Lecture october 29

PCA Theorem

$X \in \mathbb{R}^{n \times p}$

$$\Sigma_x = \frac{1}{n} X^T X \in \mathbb{R}^{p \times p}$$

$$= \mathbb{E}[X^T X]$$

$X \in \mathbb{R}^{p \times n}$

$$\Sigma_x = \mathbb{E}[X X^T]$$

Define an orthogonal transformation

$$S = [s_0 \, s_1 \, s_2 \, \dots \, s_{p-1}]$$

$$S \in \mathbb{R}^{p \times p}$$

$$S S^T = S^T S = \mathbb{1}$$

$$SX = y$$

X is a scaled matrix with subtracted mean value

$$\mathrm{var}(SX) =$$

$$= \mathbb{E}[(SX)^2] = \mathbb{E}[S X X^T S^T]$$

$$X \in \mathbb{R}^{p \times n}$$

$$= S \Sigma_X S^T = \Sigma_Y$$

$$= \mathrm{diag}(\lambda_0, \lambda_1 \ldots \lambda_{p-1})$$

$$\Sigma_X \cdot s_0 = \underset{\substack{\parallel \\ \mathrm{var}(y_0)}}{\lambda_0} s_0 = \mathrm{var}(y_0)\, s_0$$

PCA - theorem:

want $d < p$

$$y_i = s_i X \qquad i = 0, 1, 2 \ldots, p-1$$

$$\lambda_0 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_p$$

$$\mathrm{var}(y_0) = \lambda_0 \geq \mathrm{var}(y_1) \geq \ldots \mathrm{var}(y_{p-1})$$

$\mathrm{var}(y_0) = \lambda_0$ with eigenvector s_0

$$\max_{s_0 \in \mathbb{R}^p} \; s_0 \Sigma_X s_0^T$$

subject to $s_0^T s_0 = \underline{1}$

Constrained optimization (support vector machines)

Define Lagrangian

$$\mathcal{L} = s_0 \Sigma_X s_0^T + \lambda_0 (1 - s_0^T s_0)$$

compute derivatives wrt s_0 and λ_0

λ_0 ; $\quad s_0^T s_0 = \underline{1}$

$\Sigma_x s_0 = \lambda_0 s_0$

optimal solution for s_0 is given by the eigenvector of Σ_x.

First principal component.

To find the second principal component;

$$s_1^T s_0 = \underline{1}$$

$$\mathbb{E}[s_0 x x^T s_1^T] = s_0 \Sigma_x s_1^T$$

$$= \lambda_0 s_0 s_1^T = 0$$

$$\mathcal{L} = s_1^T \Sigma_x s_1 + \lambda_1 (1 - s_1^T s_1) + \gamma s_1^T s_0$$

s.t. $\quad s_1^T s_1 = \underline{1} \qquad s_1^T s_0 = 0$

Taking derivatives wrt s_1, λ_1 γ

$\gamma = 0$

$\Sigma_x s_1 = \lambda_1 s_1$. The best choice i.e. $\lambda_1 = var(y_1)$

We can construct remaining by induction.

SVD

$X \in \mathbb{R}^{n \times p}$

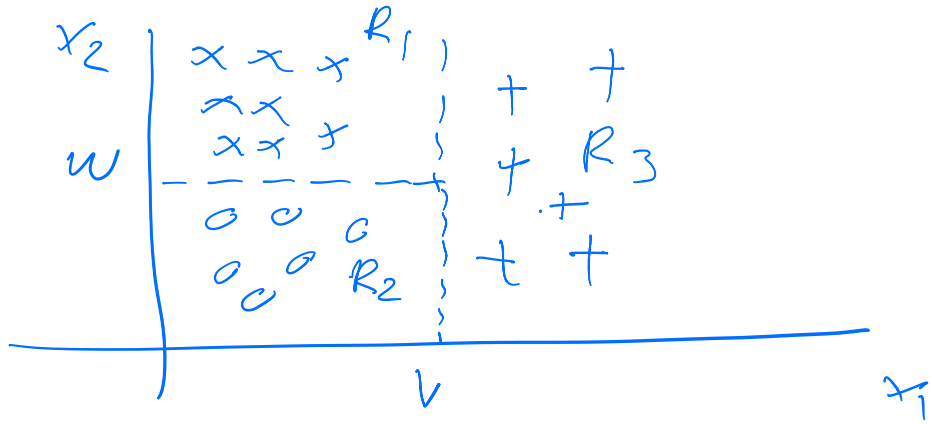
$$S = \begin{bmatrix} \sigma_0 & & \\ & \ddots & \\ & & \sigma_{p-1} \\ \hline & & 0 \\ & & 0 \\ & & \vdots \end{bmatrix}$$

