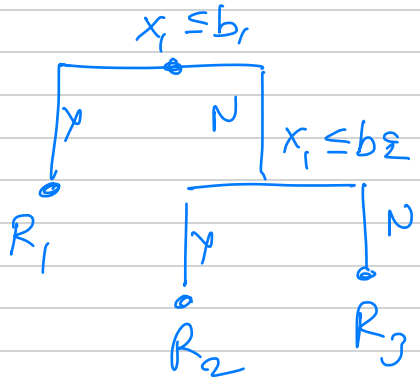


δεν μπορεί σαν γραφικό ορίζω decision tree



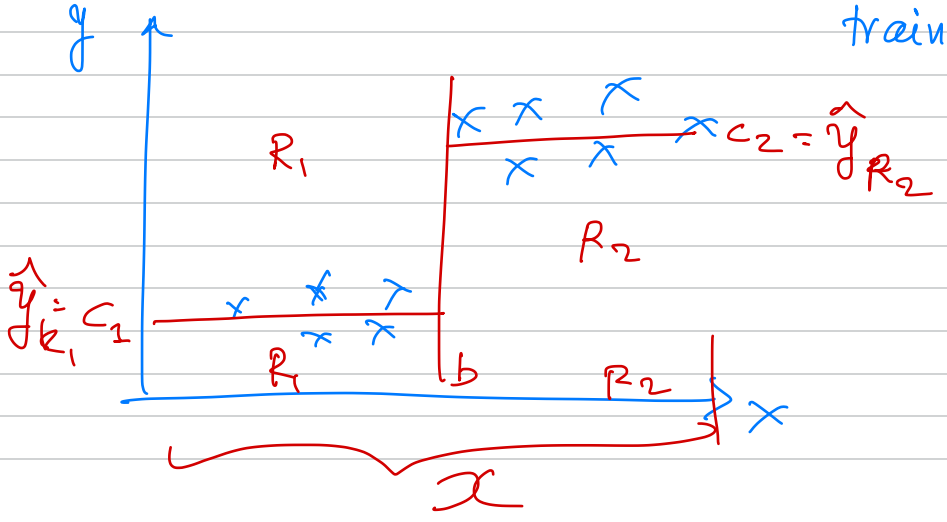






---

training set



Regression trees (y: ποσοτική)

M περιοχές $R_1, \dots, R_M$

$\hat{y}_R$ = πρόβλεψη για την περιοχή R_m

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2$$

$\left(\forall i \in R_m : \hat{y}_i = \hat{y}_{R_m} \right)$

$\forall m$: $\sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2 \to \min$ για

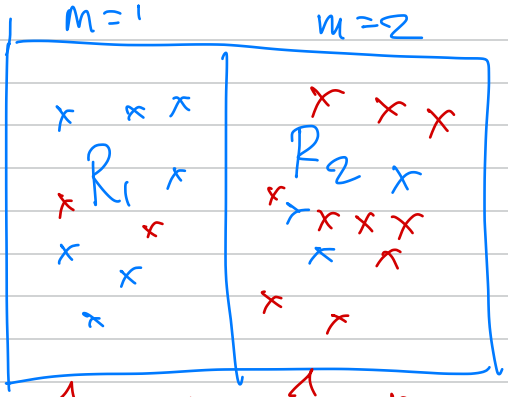
$$\hat{y}_{R_m} = \frac{\sum_{x_i \in R_m} y_i}{\sum_{i=1}^{n} 1(x_i \in R_m)}$$

$i \in R_m$ πρέπει $x_i \in R_m$

Classification Trees

Y κατηγορική $\in \{1, 2, \dots, K\}$ classes

π.χ. $Y \in \{$μπλε, κόκκινο$\}$ $(K=2)$



Gini

$\hat{p}_{1,\text{μπλε}} = \frac{9}{11}$

$\hat{p}_{1,\text{κόκ}} = \frac{2}{11}$

$\frac{2}{11} \cdot \frac{9}{11} + \frac{9}{11} \cdot \frac{2}{11}$

$\hat{y}_{R_1}$ = Μπλε $\hat{y}_{R_2}$ = Κόκ

Έστω $\hat{p}_{mk}$ = % παρατηρ. στο R_m με $y = k$

$k = 1, \dots, K$

$m = 1, \dots, M$

$$\sum_{k=1}^{K} \hat{p}_{mk} = 1 \quad \forall m$$

$\forall m$

$$\hat{y}_m = k \iff \hat{p}_{mk} = \max_{j=1,\dots,K} \hat{p}_{mj}$$

[Logistic regr:

$$\forall x \Rightarrow \hat{p}_k(x) = P(Y = k \mid X = x)$$

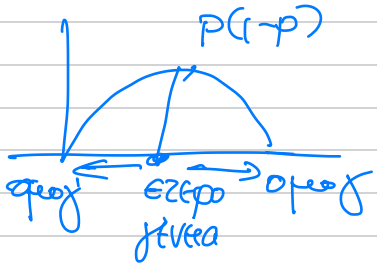
$$\hat{y} = k \iff \hat{p}_k(x) = \max_j \hat{p}_j(x)$$

