Ansambli napovednih modelov

Ljupčo Todorovski

Univerza v Ljubljani, Fakulteta za upravo
Institut Jožef Stefan, Odsek za tehnologije znanja (E-8)

April 2020

Pregled predavanja

Osnovna ideja in definicije

- Splošna metoda za zmanjševanje variance
- Komponente ansambla

Homogeni ansambli

- Vzorčenje primerov: Bagging in Boosting
- Vzorčenje spremenljivk: naključni prostori
- Vzorčenje primerov in spremenljivk: naključni gozd

Heterogeni ansambli

Stacking in heterogeni napovedni modeli

Osnovna ideja

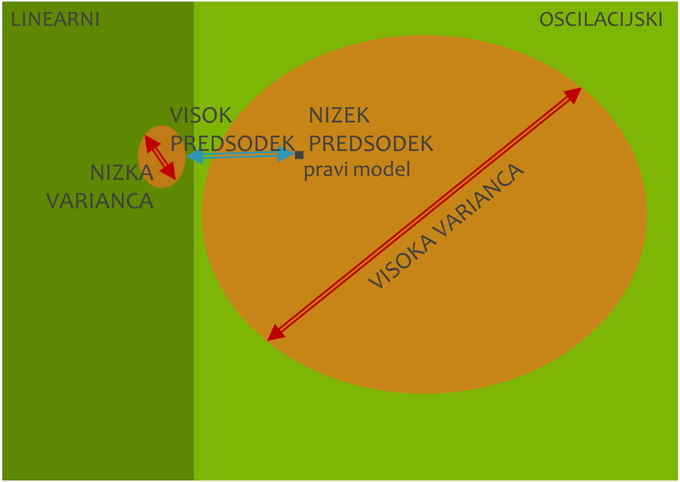
Namesto enega, se naučimo več napovednih modelov

1. Lahko spreminjamo učne podatke
2. Lahko uporabljamo različne algoritme strojnega učenja

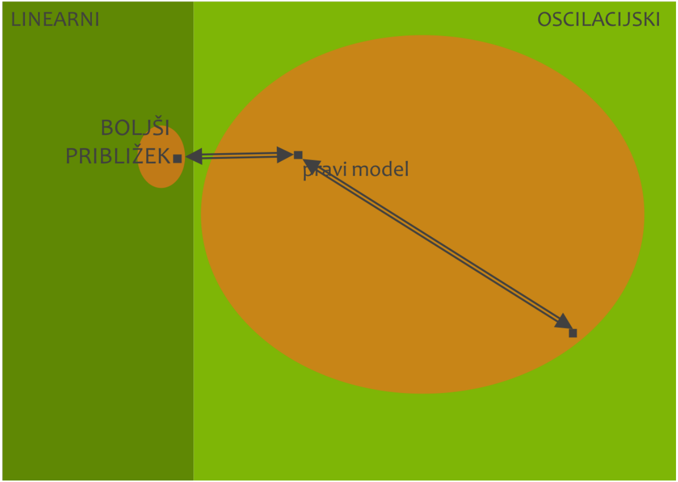
Zakaj več modelov namesto enega?

S kombiniranjem njihovih napovedi
lahko zmanjšamo varianco
in s tem tudi napovedno napako.

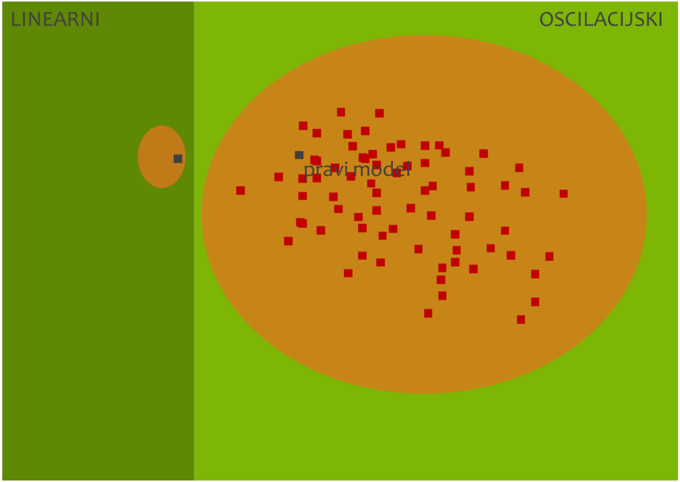
Prostora modelov: kompromis med pristranskostjo in varianco