$$X = \underbrace{U}_{n \times n} \; \underbrace{S}_{n \times p} \; \underbrace{V^T}_{p \times p}$$

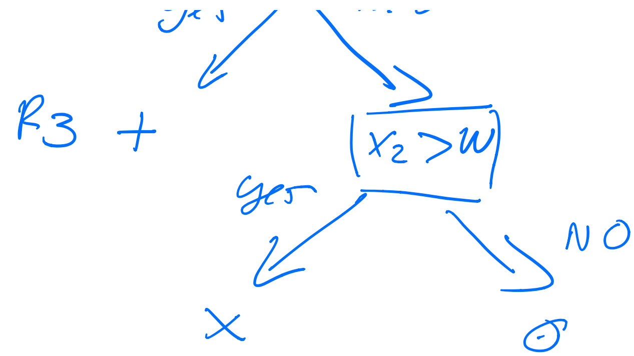
Trees, Forests, jungles and more

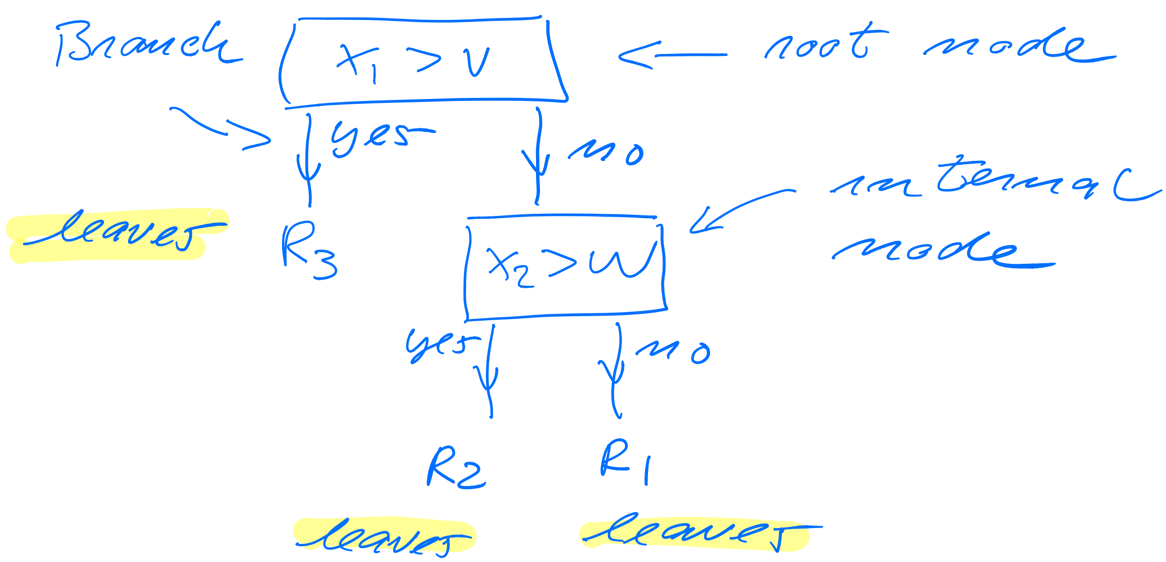
Decision Trees, Regression:

