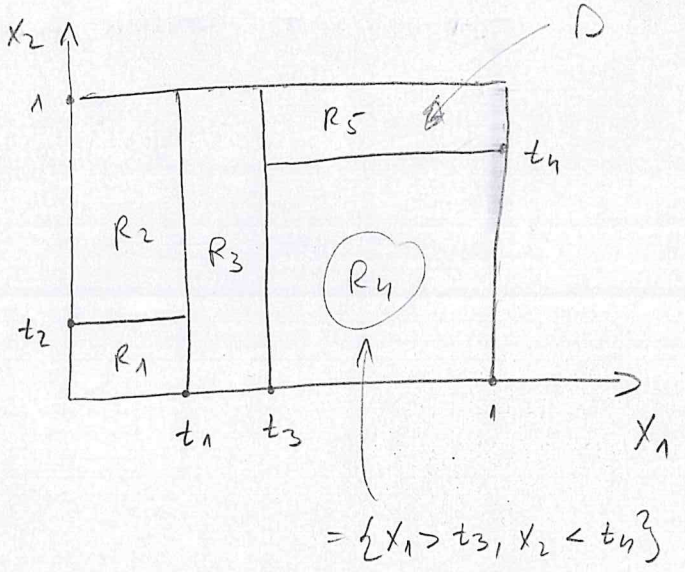
8. Metode bazirane na stablima

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



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, $\forall x \in [0,1]^2$,

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

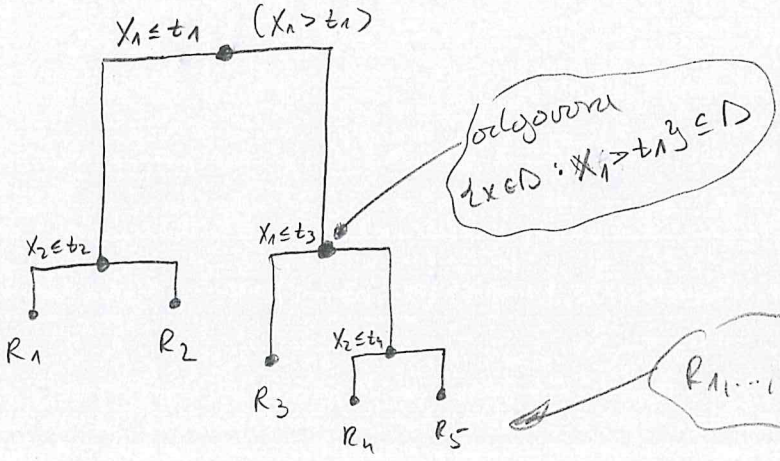
ze neke $c_1, \dots, c_5 \in \mathbb{R}$.

- ako $L(y, \hat{y}) = (y - \hat{y})^2 \longrightarrow$

$$c_m := \frac{1}{\#\{x^{(i)} \in R_m\}} \cdot \sum_{x^{(i)} \in R_m} y_i$$

↳ ekvivalentan zapis preko binarnog stabla:

ne pišemo!



[stabla možemo crtati i za $p > 3$!]

↳ Predmeti:

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

Mane:

- predikcijska sposobnost → bagging, slučajne šume, boosting

[Konstrukcija stabala u CART algoritmu?]

2.1.1 Regresijska stabla

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

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

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

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

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

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

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

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

Neka je

"cost"

$$C(T) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_T(x^{(i)}))^2 \quad (= L_I(\hat{f}_T)) \qquad (8.3)$$

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

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

$$t_L = t_L(j, s) := \{x \in t : x_j \leq s\}$$

$$t_D = t_D(j, s) := \{x \in t : x_j > s\}$$

, $j \in \{1, \dots, p\}$, $s \in \mathbb{R}$,

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

↳ tražimo [za s dovoljno gledati samo $n(t)$ unpechasti!]

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

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

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

oni samo u $\tilde{T}(t)$!

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

(iv) Vrijedi $\Delta C(t_{1,j,0}) \geq 0$, $\forall t_{i,j,s}$ te (vrijedi)

$$\Delta C(t_{1,j,0}) = 0 \iff \bar{y}(t) = \bar{y}(t_L) = \bar{y}(t_D).$$

Kriteriji zaustavljanja?

