8.3 Boosting –

- ideja: kombiniranti sekvencijalno niz jednostavnih "baznih" prediktora u jedan (jak) dobar predikt., pri čemu se svaki novi bazni prediktor više fokusira na elemente iz $\mathcal{I}$ koji su lošije predviđeni u trenutnom prediktoru
- bazni prediktori mogu biti bilo što
- mi ćemo promatrati Friedmanov (2001) "Gradient Boosting Machine" (GBM) algoritam uz regresijska stabla kao bazne prediktore
- metode poput XGBoost i LightGBM su varijante originalnog GBM algoritma, a predstavljaju jedne od najmoćnijih prediktivnih metoda u strojnom učenju

8.3.1 Gradient boosting – za regresiju ("GB") ($Y \in \mathbb{R}$)

- za danu f-ju gubitka $L: \mathbb{R}^2 \to [0, +\infty)$, želimo predvidjeti

$$f^*(x) := \underset{c \in \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}[L(Y, c) \mid X = x], \quad x \in \mathbb{R}^p, \quad (8.3$$

na temelju $\mathcal{I} = \{(x^{(i)}, y_i) : i = 1, \dots, n\}$.

↳ tipično

$$\hat{f} := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} L(y_i, f(x^{(i)}))}_{=: L_{\mathcal{I}}(f)} \quad (8.36)$$

↑ neka familija

pri čemu (8.36) rješavamo nekom optimizacijskom metodom,
, time da L u (8.35) i (8.36) ne mora nužno biti ista!

Pr. 8.7 |

(a) $L(y, f(x)) = \frac{1}{2}(y - f(x))^2$ (L_2 - gubitak) (8.37)

$\Rightarrow f^*(x) = \mathbb{E}[Y \mid X = x]$

$\Rightarrow$ pogodna za teorijsku analizu

$\Rightarrow$ ipak, ima $D_I(f)$ u (8.36) značajno utjecaju $(x^{(i)}, y_i)$ t.d. je $|y_i - f(x^{(i)})|$ velike ("outlieri")

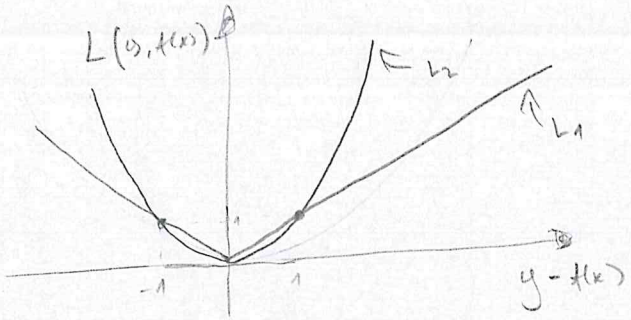
(b) $L(y, f(x)) = |y - f(x)|$ (L_1 - gub.) (8.38)

$\Rightarrow f^*(x) = \text{median}(Y \mid X = x)$

$\Rightarrow$ "robusnija" alternativa za L_2 - gub.

$\Rightarrow$ ipak, ako $y - f(x) \approx 0$,

$L_1(y, f(x)) >> L_2(y, f(x))$,



(c) $L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & \text{ako } |y - f(x)| \leq \delta, \\ \delta(|y - f(x)| - \frac{1}{2}\delta), & \text{inače} \end{cases}$ (8.39)

("Huberov" gub.)

za $\delta > 0$.

$\Rightarrow$ kombinacija L_1 i L_2 gub.: za male $y - f(x)$ se ponaša kao L_2, a za velike kao L_1 gub.

[izbor L-a ovisi o problemu i što hoćemo više penalizirati!]

- **GB metoda (za stabla):**

(i) f^* procjenjujemo modelom oblika

$$f^{(M)}(x) = f^{(0)}(x) + \sum_{m=1}^{M} b_m(x), \quad x \in \mathbb{R}^p \qquad (8.40)$$

[tj. $\mathcal{F}$ u (8.36) je familija f-ja oblika (8.40)]

gdje su $b_1, \dots, b_M$ regresijska stabla iz (8.1.1), tj.

