Odnos između pristranosti i varijance u metodama statističkog učenja

Mateša, Fabian

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Fabian Mateša

ODNOS IZMEĐU PRISTRANOSTI I VARIJANCE U METODAMA STATISTIČKOG UČENJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Hrvoje Planinić

Zagreb, veljača, 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Statističko učenje je polje znanosti koje se bavi razvojem i proučavanjem metoda za izradu modela baziranih na datom skupu podataka, tzv. skupu za učenje, u svrhu generiranja predviđanja ili odluka temeljem statističke analize. U posljednjih nekoliko desetljeća, s adventom računala sposobnih za brzo procesuiranje sve većih količina podataka, ove metode su se počele primjenjivati pri rješavanju brojnih problema u raznim granama znanosti, te su, u većini slučaja na suptilan način, postale neizbježan dio svakodnevice.
Iako ove metode nerijetko daju odlične rezultate, veliki problem predstavlja samo razumijevanje istih. Naime, posebice kada su u pitanju kompleksnije metode, njihovo korištenje se može okarakterizirati kao "crna kutija": Dobiveni model prima podatke na temelju kojih "izbacuje" rezultate bez ikakvog objašnjenja kako je do tih rezultata došlo, što uvelike otežava daljni razvoj, te postavlja etička pitanja oko njihove aplikativne primjene u slučajevima kada ti isti rezultati mogu imati izrazito velike posljedice.
U nadziranom učenju, jedan od osnovnih alata za razumijevanje modela je rastav greške modela na pristranost, varijancu i ireducibilnu grešku uzrokovanu šumom u samim podacima. Kada je to moguće, optimizacija modela se svodi na minimizaciju pristranosti i varijance, budući da se na šum ne može utjecati. Ipak, takva dekompozicija je definirana isključivo za srednjekvadratnu grešku.
Cilj ovog rada je razviti univerzalnu definiciju pojmova pristranosti i varijance neovisno o odabranoj funkciji gubitka, te proučiti kako se greška modela ponaša kao funkcija te dvije vrijednosti. Rezultati su većinom preuzeti iz Domingosovog rada *A Unified Bias-Variance Decomposition*, objavljenog 2000. godine [2].
U prvom poglavlju cilj je ustanoviti generalizirane definicije pristranosti, šuma i varijance, neovisne o odabiru funkcije gubitka, te otkriti njihove odnose s očekivanom testnom greškom. Navodi se generalizirani oblik dekompozicije greške, te se pokazuje njena primjena u slučaju srednjekvadratne i 0-1 funkcije gubitka. Također se daju za primjer funkcije za koje dekompozicija ne postoji u datom obliku. Pokazano je i da dekompozicija vrijedi za sve funkcije koje zadovoljavaju uvjete Bregmanove divergencije, među kojima je i funkcija log-gubitka. Na kraju poglavlja se spominje uloga pristranosti u efikasnosti bagging metode, te teorem u kojemu su za metrike gornja i donja granica očekivane testne greške iskazane u pojmovima šuma, pristranosti i varijance.

Drugo poglavlje sadrži primjenu rezultata dobivenih u radu na simuliranom primjeru. Promatra se odnos pristranosti i varijance modela dobivenih metodom k-najbližih susjeda, za 0-1 funkciju gubitka.
Naposljetku, dodatak sadrži kratak opis metoda statističkog učenja korištenih u radu.

Poglavlje 1

Odnos pristranosti i varijance

Neka je Y slučajna varijabla koja poprima vrijednosti u skupu $S \subseteq \mathbb{R}$ zavisna o slučajnom vektoru $X = (X_1, ..., X_p)$ sa vrijednostima u u $\mathbb{R}^p$. Za dani skup za učenje $\tau = \{(x^{(i)}, y_i) : i = 1, ..., n\}$ dobiven kao realizacija uzorka $T = \{(X^{(i)}, Y_i) : i = 1, ..., n\}$, pri čemu vrijedi $(X^{(i)}, Y_i) \sim (X, Y)$, metoda statističkog učenja daje model $\hat{f} = \hat{f}(\tau)$, koji za testnu vrijednost x vraća predikciju $\hat{y} = \hat{f}(x)$. U nastavku teksta, koristimo nazive odziv za varijablu Y, te kovarijate za vektor X.
Točnost procjene za realizaciju kovarijate x se mjeri proizvoljno definiranom funkcijom gubitka $L : S^2 \to [0, +\infty)$, te je cilj statističkog učenja pronaći model koji minimizira testnu grešku od $\hat{f}$, definiranu kao

$$L(\hat{f}) := \mathbb{E}[L(Y, \hat{f}(X))] = \int_{\mathbb{R}^p} L_x(\hat{f}) \mathbb{P}_X(dx) \tag{1.1}$$