Možna posledica: napačna izbira modela



Ansambel namesto izbire enega modela



Struktura ansambla

Osnovni modeli $m_i : i = 1..r$, tudi sestavine ansambla $\hat{M}$

Vsi napovedujejo vrednost ciljne spremenljivke Y, $\hat{m}_i : \cdot \to D_Y$.

Kombiniranje napovedi sestavin $\hat{y}_i$ v napoved $\hat{y}$ ansambla M

- Večinsko glasovanje: vrne vrednost, ki se največkrat pojavi v množici $\{\hat{y}_i : i = 1..r\}$
- Povprečje: $\hat{y} = \frac{1}{r} \sum_{i=1}^{r} \hat{y}_i$
- V splošnem funkcija $C_M : \bigtimes_{i=1}^{r} D_Y \to D_Y$, $\hat{y} = C_M(\hat{y}_1, \ldots \hat{y}_r)$

Heterogeni in homogeni ansambli

Tipologija ansamblov glede načina učenja sestavin.

Homogeni ansambli

- Vse sestavine naučene z istim algoritmom
- Vsaka sestavina naučena iz spremenjene učne množice

Heterogeni ansambli

- Sestavine naučene z različnimi algoritmi
- Običajno vse sestavine naučene iz iste učne množice

Spomnimo se: Komponente napovedne napake

σ_ϵ^2 je fiksna komponenta

Na to komponento nimamo vpliva.

Pristranskost $(E_S[\hat{m}(\boldsymbol{x}_0)] - m(\boldsymbol{x}_0))^2$

- Razlika med pričakovano vrednostjo napovedi $\hat{Y}$ in vrednostjo Y
- Za podani prostor modelov nam pove koliko se lahko $\hat{Y}$ približa Y (model $\hat{m}$ približa m)

Varianca $Var_S[\hat{m}(\boldsymbol{x}_0)]$

- Je varianca napovedi $\hat{Y}$ okoli pričakovane vrednosti $E_S[\hat{Y}]$
- Stabilnost napovedi $\hat{Y}$ modela $\hat{m}$

Dekompozicija napake za homogene ansamble

σ_ϵ^2 je fiksna komponenta

Ta komponenta ostane nespremenjena tudi pri homogenem ansamblu.

Pristranskost ansambla je enaka pristranskosti osnovnih modelov

$$(E_S[\hat{M}(\boldsymbol{x}_0)] - m(\boldsymbol{x}_0))^2 = (E_S[\hat{m}(\boldsymbol{x}_0)] - m(\boldsymbol{x}_0))^2$$

Izpeljava na tabli.

Faktor zmanjševanja variance homogenega ansambla

$$Var_S[\hat{M}(\boldsymbol{x}_0)] = \left(\rho(x_0) + \frac{1-\rho(x_0)}{r}\right) Var_S[\hat{m}(\boldsymbol{x}_0)]$$

$\rho(x_0)$ je korelacija med napovedmi osnovnih modelov, izpeljava na tabli.

Faktor zmanjševanja variance

$$\rho + \frac{1-\rho}{r}$$

Če je korelacija ρ blizu 1

- Se faktor zmanjševanja variance približa ρ, torej 1
- **Raznovrstnost napovedi** sestavin **zmanjšuje** varianco ansambla

Če je korelacija ρ blizu 0

- Se faktor zmanjševanja variance približuje $1/r$
- **Velikost ansambla** r **zmanjšuje** njegovo varianco

Učenje osnovnih modelov $m_i = \mathcal{A}(V_i)$

r vzorcev $V_i : i = 1..r$ učne množice S

- $V_i : |V_i| = |S|$, vzorčenje s ponavljanjem
- Verjetnost $p(e \notin V_i) = (1 - 1/|S|)^{|S|}$

$$\lim_{|S| \to \infty} p(e \notin V_i) = \frac{1}{e} = 0.368$$

Napovedi izven vreče (*out-of-bag*, OOB)