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

↳ Ne elji $t \in \tilde{T}$ ako

(1.) $\Delta C(t_{i,j^*,s^*}) \leq \delta$ za neki $\delta \geq 0$, [mindec]

[...previše "kratkosežno" za velike δ]

ili

(2.) $\min\{n(t_L^*), n(t_D^*)\} < q_1$ [mincut]

za $q_1 \in \mathbb{N}$

ili

(3.) $n(t) < q_2$ [minsize]

za $q_2 \in \mathbb{N}$

↳ Kako odrediti δ, q_1, q_2?

Alternativa

Obrezivanje stabla (engl. "pruning")

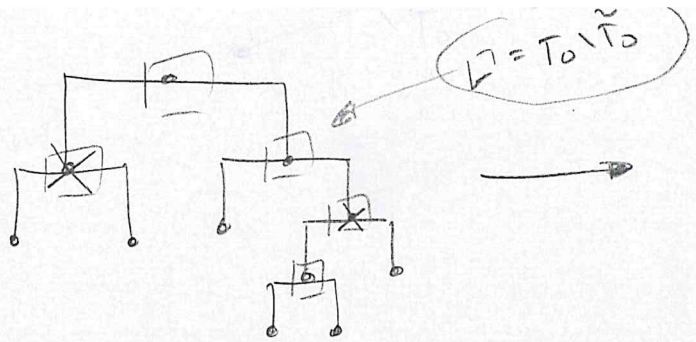
- napravimo "veliko" stablo (npr. $\delta = 0, q_1 = 1, q = 5$ gore)

$\longrightarrow T_0$

- $T \subseteq T_0$ 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.



$T \preceq T_0$

T_0

[spajamo neke elemente particije (unutrag)].

- 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 postoje $k \in \mathbb{N}$ te

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

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

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

↳ (vidi Ripley (1996))

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

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

[za CV $\rightarrow$ $\forall T_k$, $k=1,\dots,K$, ponovno računamo

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

te grešku procjenjujemo na $T \setminus T_k$.]

$\rightarrow$ California Housing. R

8.1.2 Klasifikacijska stabla

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

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

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

te

$$\hat{P}(Y=k \mid X \in t) = p_k(t) := \frac{n_k(t)}{n(t)} . \quad (8.11)$$

$\hookrightarrow$

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

te ako je gubitak 0-1, otprilike

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

$$= \underset{k \in S}{\operatorname{argmax}}\, n_k(t)$$

(8.11)

Kako dijelimo čvorove?

Uočimo, za regresijska stabla,

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

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

za

- $\hat{p}(t) := \frac{n(t)}{n} \quad (= \hat{P}(X \in t))$ (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 "nečistoće" (engl. "impurity") čvora t

Za konstrukciju stabla u klasifikaciji koristimo

(1.) Ginijev indeks:

$$i(t) = \sum_{\substack{h \neq h', \\ h,h' \in s}} \hat{p}_h(t) \hat{p}_{h'}(t) = \sum_{h \in s} \hat{p}_h(t) (1 - \hat{p}_h(t)) \qquad (8.15)$$

ili

(2.) Entropiju:

$$i(t) = - \sum_{h \in s} \hat{p}_h(t) \log (\hat{p}_h(t)) \qquad (8.16)$$

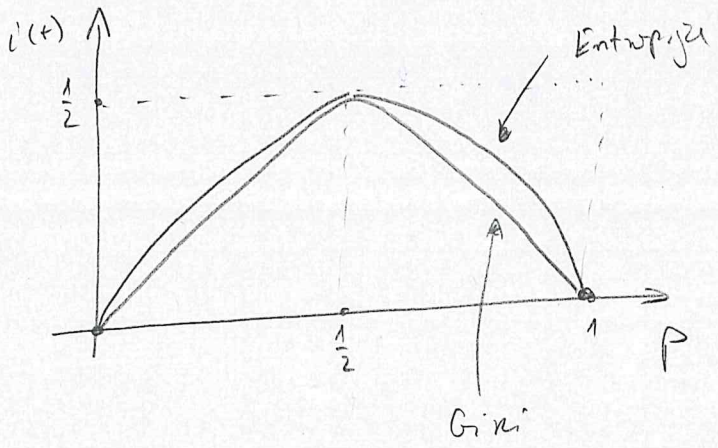
↳ uvjeti • $i(t)$ maksimalna ako $\boxed{(\hat{p}_0(t), \dots, \hat{p}_{K-1}(t)) = (\frac{1}{K}, \dots, \frac{1}{K})}$

• $i(t)$ minimalna ako $\boxed{\hat{p}_h(t) = 1}$ za neki $h \in s$.

