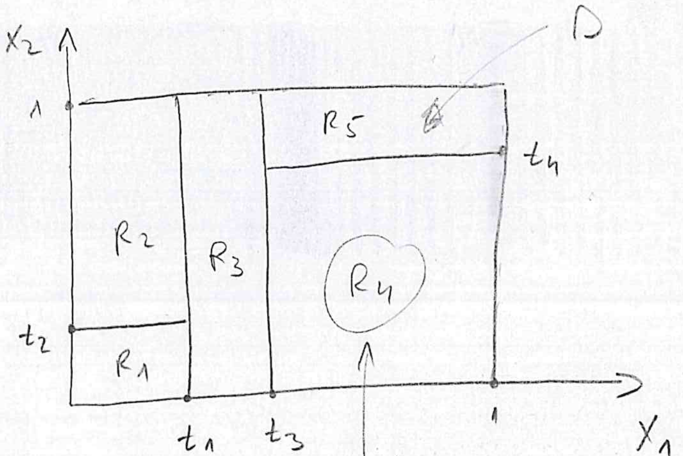
8.) Metode bazirane na stablima

8.1) CART (Classification and Regression Trees)

- dijelimo prostor prediktora $D \subseteq \mathbb{R}^p$ na p-dimenz. pravokutnike te prilagođavamo jednostavan model na svakom pravokutniku
- koristimo rekurzivne binarne particije

Pr. 8.1 $Y \in \mathbb{R}$, $X = (X_1, X_2) \in [0,1]^2 =: D$, te X_1 i X_2 kvantitativne.



$= \{X_1 > t_3, X_2 < t_4\}$

1. D dijelimo u $X_1 = t_1$
2. - $\{X_1 \le t_1\}$ — " — $X_2 = t_2$
   - $\{X_1 > t_1\}$ — " — $X_1 = t_3$
3. $\{X_1 > t_3\}$ — " — $X_2 = t_4$

↳ particija $R_1, \dots, R_5$ od D

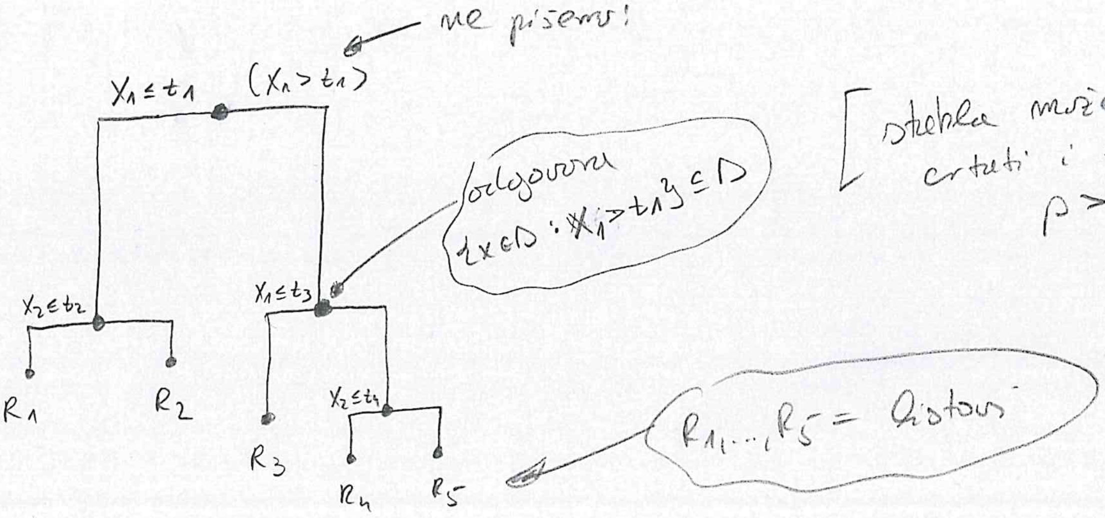
↳ za danu particiju, $\forall x \in [0,1]^2$,

$$\hat{f}(x) := \sum_{m=1}^{5} c_m \mathbb{1}_{\{x \in R_m\}} = c_k \text{ ako } x \in R_k,$$

za neke $c_1, \dots, c_5 \in \mathbb{R}$.

- ako $L(y, \hat{y}) = (y - \hat{y})^2 \longrightarrow \boxed{c_m := \frac{1}{\#\{x^{(i)} \in R_m\}} \cdot \sum_{x^{(i)} \in R_m} y_i}$

↳ ekvivalentan zapis preko **binarnog stabla**:



[stabla možemo crtati i za $p > 3$!]