Napoved OOB za en primer e

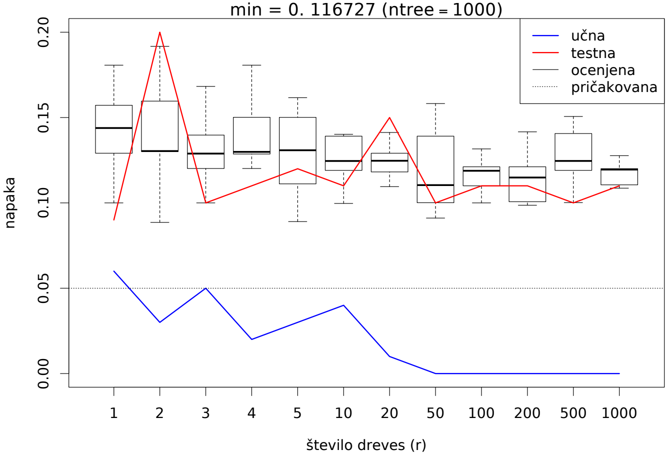
$$\hat{y}_{OOB} = C_M(\hat{y}_{i_1}, \hat{y}_{i_2}, \ldots \hat{y}_{i_s})$$

- $\hat{y}_i$ je napoved i-tega osnovnega modela
- Za vse $i_j : j = 1..s$ velja $e \notin V_{i_j}$
- Upoštevamo torej le napovedi sestavin, ki niso naučene iz primera e

Ocena napake OOB za množico primerov S

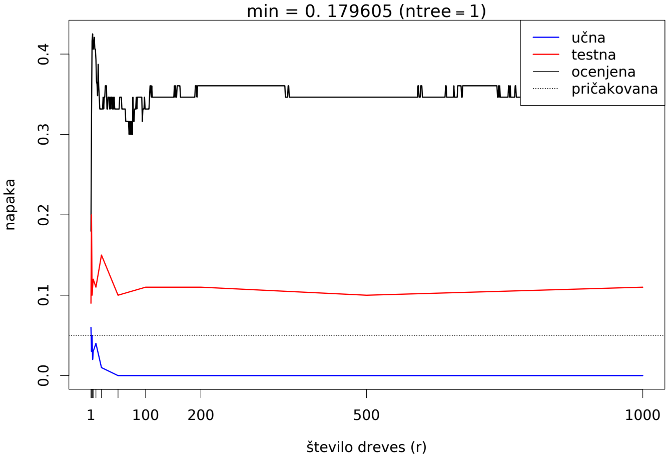
$$Err_{OOB} = \sum_{(\boldsymbol{x},y)\in S} L(\hat{y}_{OOB}, y)$$

Klasifikacija z vrečenjem: število osnovnih modelov r

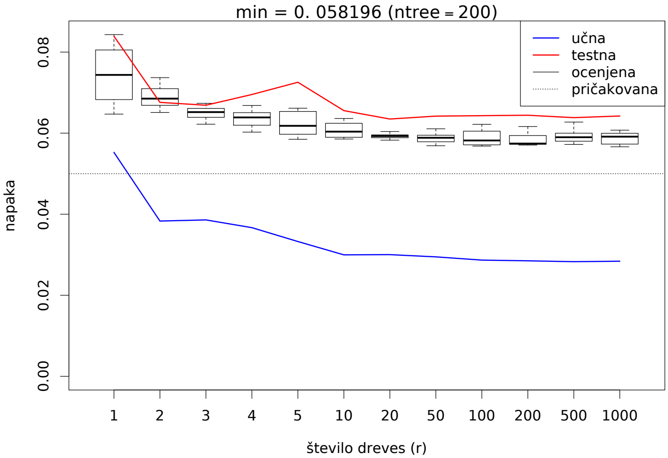


$Y = I((1 + X_1 + X_1X_2)/3 \geq 0.5), D_i = [0, 1], i = 1..2, D_Y = \{0, 1\}$
zamenjamo vrednost Y 0.05 naključno izbranim primerom

