Lecture November 5

K-means clustering

Data set $\{x_0, x_1 \dots x_{m-1}\}$ of n-observations of a D-dimensional random variable X

The aim is to find a partition of the data into some numbers of clusters $-K-$

- Need to find assignments of points belonging to a given cluster $-k-$
- Define the cluster by its position μ_K (center)

Optimize

$$C = \sum_{i=0}^{n-1} \sum_{k=0}^{K-1} r_{ik} \|x_i - \mu_k\|_2^2$$

r_{ik}

$r_{ik} = \begin{cases} 1 & \text{if within} \\ 0 & \text{else,} \end{cases}$

$$r_{ik} = \begin{cases} 1 & \text{if } k = \arg\min \|x_i - \mu_k\| \\ 0 & \text{else} \end{cases}$$

- First choose some initial value μ_k (K) optimize r_{ik} while keeping μ_k fixed.

- with r_{ik} fixed, optimize μ_k

Derivatives wrt μ_k

$$2 \sum_{i=0}^{n-1} r_{ik} (x_i - \mu_k) = 0$$

$$\mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

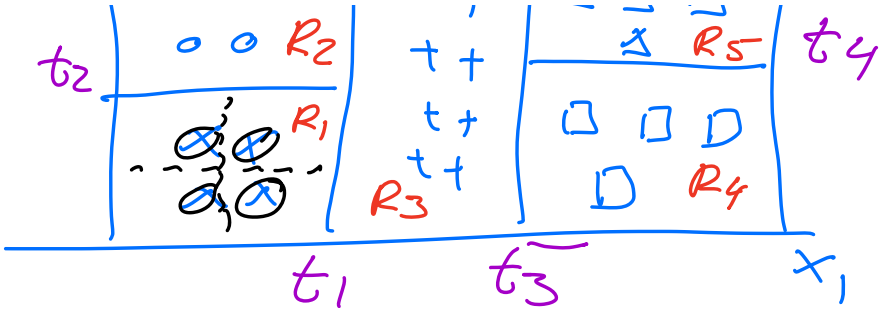
$\sum_i r_{ik}$: # points in a cluster.

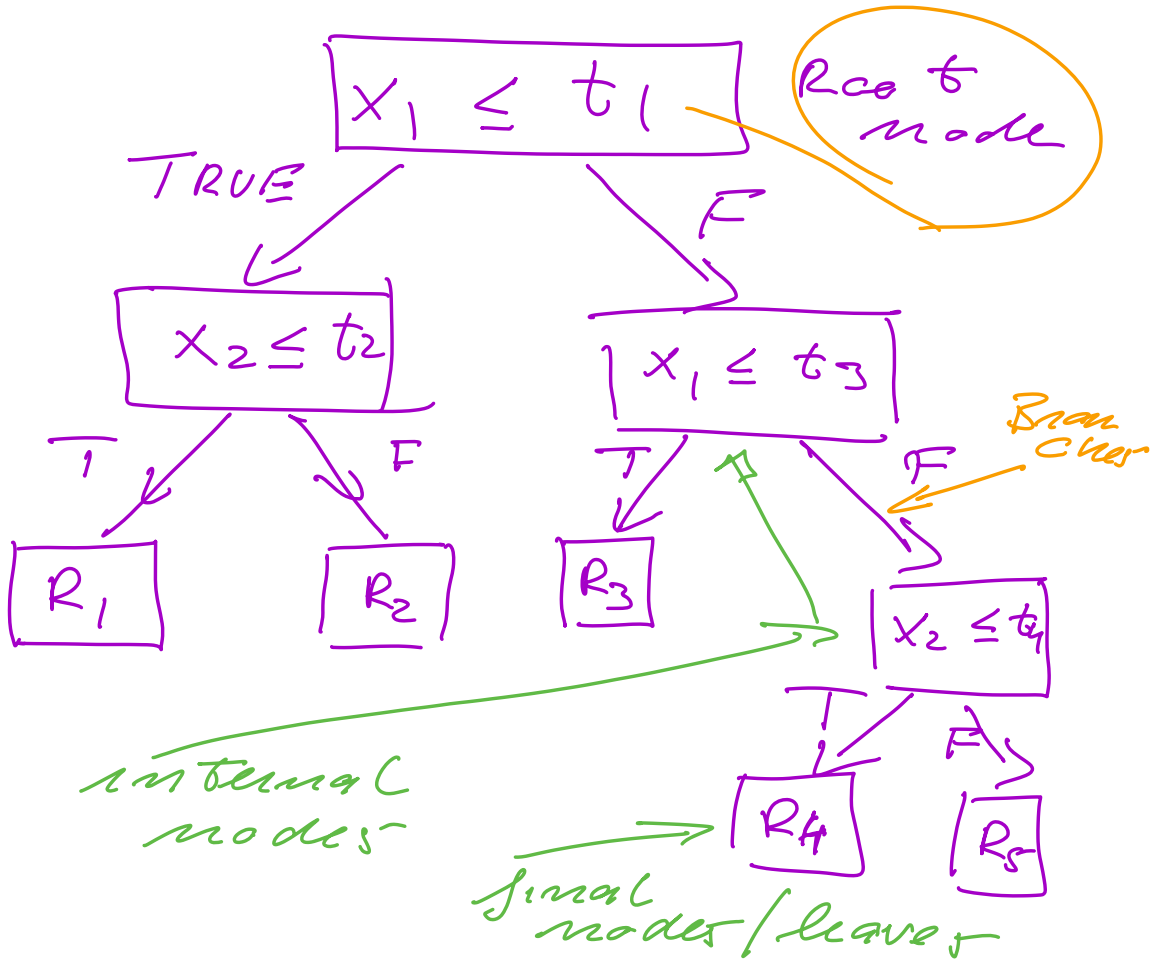
The values of μ_k are defined by the mean values defined by the data points in each cluster.

Decision Trees

ensemble methods:
- Bagging (mean of forest)
- random forests (different trees)
- Voting
- Boosting, gradient boosting

Typical Regression case


t_2
R_2
R_1
R_3
R_5
t_4
R_4
t_1
t_3
x_1


$x_1 \leq t_1$
Root node
TRUE
F
$x_2 \leq t_2$
$x_1 \leq t_3$
T
F
T
F
Branches
R_1
R_2
R_3
$x_2 \leq t_4$
T
F
internal nodes
R_4
R_5
final nodes / leaves