Κριτήριο για fit

1) Classification Error

$$\frac{1}{n} \sum_{i=1}^{n} 1(y_i \neq \hat{y}_i)$$

2) Gini Index $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, $m = 1, \dots, M$



3) Εντροπία $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Δημιουργία Διαμερίσεων

Recursive Binary Splitting (greedy algorithm)

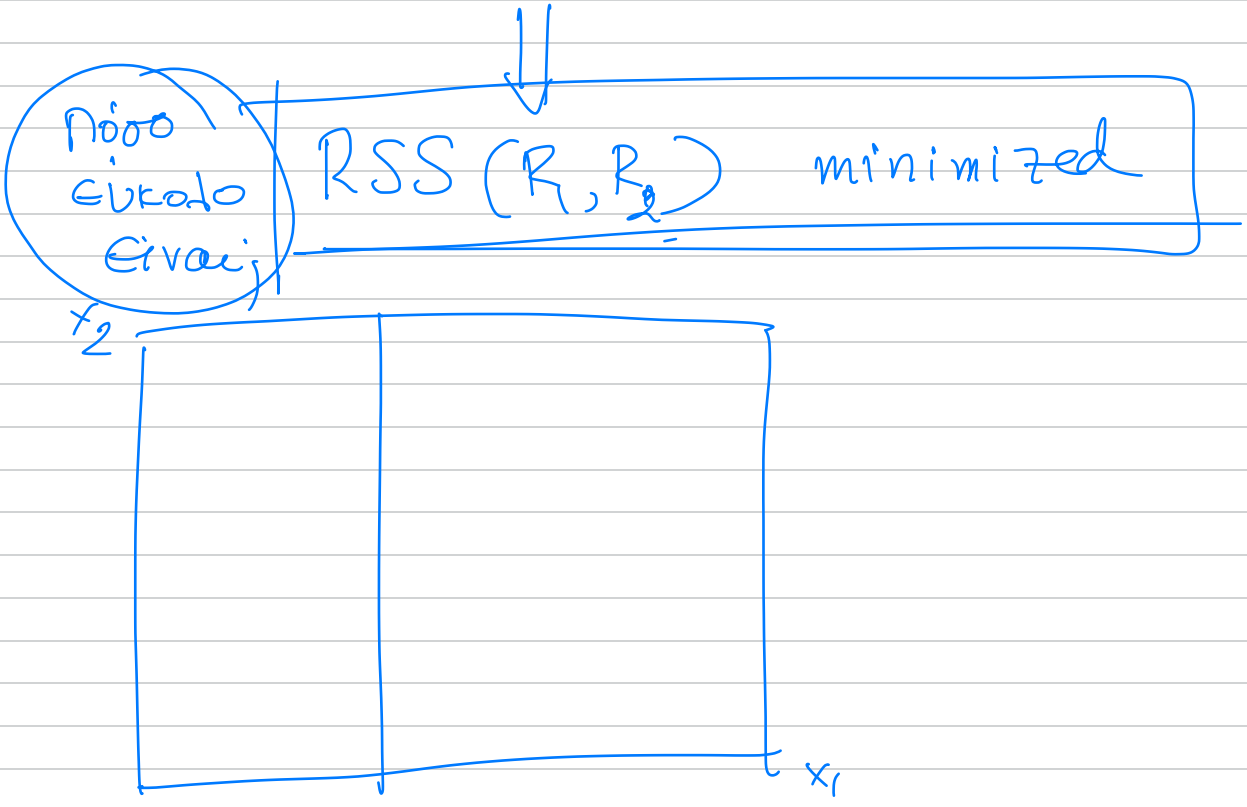
Αρχικά

Επιλέγουμε μια μεταβλητή X_j } ταυτόχρονα
κ' ένα σημείο αποκοπής S_j }

δηλαδή αν κάνουμε τη διαμέριση

$X_j \le S_j$ κ' $X_j > S_j$

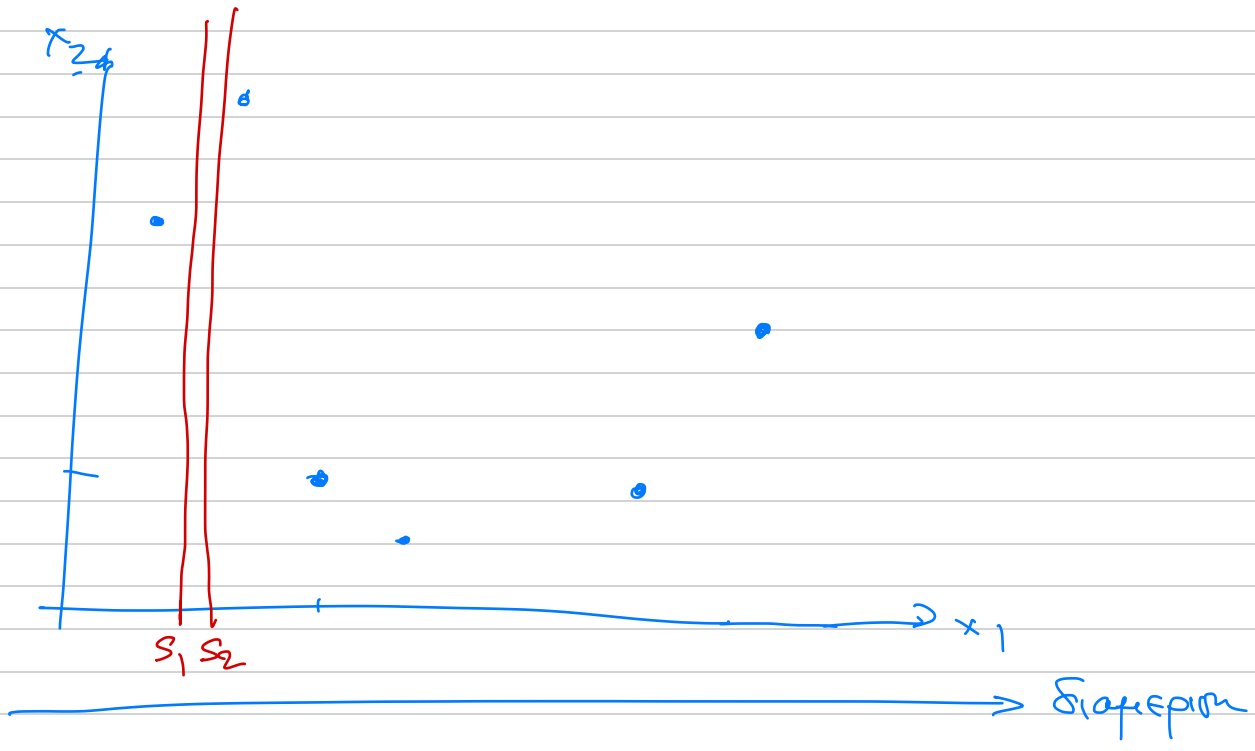
R_1 R_2