Klasifikacija z vrečenjem: r, napaka OOB

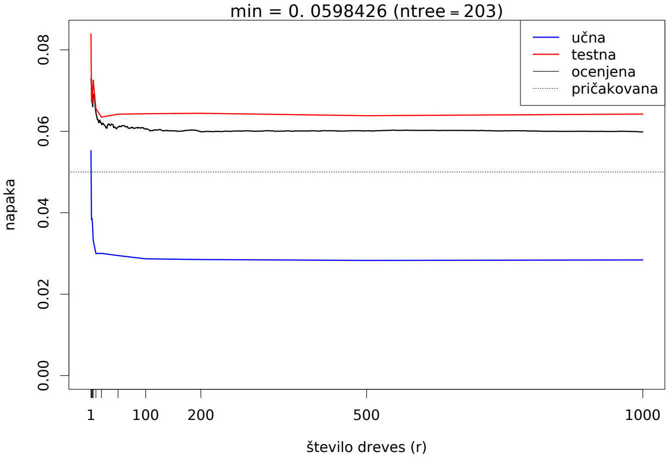


Regresija z vrečenjem: število osnovnih modelov r



$$Y = (1 + X_1 + X_1 X_2)/3 + \mathcal{N}(0, 0.05),\ D_i = [0,1], i = 1..2, D_Y = [0,1]$$

Regresija z vrečenjem: r, napaka OOB



Naključni pod-prostori (*random subspaces*)

Učenje osnovnih modelov $m_i = \mathcal{A}(V_i)$

- Vsak iz učne množice z naključno izbranim vzorcem napovednih spremenljivk $V_{\mathbf{X}} \in \mathbf{X} = \{X_1, X_2, \ldots X_p\}$, vzorčenje **brez** ponavljanja
- $V_i = S|_{V_{\mathbf{X}}}$ je vzorec S, ki vsebuje le vhodne spremenljivke iz $V_{\mathbf{X}}$
- Parameter $m \leq p$: velikost vzorcev $|V_{\mathbf{X}}| = m$

Kaj pa ocena napake OOB?

Ker ni vzorčenja primerov, ni na voljo napovedi OOB.

Naključni gozd (*random forest*)

Kombinacija metod vrečenja in naključnih pod-prostorov

- m_i je odločitveno drevo, običajno naučeno brez rezanja
- V_i je vzorec s ponavljanjem velikosti $|S|$, tako kot pri vrečenju
- Pred vsako izbiro testa v drevesu, naključni izbor $m \leq p$ spremenljivk

Večja različnost sestavin kot pri vrečenju

- Če pri vrečenju ima ena spremenljivka X veliko napovedno točnost, bodo vsi osnovni modeli podobni
- Pri naključnem gozdu pa ne, saj lahko pričakujemo drevesa brez X

Kontrola pristranskosti in variance

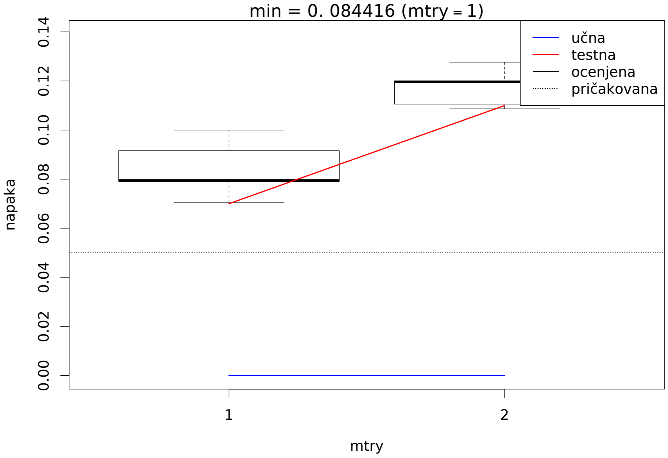
Velikost ansambla: število osnovnih modelov r