za $L_x(\hat{f}) := \mathbb{E}[L(Y, \hat{f}(X))|X = x]$.
Neke od često korištenih funkcija gubitka su srednjekvadratna ($L(y, \hat{f}(x)) = (y - \hat{f}(x))^2$) i apsolutna ($L(y, \hat{f}(x)) = |y - \hat{f}(x)|$) u slučajevima kada je $S = \mathbb{R}$ (problem regresije), te 0-1 funkcija gubitka ($L(y, \hat{f}(x)) = 0$ ako je $y = \hat{f}(x)$, 1 inače) kada je S konačan podskup od $\mathbb{R}$ (problem klasifikacije).

1.1 Pristranost, varijanca i šum

Neka je $x \in \mathbb{R}^p$ realizacija kovarijate X za odziv Y.

Definicija 1.1.1. *Za danu funkciju gubitka $L : S^2 \to [0, +\infty)$, optimalnu predikciju od x definiramo kao*

$$y_* = y_*(x) = \arg\min_{i \in S} \mathbb{E}_x[L(Y, i)], \tag{1.2}$$

Optimalan model je tada onaj model za koji vrijedi $\hat{f}(x) = y_*, \forall x \in X$, no budući da je odziv slučajna varijabla koja ne mora biti u potpunosti određena vrijednostima kovarijata, očito je da i za takav model može vrijediti $L(\hat{f}) > 0$.

Napomena 1.1.2. *U gornjoj definiciji, kao i u ostatku poglavlja, koristimo oznaku* $\mathbb{E}_x[\cdot] := \mathbb{E}[\cdot|X = x]$.

Budući da $\hat{f}$ ovisi o τ, možemo promatrati $\hat{f} = \hat{f}(T)$, u kojem slučaju $\hat{f}$ postaje slučajna funkcija, a $L_x(\hat{f})$ slučajna varijabla.

Definicija 1.1.3. *Za danu funkciju gubitka* $L : S^2 \to [0, +\infty)$*, glavnu predikciju od x definiramo kao*

$$y_m^{L,T} = y_m^{L,T}(x) = \arg\min_{i \in S} \mathbb{E}_T[L(i, \hat{f}(x))] \tag{1.3}$$

U slučajevima kada je jasno o čemu se radi, $y_m^{L,T}$ označavamo jednostavno kao y_m. Intuitivno, to je vrijednost koja se po funkciji gubitka L u prosjeku "najmanje razlikuje" od predikcija modela naučenih na svim mogućim realizacijama τ slučajnog uzorka T.
Sada možemo definirati sve pojmove potrebne za dekompoziciju greške modela.

Definicija 1.1.4. *Neka je* $x \in \mathbb{R}^p$*. Pristranost modela u x definiramo kao*

$$B(x) = L(y_*, y_m), \tag{1.4}$$

varijancu u x kao

$$V(x) = \mathbb{E}_T[L(y_m, \hat{f}(x))], \tag{1.5}$$

te šum u x kao

$$N(x) = \mathbb{E}_x[L(Y, y_*)]. \tag{1.6}$$

Ove definicije odgovaraju dosadašnjim intuitivnim razumijevanjima tih pojmova. y_m je centralna predikcija metode, odnosno predikcija kojoj teži model dobiven metodom kada se isključi utjecaj slučajnosti odabira skupa za učenje, pa pristranost označava odstupanje te predikcije od one optimalne. S druge strane, varijanca promatra isključivo utjecaj slučajnosti odabira skupa za učenje, te mjeri prosječno odstupanje predikcija dobivenih modelima učenih na svim mogućim realizacijama skupa za učenje od glavne predikcije. Šum pak ne ovisi o samom modelu, već označava udio greške uzrokovan slučajnošću u samim podacima. Na njega se može gledati i kao na očekivanu vrijednost greške izmjerene nad optimalnim modelom.
Ponekad želimo promatrati prosjek pristranosti i varijance po svim vrijednostima x, u kojem slučaju se na njih referiramo kao *prosječna pristranost* $\mathbb{E}[B(X)]$ i *prosječna varijanca* $\mathbb{E}[V(X)]$.

1.2 Univerzalna dekompozicija greške modela

U nastavku, pokazat ćemo da je za određene funkcije gubitka L moguće očekivanu testnu grešku za dani x raspisati na sljedeći način:

$$\begin{aligned}\mathbb{E}_T[L_x(\hat{f})] &= c_1\mathbb{E}_x[L(Y, y_*)] + L(y_*, y_m) + c_2\mathbb{E}_T[L(y_m, \hat{f}(x))] \\ &= c_1 N(x) + B(x) + c_2 V(x),\end{aligned} \tag{1.7}$$

gdje su $c_1 = c_1(x)$ i $c_2 = c_2(x)$ multiplikativni faktori koji poprimaju različite vrijednosti ovisno o odabiru funkcije gubitka.

