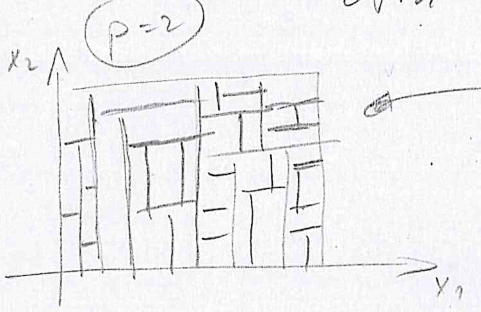
8.2 Bagging i slučajne šume [random forests] → Breiman (1996, 2001)

- Stabla odlučivanja → interpretabilni modeli, ali loša prediktivna sposobnost

- npr., velika stabla → mala pristranost, ali velika varijanca jer jako ovisi o (T)



← prilagođavamo konstantu ∀ element particije.

↳ ideja je uprosječiti predikciju velikog broja različitih stabala dobivenih iz bootstrap uzoraka, te tako smanjiti varijancu! Ipak, tako direktno gubimo interpretab.!

8.2.1 Bootstrap uzorci

- imamo

$$I = \{ \underbrace{(x^{(i)}, y_i)}_{=: z_i} : i = 1, \dots, n \},$$ [malo z_i]

koji je realizacija mjd uzorka

$$T = \{ (X^{(i)}, Y_i) : i = 1, \dots, n \}$$

iz $\mathbb{P}((X, Y) \in \cdot)$

- za dani (I), jedan bootstrap uzorak (duljine n), je

$$T_{boot} = \{ z_1^*, \dots, z_n^* \},$$

pri čemu su $z_1^*, \dots, z_n^*$ nezavisni t.d.

$$\mathbb{P}(z_1^* = z_i) = \frac{1}{n}, \quad \forall i = 1, \dots, n.$$

↳ tj., T_{boot} je nezavisan uzorak iz empirijske razdiobe $\hat{F}_n$ uzorka T

(1.) Budući da je $\hat{F}_n \approx \mathbb{P}((X,Y) \in \cdot)$ za velike n, T_{boot} je uzorak koji je približno iz $\mathbb{P}((X,Y) \in \cdot)$ [tj. stvarne razdiobe!]

(2.) Neki $z_i \in T$ će se u T_{boot} ponoviti više puta, a neki niti jednom

$\Rightarrow$ T_{boot} "drugačiji" od T

npr. $\forall z_i \in T$,

$$\mathbb{P}(z_i \notin T_{boot}) = \underbrace{\left(1 - \frac{1}{n}\right)^n}_{\mathbb{P}(z_1^* \neq z_i)} \xrightarrow{n \to \infty} e^{-1} \approx 0.368,$$

te

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{z_i \notin T_{boot}\}}\right] \xrightarrow{n \to \infty} 0.368 \qquad (3.19)$$

$\Rightarrow$ za velike n, T_{boot} u prosjeku izostavlja 36.8% početnih iz T!

Napl Bootstrap metoda se najviše koristi za konstrukciju pouzdanih intervala za prigenitelje $S = S(z_1, \dots, z_n)$ čiju je distr. preteško odrediti analitički.

9.2.2 Bagging - (Bootstrap aggregation)

1. Napravimo B (B "velik") bootstrap uzoraka

$$T_{boot}^{(1)}, \dots, T_{boot}^{(B)}$$

2. $\forall b = 1, \dots, B$, konstruiramo (tipično velika) stablo $T^{(b)}$ na temelju $T_{boot}^{(b)}$ [bez obrezivanja]

$$\Rightarrow \boxed{\hat{f}_b := \hat{f}_{T^{(b)}}}$$

3. Bagging prosjenitelj je

i) ako $(Y \in \mathbb{R})$,

$$\hat{f}_{bag}(x) := \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x), \quad x \in \mathbb{R}^p. \qquad (8.20)$$

ii) ako $(Y \in \{0, 1, \dots, K-1\} =: S)$,

$$\hat{f}_{bag}(x) := \underset{k \in S}{\operatorname{argmax}} \left| \{ b \in \{1, \dots, B\} : \hat{f}_b(x) = k \} \right| \qquad (8.21)$$

["majority vote"].

ili

$$\hat{P}(Y = k \mid X = x)$$

$$\hat{p}_k^{bag}(x) := \frac{1}{B} \sum_{b=1}^{B} \hat{p}_k^{(b)}(x), \quad x \in \mathbb{R}^p \qquad (8.22)$$

te

$$\hat{f}_{bag}(x) := \underset{k \in S}{\operatorname{argmax}}\ \hat{p}_k^{bag}(x), \quad x \in \mathbb{R}^p \qquad (8.23)$$

[Ako $L = 0-1$]

[$\hat{f}_{bag}(x)$ je tzv. plug-in prosjenitelj!]