- Več sestavin, manj variance ob nespremenjeni pristranskosti
- Torej r nastavimo na čim višjo vrednost

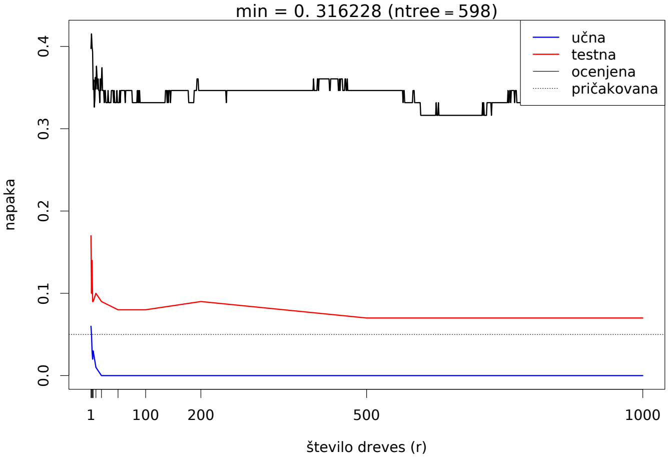
Parameter *mtry*: število izbranih spremenljivk za izbiro testa

- Ni jasne povezave med vrednostjo *mtry* in varianco
- Privzeta vrednost $mtry = \sqrt{p}$
- Izbira z opazovanjem napake OOB