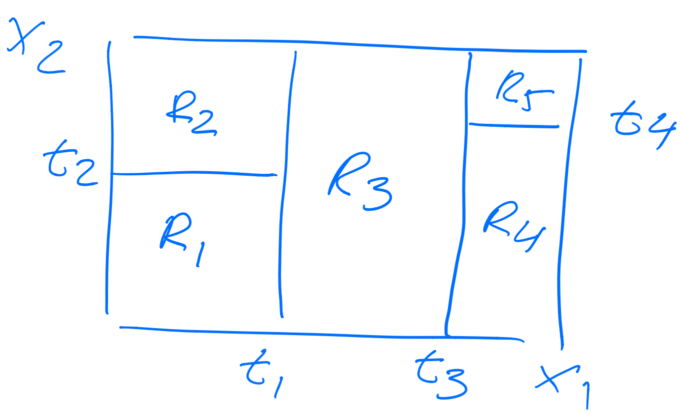
x_1, x_2

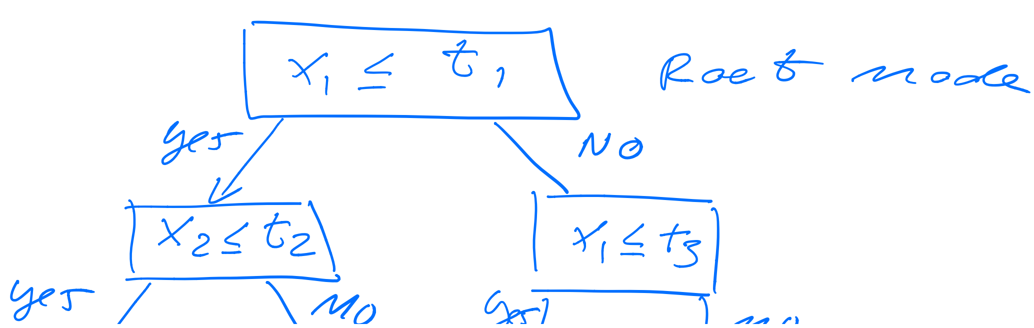


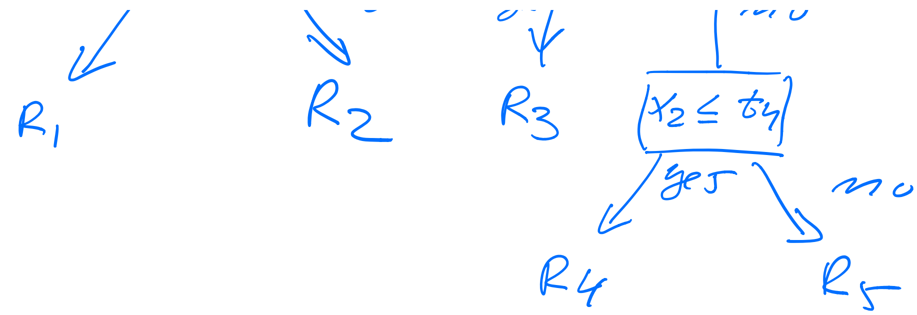
if $x_1 > v$


R3 +
x2 > W
yes
NO
X
O


Branch
x1 > V
root node
yes
no
internal node
leaves
R3
x2 > W
yes
no
R2
R1
leaves
leaves


X2
R2
t2
R1
R3
R5
R4
t4
t1
t3
x1


x1 ≤ t1
Root node
yes
NO
X2 ≤ t2
x1 ≤ t3
yes
no
yes



Algorithm for regression case

i) Divide the space into sets of possible values for x_1, x_2 .... x_p into K distinct and non-overlapping regions $R_1, R_2, \ldots R_K$

(ii) For every observation that falls into region R_i, the prediction is given by the mean value of the observations in R_i

Example: R_1, R_2

if mean of $R_1 = \mu_1 = 5$

and $R_2 = \mu_2 = 10$

an observation that falls in R_1 gets a prediction of 5

For R_2, the prediction is 10

How do we construct $R_1, R_2 .. R_K$?

Basic approach is to find boxes where we minimize

$$\sum_{j=1}^{K} \sum_{i \in R_j} (y_i - \overline{y}_{R_j})^2$$

$\overline{y}_{R_j}$ = mean value within the jth-box.

Time consuming.

Rather: start with all observations and define single region. Then successively split into smaller regions.

- Define a cut point s
  Define regions

$$R_1(j,s)\{X \mid X_j < s\} \text{ and } \underbrace{\{X \mid X_j \geq s\}}_{R_2(j,s)}$$

- Minimize

$$\sum_{j \in R_1} (y_j - \overline{y}_{R_1})^2 + \sum_{j \in R_2} (y_j - \overline{y}_{R_2})^2$$