↳ [Prednosti:

- interpretacija [npr. primjene u medicini]
- automatski odabir kovarijata i interakcije među njima
- invarijantnost na monotone transf. kovarijata

Mane:

- prediktivna sposobnost → (bagging, slučajne šume, boosting)

]

[Konstrukcija stabala u CART algoritmu?]

8.1.1 Regresijska stabla

- $Y \in \mathbb{R}$, te pretp. $X \in D \subseteq \mathbb{R}^p$ (sve X_j kont.)

Za svako binarno stablo T, svaki čvor $t \in T$ identificiramo s odgovarajućim područjem od D, te

$$\tilde{T} := \{ t \in T : t \text{ je list} \}.$$

↳ $$I(t) := \{ (x^{(i)}, \overset{\in T}{y_i}) : x^{(i)} \in t \} \qquad (8.1)$$

- $n(t) := |I(t)|$

- $$\bar{y}(t) := \frac{1}{n(t)} \sum_{x^{(i)} \in t} y_i, \qquad (8.2)$$

$$\hat{f}_T(x) := \sum_{t \in \tilde{T}} \bar{y}(t) \mathbb{1}_{\{x \in t\}}, \quad x \in D$$

[dakle, pretpostavili smo $L(y, \hat{y}) = (y - \hat{y})^2$]

Neka je

$$\overset{\text{"cost"}}{C(T)} := \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_T(x^{(i)}))^2 \left(= L_{\mathcal{I}}(\hat{f}_T)\right) \qquad (8.3)$$

$$= \boxed{\frac{1}{n} \sum_{t \in \tilde{T}} \underbrace{\sum_{x^{(i)} \in t} (y_i - \bar{y}(t))^2}_{=: RSS(t)}}$$

Kriterij dijeljenja: u svakom listu $t \in \tilde{T}$, neka su

$$\left.\begin{aligned} t_L = t_L(j,s) &:= \{x \in t : x_j \leq s\} \\ t_D = t_D(j,s) &:= \{x \in t : x_j > s\} \end{aligned}\right\}, \quad j \in \{1, \dots, p\}, \ s \in \mathbb{R},$$

sve moguće podjele, te $T(j,s)$ rezultirajuće stablo.

[za s dovoljno gledati samo $n(t)$ vrijednosti!]

$\hookrightarrow$ tražimo

$$(j^*, s^*) := \underset{j,s}{\operatorname{arg\,max}} \{\underbrace{C(T) - C(T(j,s))}_{=: \Delta C(t,j,s)}\} \qquad (8.4)$$

Nap. 8.2

(a) iz (8.3) odmah slijedi

$$\Delta C(t,j,s) = \frac{1}{n} (RSS(t) - RSS(t_L) - RSS(t_D)) \qquad (8.5)$$

te

$$(j^*, s^*) = \underset{j,s}{\operatorname{arg\,min}} \{RSS(t_L) + RSS(t_D)\} \qquad (8.6)$$

ovisi samo o $\mathcal{I}(t)$!

$$\frac{1}{n(t_L)} RSS(t_L) = \widehat{Var}(y_i : x^{(i)} \in t_L)$$

(b) Vrijedi $\Delta C(t, j, s) \geq 0$, $\forall t, j, s$ te (DZ)

$$\Delta C(t, j, s) = 0 \iff \bar{y}(t) = \bar{y}(t_L) = \bar{y}(t_D).$$

Kriterij zaustavljanja?

[broj listova $|\tilde{T}|$ kontrolira kompleksnost!]

Ne dijeli $t \in \tilde{T}$ ako

(1.) $\boxed{\Delta C(t, j^*, s^*) \leq \delta}$ za neki $\delta \geq 0$, [mindev]

[„previše "kratkovidno" za velike δ]

ili

(2.) $\boxed{\min\{n(t_L^*), n(t_D^*)\} < g_1}$ [mincut]

za $g_1 \in \mathbb{N}$

ili

(3.) $\boxed{n(t) < g_2}$ [minsize]

za $g_2 \in \mathbb{N}$

Kako odrediti δ, g_1, g_2?