Pr. 8.4 | $S = \{0, 1\}$ $\longrightarrow$ $p := p_1(t)$, $p_0(t) = 1 - p$

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

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



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

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

$$(8.17) \quad = \text{"} - \underset{j,s}{\operatorname{argmin}} \{ \hat{p}(t) i(t) - \hat{p}(t_L) i(t_L) - \hat{p}(t_D) i(t_D) \}$$

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

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

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

tj. $\boxed{C(T) = L_T(\hat{f}_T)}$.

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

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

$$= \boxed{1 - \hat{p}_k(t)} \left(= \hat{P}\left(Y \neq \hat{f}_T(x) \mid x \in t\right)\right) \quad (8.19)$$

(8.11)

?

npr. $S = \{0,1\}$, $\boxed{t = (400, 400)}$ $\rightarrow n(t) = 800$

$n_0(t)$ $n_1(t)$

↳ podjele (t_L, t_D):

1° $t_L = (300, 100)$
$t_D = (100, 300)$

2° $t_L = (200, 400)$
$\boxed{t_D = (200, 0)}$

↳ u 1° imamo $\hat{p}(t_L) = \hat{p}(t_D) = \frac{400}{n}$, te

$$\hat{p}(t_L)\, i(t_L) + \hat{p}(t_D)\, i(t_D) \underset{Dt}{=} \frac{1}{n} \begin{cases} 300, & \text{Gini} \\ 449.9, & \text{Entropija} \\ 200, & 0-1 \end{cases}$$

u 2° imamo $\hat{p}(t_L) = \frac{600}{n}$, $\hat{p}(t_D) = \frac{200}{n}$ te

$$\text{---------- } // \text{ ----------} \underset{Dt}{=} \frac{1}{n} \begin{cases} 266.7, & \text{Gini} \\ 381.9, & \text{entropija} \\ 200, & 0-1 \end{cases}$$

$\Rightarrow$ $0-1$ mjera ne preferira 2° iako daje

(8.17) "čist" čvor [koji ne treba dalje dijeliti]

Ipak, 0-1 mjera kvantitativno pri obrezivanju stabla.

[zanima nas predikcija]

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. $t_L = \{x \in t : x_j \in \{0,1,3\}\}$, $t_D = \{$ — " — $x_j \in S_j \setminus \{0,1,3\}\}$

$\longrightarrow$ za velike K_j imamo prostranost za izbor X_j za podjelu!

Funkcija gubitka

$\hookrightarrow$ pretp. $S = \{0,1\}$

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

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

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

$\Rightarrow$ želimo što veću sens.

(b) 0 = nije spam, 1 = spam

$\rightarrow$ — " — specif.

↳ Matrica gubitka je $L = (L_{k,h'} : h,h' \in \{0,1\})$, gdje je (t-ja)

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

- $L_{k,k'}$ = trošak ako stvarna mjerka (h) predviđamo da je (h'), $(k \neq h')$

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

imamo

$$\hat{h}_T(x) = h(t) := \underset{k \in \{0,1\}}{\operatorname{argmin}} \underbrace{\sum_{\ell \in \{0,1\}} L_{\ell,k}\, \hat{p}_\ell(t)}_{= \hat{E}[L_{Y,k} \mid X \in t]} \quad \text{(nav u (7.2))}$$

$$\underbrace{\hat{p}_\ell(t)}_{= \hat{P}(Y=\ell \mid X \in t)}$$

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

tj.

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

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

↳ - $L_{0,1} > L_{1,0}$ (npr. 0 = nije spam, 1 = spam)

$\Rightarrow \alpha > \frac{1}{2}$ te vedi specificity.

- $L_{0,1} < L_{1,0}$ (npr. 0 = nije bolest, 1 = bolest)

$\Rightarrow \alpha < \frac{1}{2}$ te vedi sensitivity

- $L_{0,1} = L_{1,0}$ (0-1 gubitak)

$\Rightarrow \alpha = \frac{1}{2}$.

- uvijek postoji "trade off" između senzit. i specif.

→ ROC krivulja

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