$\forall m$, $\exists$ stablo T_m t.d.

- $|\tilde{T}_m| = J$ (tipično J mali!), te

- $b_m(x) := \sum_{t \in \tilde{T}_m} \gamma_t \mathbb{1}_{\{x \in t\}}$, uz

$$\gamma_t := \underset{\gamma \in \mathbb{R}}{\operatorname{argmin}} \sum_{x^{(i)} \in t} L(y_i, \gamma) \qquad (8.41)$$

ovisi o L!

mpr. (a) $L = L_2 \Rightarrow \gamma_t = \bar{y}(t)$ (kao u (8.1.1))

(b) $L = L_1 \Rightarrow \gamma_t = \text{medijan}(y_i : x^{(i)} \in t)$

(ii) rješenje problema (8.36) aproksimirat ćemo sekvencijalno po uzoru na metodu

gradijentnog spusta [engl. gradient descent]:

$\forall m = 1, 2, \dots, M$, b_m biramo tako da

$$b_m \approx \underset{b}{\operatorname{argmin}} L_I(f^{(m-1)} + b) \qquad (8.42)$$

$$= \text{-}\!\!\|\!\!\text{-} \sum_{i=1}^{n} L(y_i, f^{(m-1)}(x^{(i)}) + b(x^{(i)}))$$

Gradijentni spust [digresija]

- Neka je $L: \mathbb{R}^n \to \mathbb{R}$ diferencijabilna i konveksna

t-jce, a trazimo

$$\Theta^* := \underset{\Theta \in \mathbb{R}^n}{\operatorname{argmin}}\, L(\Theta).$$

- Algoritam:

(1.) odredi $\Theta^{(0)}$ [npr. $\Theta^{(0)} := 0$]

(2.) $\forall m = 1, \dots, M$,

$$\Theta^{(m)} = \Theta^{(m-1)} + \eta_m \cdot g_m$$

gdje je

(a) $g_m := -\nabla L(\Theta^{(m-1)})$ (negativni gradijent)

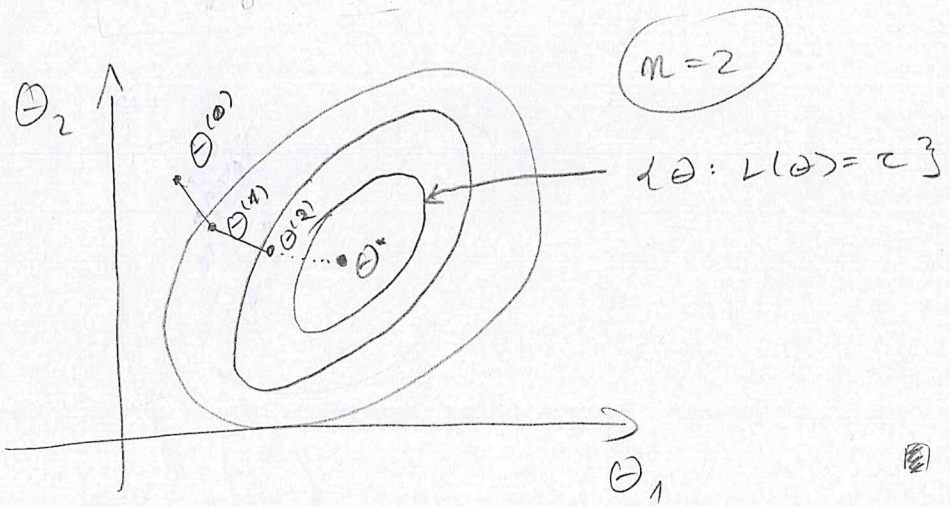
Smjer najvećeg pada (lokalno) t-je L u $\Theta^{(m-1)}$

$$= \left(\frac{-\partial L}{\partial \Theta_1}(\Theta^{(m-1)}), \dots, \frac{-\partial L}{\partial \Theta_n}(\Theta^{(m-1)}) \right) \in \mathbb{R}^n$$

(b) $\eta_m := \underset{\eta > 0}{\operatorname{argmin}}\, L(\Theta^{(m-1)} + \eta \cdot g_m)$

("line search")

(3.) Vratiti $\Theta^{(M)}$.