Alternativa

Obrezivanje stabla (engl. "pruning")

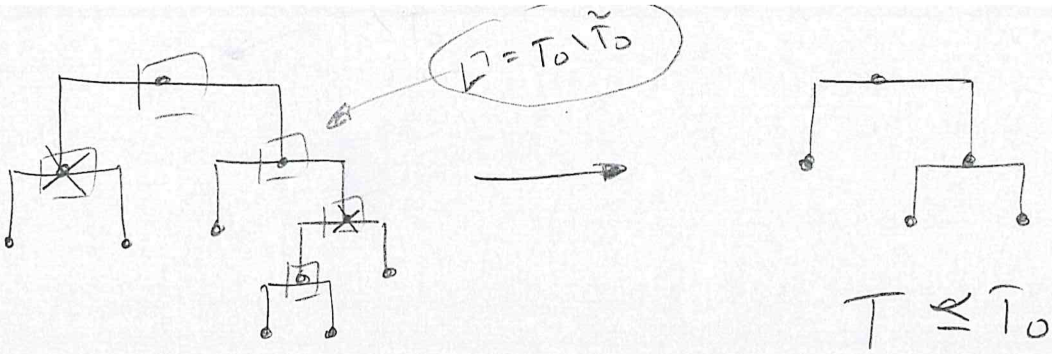
- napravimo "veliko" stablo (npr. $\boxed{\delta = 0, g_1 = 1, g = 5}$ gore)

$$\longrightarrow T_0$$

- $T \subseteq T_0$ je podstablo od T_0 ("$T \preceq T_0$")

ako je dobiveno "obrezivanjem" bilo kojeg broja čvorova $t \in T_0 \setminus \tilde{T}_0$.

npr.



T_0 $\qquad\qquad$ $T \preceq T_0$

[spajamo neke elemente particije (unutrnje)].

• za sve $\alpha \geq 0$,

$$C_\alpha(T) := C(T) + \alpha \cdot |\tilde{T}| \qquad (8.7)$$

penalizacija kompleksnosti

te

$$T(\alpha) := \underset{T \preceq T_0}{\operatorname{argmin}}\, C_\alpha(T). \qquad (8.8)$$

Nap. 8.3

(a) $\alpha = 0 \Rightarrow T(0) = T_0$

veći $\alpha \Rightarrow$ manji $|\tilde{T}(\alpha)|$

(b) Može se pokazati da postoji $k \in \mathbb{N}$ te

$$0 = \alpha_0 < \alpha_1 < \dots < \alpha_k < \alpha_{k+1} = +\infty,$$

i niz stabala $T^0 \succeq T^1 \succeq \dots \succeq T^k$ t.d.

$$\boxed{T(\alpha) = T^i \text{ za } \alpha \in [\alpha_i, \alpha_{i+1}\rangle,\ i = 0, \dots, k} \qquad (8.9)$$

→ (vidi Ripley (1996))

$$(T^0 = T_0,\ |\tilde{T}^k| = 1)$$

(c) α (tj. i) odabiremo CV metodom ili na testnom skupu.

[za CV $\to$ $\forall I_k, k=1,\dots,K$, ponovno računamo

$$T_0 = T_0(I_k), \quad T(\alpha) = T(\alpha, I_k), \ \alpha \geq 0, \quad (!)$$

te grešku procjenjujemo na $I \setminus I_k$.]

$\to$ California Housing.R

8.1.2 Klasifikacijska stabla

- $Y \in \{0, 1, \dots, K-1\} =: S$

Za stablo T, i sve $t \in T$,

$$m_k(t) := |\{x^{(i)} \in t : \boxed{y_i = k}\}|, \quad (8.10)$$

te

$$\hat{\mathbb{P}}(Y=k \mid X \in t) = p_k(t) := \boxed{\frac{m_k(t)}{m(t)}}. \quad (8.11)$$

$\hookrightarrow$

$$\hat{f}_T(x) := \sum_{t \in \tilde{T}} \overset{\in S}{k(t)} \mathbb{1}_{\{x \in t\}}, \quad x \in D,$$

te ako je gubitak $0-1$, stavimo

$$k(t) = \operatorname*{argmax}_{k \in S} \hat{p}_k(t) \quad . \quad (8.12)$$

$$\underset{(8.11)}{=} \text{—"—} \ m_k(t)$$

Kako dijelimo čvorove?