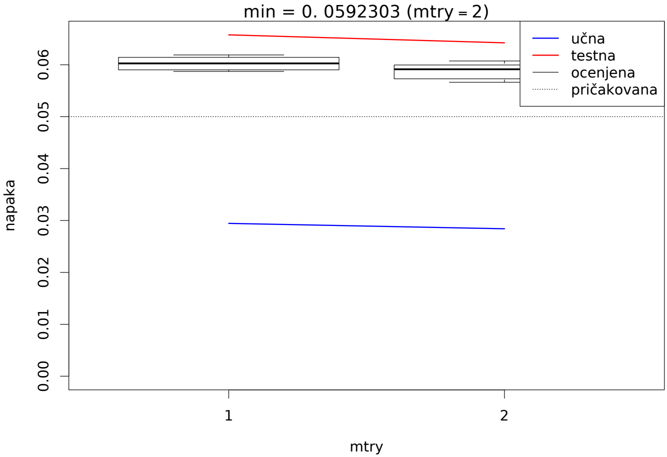
Klasifikacija z NG: parameter *mtry*



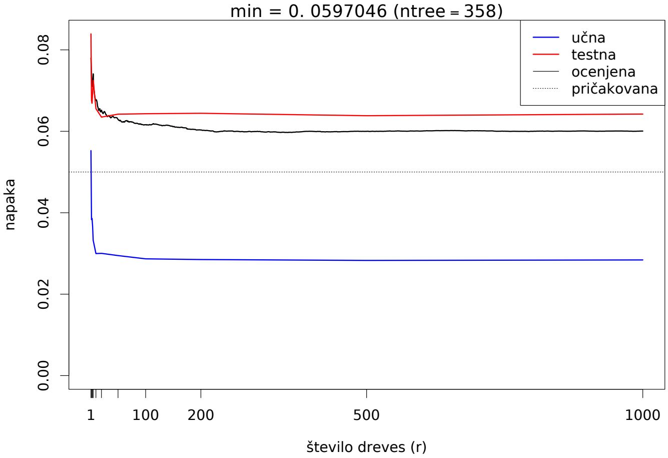
Klasifikacija z NG: r, $mtry = 1$, OOB



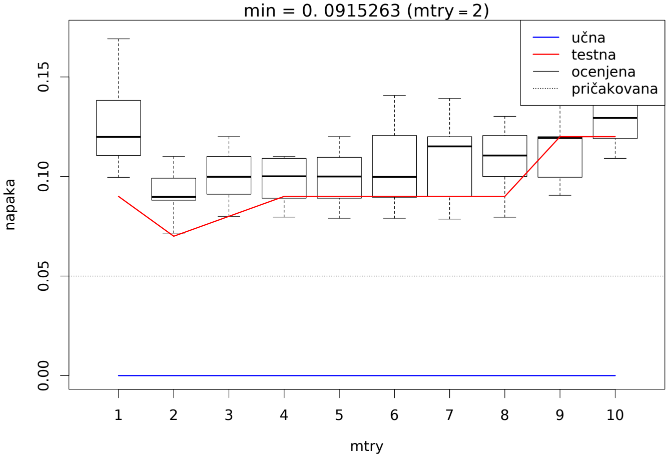
Regresija z NG: parameter *mtry*



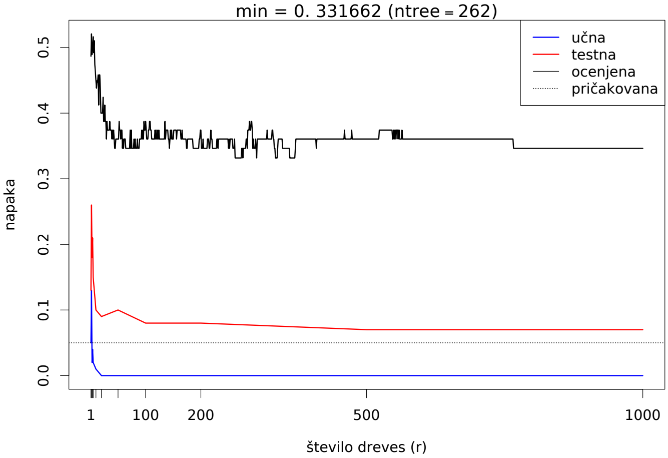
Regresija z NG: r, $mtry = 2$, OOB



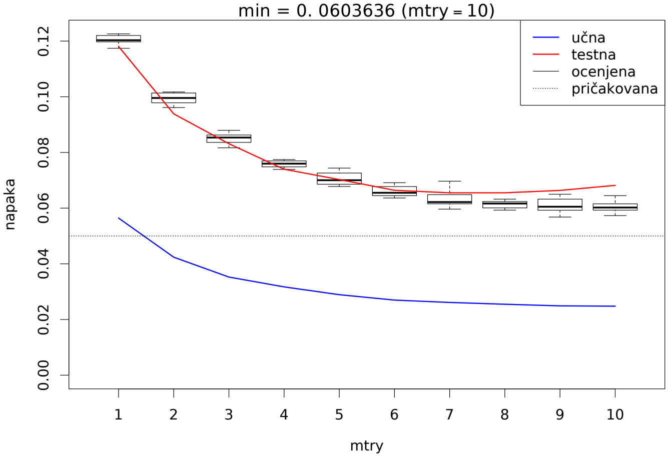
Klasifikacija z NG ($p = 10$): parameter *mtry*



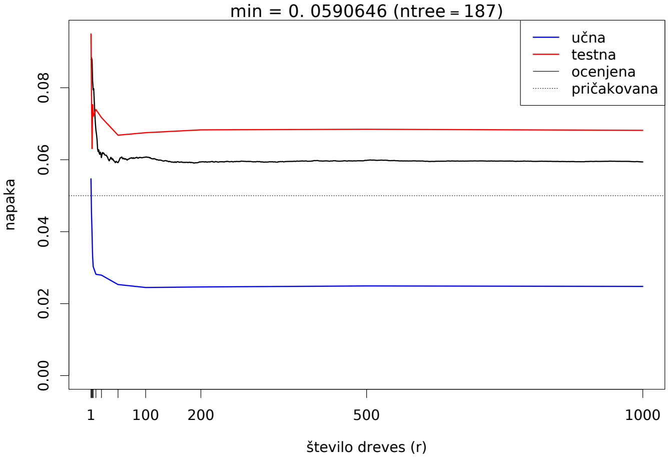
Klasifikacija z NG ($p = 10$): r, $mtry = 2$, OOB



Regresija z NG ($p = 10$): parameter *mtry*



Regresija z NG ($p = 10$): r, $mtry = 10$, OOB



Razlaga napovedi ansambla

Moč razlage napovednega modela

- Ansambli ne ponujajo razumljive razlage napovedi
- Za razliko od odločitvenega drevesa ali liste pravil
- Tudi linearna regresija: predznak in velikosti parametrov

Poskus razlage ansamblov

Dva načina izračuna relevantnosti napovednih spremenljivk za napoved.