Nap. 8.6 iz (8.20) i JZVB-a slijedi

$$\hat{f}_{bag}(x) \xrightarrow{B \to \infty} \mathbb{E}_{T_{boot}^{(1)}} \left[\hat{f}_{T^{(1)}}(x) \right] \qquad (8.24)$$

$\Rightarrow$ već B ne znači kompleksniji model! ☺

Zašto bagging pomaže? [heuristički]

- pretp. $Y \in \mathbb{R}$, te $\hat{f} := \hat{f}_{T_0}$ za veliko stablo $T_0 = T_0(T)$ (slučajno!)

↳
$$\mathbb{E}_T[L_x(\hat{f})] = \sigma^2 + Var_T(\hat{f}(x)) + Bias_T[\hat{f}(x)]$$
$$\underset{\text{pretp.}}{\approx} \sigma^2 + \underbrace{Var_T(\hat{f}(x))}_{=: \sigma^2_{T_0}}$$

- $\hat{f}_1(x), \dots, \hat{f}_B(x)$ su jednako distrib. (ali zavisne!) [jer ovise o T!]

te pretp. da je

$$Var(\hat{f}_b(x)) \approx \sigma^2_{T_0}, \quad Bias(\hat{f}_b(x)) \approx 0, \qquad (8.25)$$

[slučajnost još od $T^{(1)}_{boot}, \dots, T^{(B)}_{boot}$!]

te neka je

$$\rho(x) := \rho(\hat{f}_b(x), \hat{f}_{b'}(x)), \quad b \neq b' \qquad (8.26)$$

[korelaciju]

↳ (1.) $\mathbb{E}[\hat{f}_{bag}(x)] = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}[\hat{f}_b(x)]$
$$= \mathbb{E}[\hat{f}_1(x)],$$

tj. $Bias(\hat{f}_{bag}(x)) = Bias(\hat{f}_1(x)) \overset{(8.25)}{\approx} 0.$

(2.) $Var(\hat{f}_{bag}(x)) \overset{DZ}{=} \left(\rho(x) + \frac{1-\rho(x)}{B}\right) \cdot Var(\hat{f}_1(x))$

$$\underset{(8.25)}{\approx} \text{———} \;\; '' \;\; \text{———} \; \sigma^2_{T_0}$$

$\rho(x) \cdot \sigma^2_{T_0}$ [$\Rightarrow \rho(x) \geq 0$] (8.27)

za veliki B

$< \sigma^2_{T_0}$!!!

$\rho(x) < 1$

8.2.3 Slučajne šume (random forests) → (RF)

↳ ideja je smanjiti $\rho(x)$ u (8.25)

- u odnosu na bagging, jedina promjena je u konstrukciji stabla $T^{(b)}$ iz $T^{(b)}_{boot}$, $\forall b = 1, \dots, B$:

> Prije svakog dijeljenja čvora, izaberi slučajno (m) kovarijata $(m \leq p)$ kao kandidate za dijeljenje.

- intuitivno, manji (m) ⇒ manja $\rho(x)$

[stabla primorana koristiti različite kovarijate za dijeljenje]

(ipak, mijenjaju se i $Var(f_b(x))$, $Bias(f_b(x))$)

→ [vidi ESL, Pogl. 15.4]

→ (m) je najvažniji hiperparametar. ("mtry")

[RF često rade odlično, a jednostavne su za prilagodbu ("out-the-shelf")]

Out-of-Bag ("OOB") procjena greške

- $\forall b \in \{1, \dots, B\}$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{z_i \notin T_{boot}^{(b)}\}\right] \overset{(8.19)}{\approx} 0{,}308$$

"OOB uzorak" $\Rightarrow$ nisu korišteni za konstr. $\hat{f}_b$!