Uočimo, za regresijska stabla,

$$C(T) = \frac{1}{n} \sum_{t \in \tilde{T}} \sum_{x^{(i)} \in t} (y_i - \bar{y}(t))^2$$

$$= \boxed{\sum_{t \in \tilde{T}} \hat{p}(t) \cdot i(t)} \qquad (8.13)$$

za

- $\hat{p}(t) := \frac{n(t)}{n} \quad \left(= \hat{P}(X \in t) \right) \qquad (8.14)$
- $i(t) := \frac{1}{n(t)} \sum_{x^{(i)} \in t} (y_i - \bar{y}(t))^2 = \boxed{\widehat{Var}(y_i : x^{(i)} \in t)}$

mjera $\boxed{\text{"nečistoće"}}$ (engl. "impurity") čvora t

Za konstrukciju stabla u klasifikaciji koristimo

1. Ginijev indeks:

$$i(t) = \sum_{\substack{k \neq k', \\ k,k' \in S}} \hat{p}_k(t) \hat{p}_{k'}(t) = \sum_{k \in S} \hat{p}_k(t)(1 - \hat{p}_k(t)) \qquad (8.15)$$

ili

2. Entropiju:

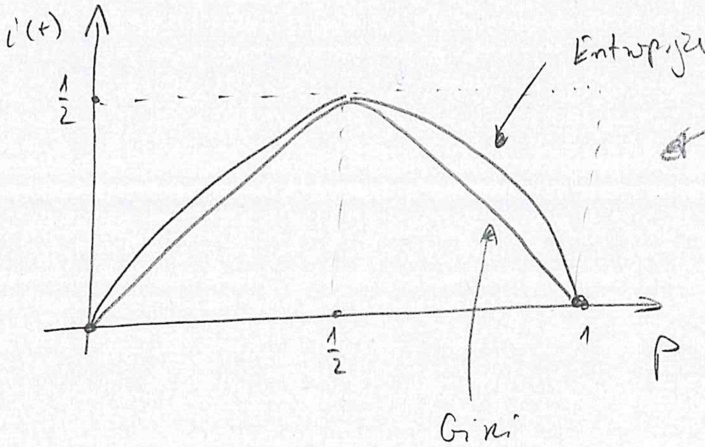
$$i(t) = - \sum_{k \in S} \hat{p}_k(t) \log(\hat{p}_k(t)) \qquad (8.16)$$

↳ vrijedi
- $i(t)$ maksimalna ako $\boxed{(\hat{p}_0(t), \ldots, \hat{p}_{K-1}(t)) = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)}$
- $i(t)$ minimalna ako $\boxed{\hat{p}_k(t) = 1}$ za neki $k \in S$.

Pr. 8.4 | $S = \{0, 1\} \longrightarrow \boxed{p := \hat{p}_1(t), \quad \hat{p}_0(t) = 1 - p}$

$\Rightarrow$ Gini $\longrightarrow$ $i(t) = 2p(1-p)$

Entropija $\longrightarrow$ $i(t) = -p\log(p) - (1-p)\log(1-p)$



otvorene konkavne!

Kriterij dijeljenja: za $t \in \tilde{T}$, tražimo

$$(j^*, s^*) = \underset{j,s}{\operatorname{argmax}} \{ \Delta C(t, j, s) \}$$

$$(8.17) \quad = \text{—}\,\text{''}\,\text{—} \{ \hat{p}(t)i(t) - \hat{p}(t_L)i(t_L) - \hat{p}(t_D)i(t_D) \}$$

$$= \underset{j,s}{\operatorname{argmin}} \{ \hat{p}(t_L)i(t_L) + \hat{p}(t_D)i(t_D) \}$$

Nap. 8.5 | U regresiji imamo $L(y, \hat{y}) = (y - \hat{y})^2$ te koristimo

$$i(t) = \frac{1}{n(t)} \sum_{x^{(i)} \in t} L(y_i, \underbrace{\hat{h}_T(x^{(i)})}_{= \bar{y}(t)}),$$

tj. $\boxed{C(T) = L_I(\hat{h}_T)}$.