Povprečno zmanjševanje nečistoče

Relevantnost napovedne spremenljivke X v odločitvenem drevesu t

$IR(t, X)$ je povprečna vrednost zmanjševanja nečistosti IR v notranjih vozliščih drevesa, ki testirajo vrednost neodvisne spremenljivke X.

Relevantnost spremenljivke X v ansamblu M

$$IR(M, X) = \frac{1}{r} \sum_{i=1}^{r} IR(m_i, X)$$

- Povprečje relevantnosti X v vseh sestavinah ansambla
- Normalizacija: največja vrednost pomena enaka 1 (ali 100%)

Povprečno zmanjševanje točnosti

Kaj pa če osnovni modeli niso odločitvena drevesa?

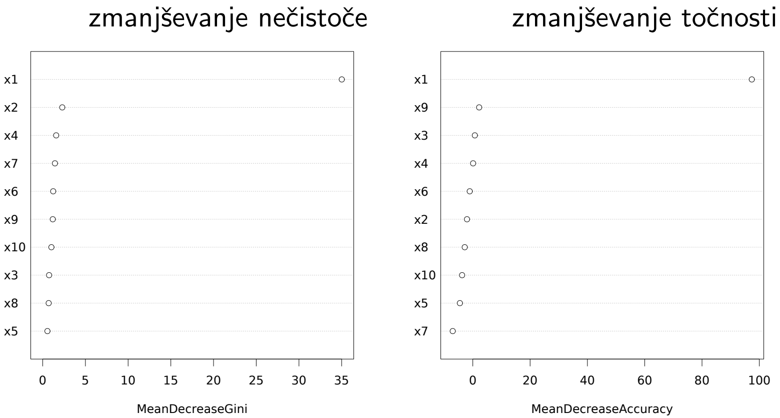
Izračunamo razliko $D = Err(M_P, S) - Err(M, S)$ med

- Napako modela M na množici S
- Napako modela M_P naučenega iz podatkovne množice S_P, kjer so vrednosti spremenljivke X naključna permutacija vrednosti X v S
- $Err(M_P, S)$ je napaka modela, kjer je vpliv spremenljivke X izničen

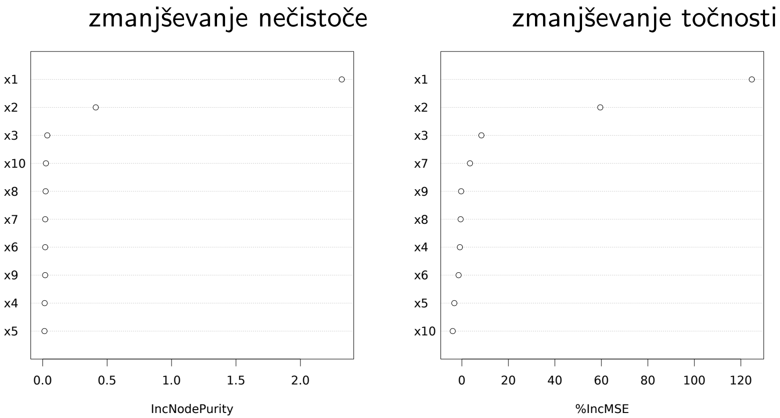
Vrednost razlike D je povečanje napake (in zmanjševanje točnosti)

- Visoka, če je X za napoved relevantna spremenljivka
- Nizka, če je X nepomembna za napovedovanje

Klasifikacija z NG: relevantnost napovednih spremenljivk



Regresija z NG: relevantnost napovednih spremenljivk



Osnovna ideja

- Zgradimo m_1 na učni množici S
- Ponavljaj $r-1$ krat: za $i = 2..r$
  - Opazuj napako ansambla do sedaj zgrajenih modelov na primerih iz S
  - Vzorči S s ponavljanjem: primeri z visoko napako bolj zastopani v V
  - $m_i = \mathcal{A}(V)$
  - Namesto vzorčenja, spreminjanje uteži učnih primerov

Vzorčenje primerov ne popolnoma naključno

Boosting se osredotoča na učne primere,
kjer ima model visoko napovedno napako.