- Shvatimo svaku vrijednost t-ja $\boxed{t : \mathbb{R}^p \to \mathbb{R}}$ kao vektor

$\boxed{t \in \mathbb{R}^n}$ uz preslikavanje

$$t \longmapsto \left(t(x^{(1)}), \dots, t(x^{(n)}) \right) =: (t_1, \dots, t_n),$$

a L_I kao t-ju na $\mathbb{R}^n$ uz

$$L_I(t) := \sum_{i=1}^{n} L(y_i, t_i), \quad t = (t_1, \dots, t_n) \in \mathbb{R}^n. \tag{8.43}$$

↳ ključna ideja: u (8.42) b_m biramo tako da

$(b_m(x^{(1)}), \dots, b_m(x^{(n)}))$ **aproksimira**

$$-\underbrace{\nabla L_I(t^{(m-1)})}_{= g_m} = (-g_{1,m}, \dots, -g_{n,m})$$

gdje je $\forall i = 1, \dots, n$,

$$g_{i,m} := \frac{\partial L_I}{\partial t_i}(t)\Big|_{t = t^{(m-1)}}$$

$$\overset{(8.43)}{=} \frac{\partial L(y_i, t_i)}{\partial t_i}\Big|_{t_i = t^{(m-1)}(x^{(i)})},$$

pri čemu stablo T_m konstruiramo koristeći CART algoritam iz 8.1.1. na podatke

$$\{(x^{(i)}, \boxed{-g_{i,m}}) : i = 1, \dots, n\}$$

(dakle, uz $L = L_2$) → ∃ jako efikasne implem.!

- "Gradient tree boosting" algoritam

Friedman (2001)

(1.) Inicijaliziraj $f^{(0)}$. [mpr. $\equiv$ const.]

(2.) $\forall m = 1, \dots, M$,

(a) $\forall i = 1, \dots, n$, izračunaj

$$g_{im} := \left. \frac{\partial L(y_i, f_i)}{\partial f_i} \right|_{f_i = f^{(m-1)}(x^{(i)})}$$

(b) Konstruiraj regresijsko stablo $\tilde{T}_m$ (kao u (8.1.1)) za podatke

$$\{ (x^{(i)}, -g_{im}) : i = 1, \dots, n \}$$

"pseudo-reziduali"

t.d. $|\tilde{T}_m| = J$,

(c) $\forall t \in \tilde{T}_m$, izračunaj

$$\gamma_t := \underset{\gamma \in \mathbb{R}}{\operatorname{argmin}} \sum_{x^{(i)} \in t} L(y_i, \gamma),$$

"line search"

te definiraj

$$b_m(x) := \sum_{t \in \tilde{T}_m} \gamma_t \mathbb{1}_{\{x \in t\}}, \quad x \in \mathbb{R}^p$$

(d) $f^{(m)}(x) := f^{(m-1)}(x) + \lambda \cdot b_m(x), \quad x \in \mathbb{R}^p$

(3.) Vrati $f^{(M)}$. ▫

gbm paket u R-u

Tri su hiperparametra:

(i) M – broj stabala. Iako veći M može overfittingu, ali u praksi se nije toga dogodilo. → CV dovoljno za odabir

[još nije skroz jasan ovaj teorem!]

(ii) $\lambda \in \langle 0, 1]$ - "shrinkage" parametar [regularizacija!]