↓

- $\forall i = 1, \dots, n$

$$OOB_i := \{b \in \{1, \dots, B\} : z_i \notin T_{boot}^{(b)}\}, \quad (8.28)$$

- ako je $(y \in \mathbb{R})$, definiramo

$$\hat{y}_i^{OOB} := \frac{1}{|OOB_i|} \sum_{b \in OOB_i} \hat{f}_b(x^{(i)}), \quad i = 1, \dots, n$$

te OOB procjenu greške

$$OOB(\hat{f}_{rf}) := \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i^{OOB})^2 \quad (8.29)$$

↳ slične procjene kao i LOOCV, ali dobivamo ih besplatno, tj. u sklopu treniranja!

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

$$\hat{y}_i^{OOB} := \underset{h \in S}{\operatorname{argmax}} \left| \{b \in \{1, \dots, B\} \cap OOB_i : \hat{f}_b(x^{(i)}) = h\} \right|,$$

te $$OOB(\hat{f}_{rf}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \neq \hat{y}_i^{OOB}\}. \quad (9.30)$$

Utjecaj kovarijeta (engl. variable importance) ($\eta \in \mathbb{R}$)

[· $\hat{t}_{rt}$ (i specijalno, $\hat{t}_{bag}$) više nemaju strukturu stabla pa se gubi njihova sposobnost interpretacije!]

- za jedno stablo T, neka je $\forall j = 1, \dots, p$,

$$I_j(T) := \sum_{t \in T \setminus \tilde{T}:\ j^*(t) = j} \Delta C(t, j^*(t), s^*(t)) \qquad (8.31)$$

(I_j → "importance")

uz

- $X_{j^*(t)}$ je kovarijeta korištena za dijeljenje čvora t,
- a $\Delta C(t, j^*(t), s^*(t))$ je smanjenje "troška" C (vidi (8.4)) pri dijeljenju čvora t.

- za RF definiramo

$$I_j := \frac{1}{B} \sum_{b=1}^{B} I_j(T^{(b)}). \qquad (8.32)$$

Gradovi djelomične zavisnosti (partial dependence plots)

↳ (PDP)

- pretp. ($Y \in \mathbb{R}$)
- za $A, B \subseteq \{1, \dots, p\}$ t.d. $A \cap B = \emptyset$, $A \cup B = \{1, \dots, p\}$, $\forall x \in \mathbb{R}^p$

  pišemo $x_A := (x_j)_{j \in A}$, $x_B := (x_j)_{j \in B}$
- za $f := \hat{f}_{rf} : \mathbb{R}^p \to \mathbb{R}$, projekcijsku ili djelomičnu ovisnost f-je f na X_A definiramo kao

$$f_A(x_A) := \mathbb{E}\left[f(x_A, X_B) \right], \quad x_A \in \mathbb{R}^{|A|} \qquad (8.33)$$

(x_A: brojčev, $|A| = 1$ ili 2; X_B: slučajno)

[f_A predstavlja utjecaj kovarijeta X_A na $f(x)$ nakon što smo u obzir uzeli projekcijski utjecaj ostalih kovarijeta X_B na $f(x)$]

(npr.) - ako je $f(x) = h_1(x_A) + h_2(x_B)$,

$$f_A(x_A) = h_1(x_A) + \underbrace{\mathbb{E}\left[h_2(X_B) \right]}_{= \text{const.}!}$$

- ako je $f(x) = h_1(x_A) \cdot h_2(x_B)$,

$$f_A(x_A) = \text{const.} \cdot h_1(x_A).$$

- $f_A(x_A)$ procjenjujemo s

(iz 7!)

$$\hat{f}_A(x_A) := \frac{1}{n} \sum_{i=1}^{n} f(x_A, x_B^{(i)}). \qquad (8.34)$$

[za RF moguće efikasno računati] [(8.33) nije ograničen samo na RF!]

Nap. 1 Ako su X_A i X_B **nezavisne**,

$$\mathbb{E}\left[f(X_A, X_B) \mid X_A \right]$$

$$= f_A(X_A), \text{ tj.}$$

$$f_A(x_A) = \mathbb{E}\left[f(\underbrace{X_A, X_B}_{=X}) \mid X_A = x_A \right].$$

→ Boston_bag_i_rt.R

općenito, to ne možemo provjeriti jer nemamo dovoljno $(x^{(i)}, y_i) \in \mathcal{L}$ t.d. $x_A^{(i)} = x_A$ ∎