Klasifikacija, AdaBoost

Spreminjanje uteži učnih primerov za model m_i

$$E_w = E_w(m_i) = \sum_{(w,\boldsymbol{x},y)\in V : m_i(\boldsymbol{x})\neq y} w$$

- Če je $E_w > 1/2$ prekinemo delovanje algoritma
- Uteži primerov e, ki jih trenutni ansambel pravilno razvrsti, zmanjšamo $w_e \leftarrow w_e\, E_w/(1-E_w)$
- Uteži vseh primerov normaliziramo, tako da $\sum_{e\in S} w_e = 1$

Utež modela m_i in trenutni ansambel M_i

$$\alpha_i = \frac{1}{2}\log\frac{1-E_W}{E_W}, \quad M_i = \sum_{j=1}^{i} \alpha_j m_j$$

Končni ansambel $M = M_r$.

Regresija, GradientBoosting

Spreminjanje vrednosti ciljne spremenljivke

$$y_{i+1} = y_i - M_i(\boldsymbol{x})$$

Vrednosti ciljne spremenljivke zamenjamo z ostanki trenutnega ansambla

Utež modela m_i in trenutni ansambel M_i

$$\gamma_i = \arg\min_{\gamma} \sum_{(\boldsymbol{x},y)\in S} L(y, M_{i-1}(x) + \gamma\, m_i(\boldsymbol{x})), \quad M_i = \sum_{j=1}^{i} \gamma_j m_j$$

Končni ansambel $M = M_r$.

Kompromis med pristranskostjo in varianco

Povečanje velikosti ansambla r

- Lahko zapelje boosting k preprileganju
- Zato pozorno opazovanje učne in testne napake pri naraščajočem r

Nastavitev parametrov osnovnega algoritma $\mathcal{A}$

- Boosting lahko zmanjša pristranskost osnovnih modelov, na račun povečane variance
- Zato izberemo osnovni algoritem z visoko pristranskostjo
- Majhna, sproti porezana odločitvena drevesa ali linearne modele

Stacking

Učenje osnovnih modelov za *stacking*

- Različni algoritmi, ista podatkovna množica
- Napovedi osnovnih modelov na tesnih podatkih (prečno preverjanje)
- Tako dobljene napovedi postanejo nove napovedne spremenljivke ($\boldsymbol{X'}$)

Kombiniranje napovedi osnovnih modelov

$$C_M = \mathcal{A}(S')$$

- Novim spremenljivkam $\boldsymbol{X'}$ dodamo ciljno spremenljivko Y in tako dobimo učno množico S'
- Funkcija kombinacije C_M je napovedni model naučen iz S'

Heterogeni napovedni modeli

Posebna vrst ansamblov, ki bolj neposredno povezujejo različne modele

Nekaj primerov: kombinacija dreves in drugih napovednih modelov

- Modelska drevesa: regresijska drevesa, kjer napovedi v končnih vozliščih podajo linearni modeli
- Drevesa najbližjih sosedov: odločitvena drevesa, kjer napovedi v končnih vozliščih poda metoda najbližjih sosedov
- Meta drevesa: odločitvena drevesa, kjer napovedi v končnih vozliščih določajo kateri model bo podal napoved

Nevarnost heterogenih modelov: povečanje variance

- V nasprotju z običajno lastnostjo ansamblov
- Zaradi zapletenih modelov v listih se sicer zmanjša pristranskost modela, a hkrati se povečuje varianca

Znani algoritmi in implementacije

Boosting (Schapire 1990) in AdaBoost (Freund and Schapire 1996)

Kako iz slabih napovednih modelov zgraditi dobre?

Vrečenje oz. bagging (Breiman 1994)

Kako zmanjšati varianco?

Naključni gozdovi (Breiman 2001)

Kombinacija vrečenja z idejami naključnih pod-prostorov