- manji λ $\Rightarrow$ veći "optimalni" M,
  te tipično i bolja predikt. sposobnost!
  $\rightarrow$ problem ak je M prevelik

[alg. "sporije" uči]

[= tipično biramo λ što manji, npr. iz $\langle 0.001, 0.1 \rangle$]

(iii) J = broj čvorova u svakom stablu.

- kontrolira interakcije
- $\forall T_m$ imamo ukupno $(J-1)$ odvajanja
  $\Rightarrow$ moguće interakcije od najviše $(J-1)$ varijabli
- tipično $J \in \{4, \dots, 8\}$

Nap. 8.8| Po uzoru na slučajne šume, u svakom koraku $m = 1, \dots, M$, koraci (a) – (c) mogu se provesti na slučajno odabranom podskupu od p veličine $(p \cdot n)$

za $p \in \langle 0, 1]$ ("Stohastic GB")

[$\rightarrow$ default je $p = \frac{1}{2}$.]

Pr. 8.9|

(a) $L(y, t(x)) = \frac{1}{2}(y - t(x))^2$

$\Rightarrow -g_{im} = \left. \frac{-\partial L(y_i, t)}{\partial t} \right|_{t = t^{(m-1)}(x^{(i)})} = y_i - t^{(m-1)}(x^{(i)})$

"obični" reziduali za $t^{(m-1)}$

$\rightarrow$ ("L_2-boosting").

(b) $L(y, f(x)) = |y - f(x)|$

$$\Rightarrow -g_{i,m} = \operatorname{sgn}\left(y_i - f^{(m-1)}(x^{(i)})\right)$$

$$= \begin{cases} +1, & y_i > f^{(m-1)}(x^{(i)}) \\ -1, & < \end{cases}$$

← [upće ne ori'n' o $|y_i - f^{(m-1)}(x^{(i)})|$ ("robusnost")

(c) $L(y, f(x))$ = Huber iz (9.35)

$$\Rightarrow (\text{PZ}) \quad -g_{i,m} = \begin{cases} y_i - f^{(m-1)}(x^{(i)}), & \text{ako } |y_i - f(x^{(i)})| < \delta \\ \delta \cdot \operatorname{sgn}(y_i - f^{(m-1)}(x^{(i)})), & \text{inače} \end{cases}$$

↳ δ se tipično bira iz $\bar{L}$ posebno $\neq m$. ∎

Nap. 8.10 | Interpretabilnost?

↳ utjecaj kovarijata te grafove parcijalne zavisnosti radimo isto kao za RF. ∎

→ L_2 - boosting - primjer

→ Boston - boosting.R

8.3.2 GB za binarnu klasifikaciju

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

↳ želimo procijeniti $\boxed{p^*(x) := \mathbb{P}(Y=1 \mid X=x), \ \forall x \in \mathbb{R}^p}$

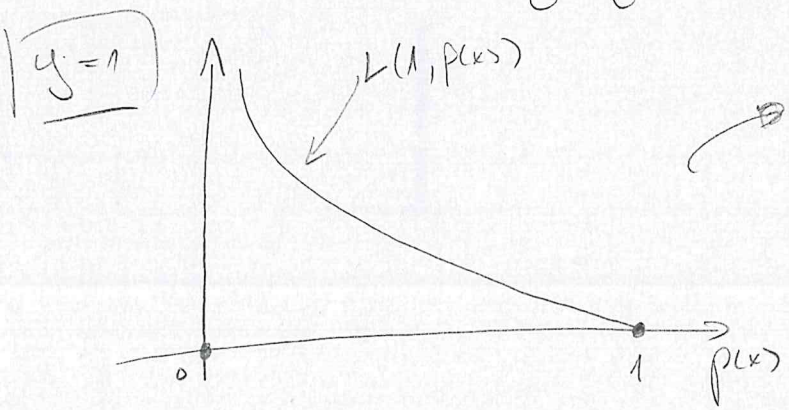
- tzv. Bernoullijer gubitak je $L : S \times [0,1] \to [0,+\infty\rangle$ (negativna log-vjerodost.)