$\longrightarrow$ zašto u klasifikaciji ne koristimo 0-1 gubitak

$$i(t) = \frac{1}{n(t)} \sum_{x^{(i)} \in t} \mathbb{1}\{ y_i \neq \underbrace{\hat{h}_T(x^{(i)})}_{= h(t)} \}$$

$$\underset{(8.11)}{=} 1 - \hat{p}_{k(t)}(t) \qquad (8.18)$$

npr. $S = 10, 10,$

$$t = (4, 10)$$

$\uparrow n_0(t)$ $\uparrow n_1(t)$ $\to n(t) = 20$

te gledamo podjelu

$$t_L = (4, 0)$$

$$\boxed{t_D = (0, 10)} \leftarrow \text{"čist čvor"}$$

$$\Delta C = \hat{p}(t)\,i(t) - \hat{p}(t_L)\,i(t_L) - \hat{p}(t_D)\,i(t_D)$$

$$= \frac{1}{n}\begin{cases} \boxed{0, \quad 0-1} \\ 1.6, \quad \text{Gini} \\ 3.27, \quad \text{entropija} \end{cases}$$

DZ

jer $k(t_L) = k(t_D) = k(t)$!
$(=1)$

[problem je u tome što $0-1$ f-ja nečistoće nije strogo konveksna!]

Npr, $0-1$ mjeru koristimo pri obrezivanju

[jer nas zanima predikcija]

∎

~~tpoh, 0-1 mjere kvalitete pri obrezivanju stabla.~~

[~~zanima nas~~ predikciju]

8.1.3 Generalizacije

Kategorijalne kovarijate

Ako je $X_j \in \{0,1,\dots,K_j-1\} =: S_j$, imamo $2^{K_j-1}-1$ mogućih podjela S_j na dvije grupe

(npr. $t_L = \{x \in t : x_j \in \{0,1\}\}$, $t_R = \{$ — '' — $x_j \in S_j \setminus \{0,1\}\}$)

$\Rightarrow$ za velike K_j imamo pristranost za izbor X_j za podjelu!

Funkcija gubitka ↳ pretp. $\boxed{S = \{0,1\}}$

Umjesto samo $L(\hat{A}) = \mathbb{P}(Y \neq \hat{A}(X))_{\mathbb{P}_{X,Y}}$, često nas više zanimaju

- $\mathbb{P}(Y = \hat{A}(X) \mid Y = 1) \rightsquigarrow$ "sensitivity"
- — '' — $\mid Y = 0) \rightsquigarrow$ "specificity"

npr. (a) 0 = nije bolest, 1 = bolest

$\Rightarrow$ želimo što veću sens.

(b) 0 = nije spam, 1 = spam

$\Rightarrow$ — '' — specif.

↳ Matrica gubitka (f-ja) je $L = (L_{k,k'} : k,k' \in \{0,1\})$, gdje je

9.1

- $L_{k,k} = 0$, $\forall k$

- $L_{k,k'}$ = trošak ako stvarnu vrijednost k (≥0) predviđamo s k', $k \neq k'$

Za stablo T, ako je $x \in t$ ($t \in \tilde{T}$), i imamo

$$\hat{f}_T(x) = k(t) \overset{\text{kao u (7.2)}}{:=} \underset{k \in \{0,1\}}{\operatorname{argmin}} \underbrace{\sum_{\ell \in \{0,1\}} L_{\ell,k} \overbrace{\hat{p}_\ell(t)}^{=\hat{\mathbb{P}}(Y=\ell \mid X \in t)}}_{=\hat{\mathbb{E}}[L_{Y,k} \mid X \in t]}$$

$$\overset{L_{k,k}=0}{=} \underset{k \in \{0,1\}}{\operatorname{argmin}} L_{1-k,k} \, \hat{p}_{1-k}(t),$$

tj.

$$\hat{f}_T(x) = 1 \iff L_{0,1} \hat{p}_0(t) \leq L_{1,0} \hat{p}_1(t)$$

$$\underset{\hat{p}_0 = 1-\hat{p}_1}{\iff} \boxed{\hat{p}_1(t) \geq \frac{L_{0,1}}{L_{1,0} + L_{0,1}} =: \alpha}$$

↳ • $L_{0,1} > L_{1,0}$ (npr. 0 = nije spam, 1 = spam)
$\Rightarrow \alpha > \frac{1}{2}$ te veći $\boxed{\text{specificity}}$.

[• $L_{0,1} < L_{1,0}$ (npr. 0 = nije bolest, 1 = bolest)
$\Rightarrow \alpha < \frac{1}{2}$ te veći $\boxed{\text{sensitivity}}$]

• $L_{0,1} = L_{1,0}$ (0-1 gubitak)
$\Rightarrow \alpha = \frac{1}{2}$.

- uvijek postoji "tradeoff" između sensit. i specif.

→ $\boxed{\text{ROC krivulja}}$ 92

- umjesto kod predikcija, L se može uzeti u obzir i pri konstrukciji stabla

→ Carseats.R