Εδώ θα επιλέγουμε τον X_1

training set

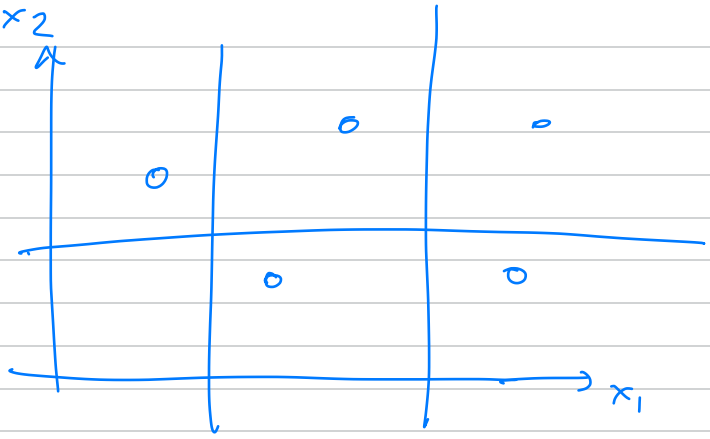
X_1	X_2	Y
3	2	.
4	1	1
1	5	1
2	4	
6	2	1
4	4	1



$$RSS(s_1) = RSS(s_2)$$

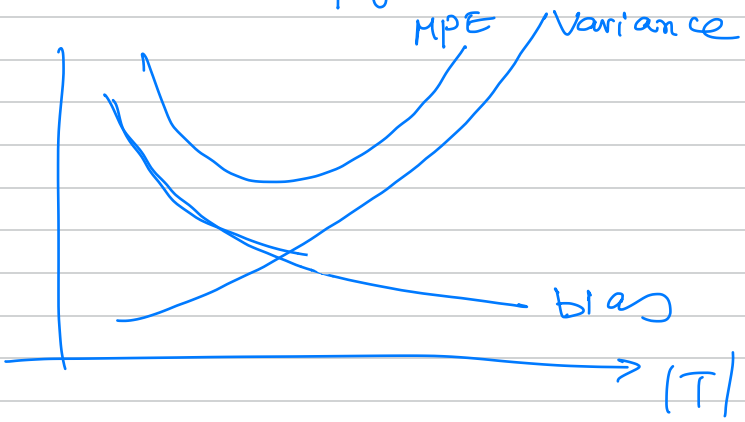
Πότε σταματάμε;

Περιορισμός ελάχιστου αριθμού παρατηρήσεων σε κάθε R_m,



$T = \{\text{τερματικοί κόμβοι}\}$ $|T| = M$ (μέγεθος δέντρου)

Μεγάλα δέντρα : μικρό bias μεγάλο variance.



Pruning

Πρώτα δημιουργούμε το (μεγάλο) δέντρο T_0

Κλάδεμα Έστω παράμετρος ποινής α

Θέλουμε $\min_{T \subset T_0} \left[RSS + \alpha |T| \right]$

α : υπερπαράμετρος