je

$$L(y, p(x)) = -\log\left(\mathbb{P}(B(p(x)) = y)\right)$$

$$= \begin{cases} -\log p(x), & \text{ako } y=1 \\ -\log(1-p(x)), & \text{ako } y=0 \end{cases}$$

$$= -y \log p(x) - (1-y)\log(1-p(x)) \qquad (8.43)$$



$L(1,p(x)) > 0$ i

u slučaju $p(x) \geq \frac{1}{2}$,

tj. kada je

$\hat{f}(x) := \mathbb{1}_{\{p(x) \geq \frac{1}{2}\}} = 1 = y$!

[↳ želimo $p(x)$ što bliže (1)!]

- za DZ pokažite da je

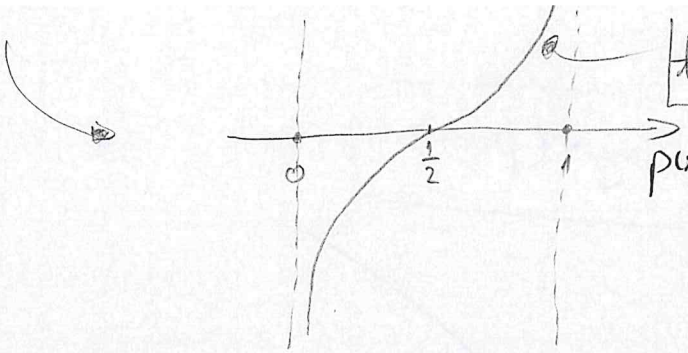
$$p^*(x) = \underset{c \in [0,1]}{\text{arg min}} \ \mathbb{E}\left[L(Y,c) \mid X=x\right] \qquad (8.44)$$

- ekvivalentno, umjesto $p^*(x)$ možemo modelirati

kao u log-regresiji!

$$f^*(x) := \log \frac{p^*(x)}{1-p^*(x)} \quad \boxed{\in \mathbb{R}} \qquad (8.45)$$

$(= \text{logit}(p^*(x)))$



$$t(x) := \text{logit}(p(x)) \qquad (8.46)$$

"skor"

Također,

$$\bullet\ p(x) = \frac{e^{t(x)}}{1+e^{t(x)}} = \frac{1}{1+e^{-t(x)}} \qquad (8.47)$$

$$\bullet\ 1-p(x) = \frac{1}{1+e^{t(x)}} \qquad (8.48)$$

↳ Bernoullijev gubitak (s obzirom na t) je

$$L(y, t(x)) = -y \log p(x) - (1-y)\log(1-p(x))$$

$$\overset{(8.47)}{\underset{(8.48)}{=}} -y\,t(x) + \log\left(1+e^{t(x)}\right), \qquad (8.49)$$

a iz (8.44) vidi da je

$$t^*(x) = \log\frac{p^*(x)}{1-p^*(x)} = \underset{c\in\mathbb{R}}{\text{argmin}}\ \mathbb{E}\left[L(y,c)\,|\,X=x\right] \qquad (8.50)$$

$$=: t^*_{\text{Bern}}(x)$$

⇒ ako u 6.3 algoritmu iz (8.3.1) uzmemo L iz (8.49), projekcijom $t^*(x) = \log\frac{p^*(x)}{1-p^*(x)}$, a pseudo-reziduali su

$$-g_{i,m} = -\left.\frac{\partial L(y_i, t)}{\partial t}\right|_{t=t^{(m-1)}(x^{(i)})}$$

$$\overset{(8.47)}{\underset{(8.49)}{=}} y_i - p^{(m-1)}(x^{(i)}) \qquad (8.51)$$

- ako je f proizvoljna za $f^*(x)$, definiramo $\hat{f}: \mathbb{R}^p \to S$ sa

$$\hat{f}(x) := \mathbb{1}_{\{f(x) \geq 0\}} \quad \left(= \mathbb{1}_{\{p(x) \geq \frac{1}{2}\}} \right) \qquad (8.52)$$

- za Bem. gubitak vrijedi

$$L(y, f(x)) \overset{(8.47) \to (8.48)}{=} \underbrace{y}_{\in\{0,1\}} \log\left(1 + e^{-f(x)}\right) + \underbrace{(1-y)}_{\in\{0,1\}} \log\left(1 + e^{f(x)}\right)$$

$$= \log\left(1 + e^{-\tilde{y} \cdot f(x)}\right), \qquad (8.53)$$

gdje je $\tilde{y} := 2y - 1 \in \{-1, 1\}$.

↳ uočimo,

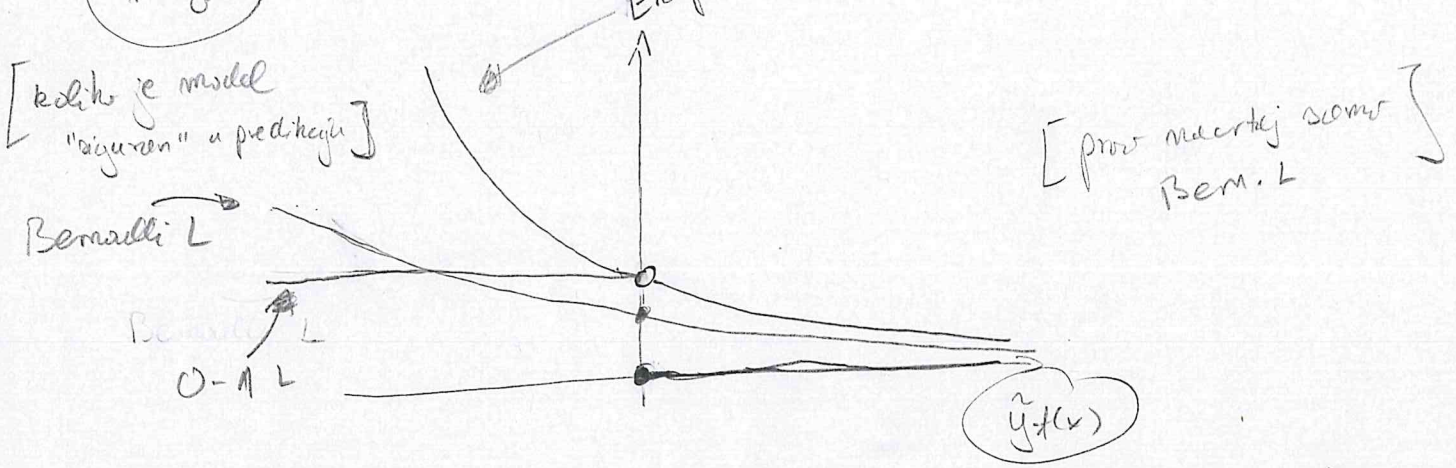
$$f(x) = y \overset{(8.52)}{\iff} \tilde{y} \cdot f(x) \geq 0$$

te

$$\tilde{y} f(x) \nearrow +\infty \Rightarrow p_y(x) = \begin{cases} p(x), & y=1 \\ 1-p(x), & y=0 \end{cases} \nearrow 1$$

$$(\searrow -\infty) \qquad (\searrow 0)$$

"margin" $= \tilde{y} f(x)$



- alternativno za Bem. gubitak je f-ja L t.d. $\forall f: \mathbb{R}^p \to \mathbb{R}$,

$$L(y, f(x)) = e^{-\tilde{y} f(x)} \qquad (8.54)$$

"Eksponencijalni gubitak"

Nap. 8.11

(i) Iz (8.53) imamo da za Bern. gubitak vrijedi:

$\tilde{y} f(x) \to -\infty \Rightarrow L(y, f(x)) \sim \boxed{-\tilde{y} f(x)}$ ← linearan!

[primjer na kojima f jako griješi!]

↳ Eksponencijalni gub. ima veći problem s "outlierima" [npr. ako imamo greške u podacima] [slično kao L_2-gub. za regresiju]

S druge strane, iz (8.53) imamo

$\tilde{y} f(x) \to +\infty \Rightarrow L(y, f(x)) \sim \boxed{e^{-\tilde{y} f(x)}}$ (Bern.)

$\log(1+x) \sim x$, kada $x \to 0$

(isto kao Eksp. L)

↳ [nacrtaj Eksp. L na slici].

(ii) Oba gubitka su monotona, konveksna i diferencijabilna aproksimacija 0-1 gubitka

$L(y, f(x)) = \mathbb{1}\{\boxed{\tilde{y} f(x) < 0}\}$

↳ [dodaj 0-1 gub. na slici]

Nap. 8.12 GB uz eksponencijalni gubitak je ekvivalentan tzv. Adaboost algoritmu (Freund, Schapire, 1996) (vidi ESL, str. 10)

→ prvi uspješan boosting algoritam

→ iako je manje robustan od GB algoritma uz Bern. gubitak, postoje jako efikasne implementacije → vidi npr. Viola-Jones algoritam za detektiranje lica

→ tu ide Adaboost zapravo optimizira eksponencijalni gubitak je u originalnom algoritmu bilo skriveno implicitno ⇒ Friedman et. al. (2000) to pokazuju, te je to zapravo dovelo do GB algoritma u Friedman (2001)! ||

13

- za DZ pokažite da za eksp. gubitak vrijedi

$$f^*_{Eksp}(x) := \underset{c \in \mathbb{R}}{\arg\min}\, \mathbb{E}[L(Y,c) \mid X=x] = \frac{1}{2} f^*(x) \qquad (8.55)$$

⇒ ako je f procjena za f^*_{Eksp}, $(2f)$ je procjena za f^*, tj. $\frac{1}{1+e^{-2f(x)}}$ je procjena za $p^*(x)$ (vidi (8.47)).

↳ specijalno, opet stavljamo

$$\hat{f}(x) := \{ h_i \; f(x) \geq 0 \} = \{ h_i \; 2f(x) \geq 0 \}.$$

⇒ GB algoritam uz eksp. gubitak procjenjuje f^*_{Eksp}, a pseudo-reziduali su

$$-g_{i,m} \overset{(DZ)}{=} \tilde{y}_i e^{-\tilde{y}_i f^{(m-1)}(x^{(i)})} \qquad (8.56)$$

↳ $\tilde{y}_i f^{(m-1)}(x^{(i)}) < 0 \Rightarrow |-g_{i,m}| = e^{|f^{(m-1)}(x^{(i)})|}$

[brine klasificirani primjeri]

[koliko smo bili "sigurni" u klasifikaciju]

Nap.

- značajnost varijabli i gradijenti parcijalne ovisnosti (za $f^{(m)}$, ne za $\Pi^{(m)}$) kao i za regresiju
- u slučaju $Y \in \{0, \dots, K-1\}$, za $K \geq 3$, modeliramo (K) f-ja $f_0, \dots, f_{K-1}$ kao u multinomijalnoj log. regresiji: npr. $f_0 \equiv 0$, a za $k \geq 1$

$$f_k(x) = \log \frac{P(Y=k \mid X=x)}{P(Y=0 \mid X=x)}$$

$$\text{tj. } P(Y=k \mid X=x) = \frac{e^{f_k(x)}}{\sum_{\ell=0}^{K-1} e^{f_\ell(x)}}$$

[izostavljamo detalje]

- ako je $Y \in \{0,1,2,\dots\} = \mathbb{N}_0$, kao u GLM-u, za Poissonovu razdiobu, GB algoritmom možemo modelirati

$$f(x) := \log(\mathbb{E}[Y \mid X=x]) \quad (\in \mathbb{R}!)$$

te koristiti

$$L = -\log\text{-vjerodostojnost}$$

. ☞ 1