1.2.1 Srednjekvadratna greška

Pretpostavimo da je $\mathbb{E}[Y^2] < +\infty$. Poznato je tada $\arg\min_{c\in\mathbb{R}} \mathbb{E}[(Y-c)^2] = \mathbb{E}[Y]$, pa u slučaju srednjekvadratne greške vrijede sljedeće jednakosti:

$$y_* = \arg\min_{i\in\mathbb{R}} \mathbb{E}_x[(Y-i)^2] = \mathbb{E}_x[Y] \tag{1.8}$$

$$y_m = \arg\min_{i\in\mathbb{R}} \mathbb{E}_T[(\hat{f}(x)-i)^2] = \mathbb{E}_T[\hat{f}(x)] \tag{1.9}$$

$$B(x) = (y_* - y_m)^2 \tag{1.10}$$

$$V(x) = \mathbb{E}_T[(y_m - \hat{f}(x))^2] = \mathrm{Var}_T(\hat{f}(x)) \tag{1.11}$$

$$N(x) = \mathbb{E}_x[(Y-y_*)^2] = \mathrm{Var}_x(Y) \tag{1.12}$$

$$\tag{1.13}$$

Teorem 1.2.1. *Neka je $S = \mathbb{R}$. Pretpostavimo da je $\mathbb{E}[Y^2] < +\infty$. Rastav* (1.7) *vrijedi za srednjekvadratnu funkciju gubitka kada je $c_1 = c_2 = 1$.*

Dokaz. Računanjem dobivamo

$$
\begin{aligned}
\mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[E_x[(Y-\hat{f}(x))^2]] \\
&= \mathbb{E}_T[E_x[(Y-\mathbb{E}_x[Y]+\mathbb{E}_x[Y]-\hat{f}(x))^2]] \\
&= \mathbb{E}_T[\mathbb{E}_x[(Y-\mathbb{E}_x[Y])^2]+2\mathbb{E}_x[(Y-\mathbb{E}_x[Y])(\mathbb{E}_x[Y]-\hat{f}(x))] \\
&+ \mathbb{E}_x[(\mathbb{E}_x[Y]-\hat{f}(x))^2]] \\
&= \mathbb{E}_x[(Y-\mathbb{E}_x[Y])^2]+\mathbb{E}_T[(\mathbb{E}_x[Y]-\mathbb{E}_T[\hat{f}(x)]+\mathbb{E}_T[\hat{f}(x)]-\hat{f}(x))^2] \\
&= \mathbb{E}_x[(Y-\mathbb{E}_x[Y])^2]+(\mathbb{E}_x[Y]-\mathbb{E}_T[\hat{f}(x)])^2+\mathbb{E}_T[(\mathbb{E}_T[\hat{f}(x)]-\hat{f}(x))^2] \\
&= \mathbb{E}_x[(Y-y_*)^2]+(y_*-y_m)^2+\mathbb{E}_T[(y_m-\hat{f}(x))^2] \\
&= N(x)+B(x)+V(x).
\end{aligned}
$$

□

1.2.2 0-1 greška

Neka je S dvočlan skup, te bez smanjenja općenitosti pretpostavimo da je $S = \{0,1\}$. Tada vrijedi

$$y_* = \arg\min_{i\in S} \mathbb{E}_x[\mathbb{1}_{(Y\neq i)}] = \begin{cases} 1 & ,\mathbb{P}_x(Y=1) \geq \frac{1}{2} \\ 0 & ,\mathbb{P}_x(Y=1) < \frac{1}{2}, \end{cases} \tag{1.14}$$

$$y_m = \arg\min_{i\in S} \mathbb{E}_T[(\mathbb{1}_{(\hat{f}(x)\neq i)}] = \begin{cases} 1 & ,\mathbb{P}_T(\hat{f}(x)=1) \geq \frac{1}{2} \\ 0 & ,\mathbb{P}_T(\hat{f}(x)=1) < \frac{1}{2}, \end{cases} \tag{1.15}$$

$$B(x) = \mathbb{1}_{(y_*\neq y_m)} = \begin{cases} 1 & ,\mathbb{P}_T(\hat{f}(x)\neq y_*) \geq \frac{1}{2} \\ 0 & ,\mathbb{P}_T(\hat{f}(x)\neq y_*) < \frac{1}{2}, \end{cases} \tag{1.16}$$

$$V(x) = \mathbb{E}_T[\mathbb{1}_{(\hat{f}(x)\neq y_m)}] = \mathbb{P}_T(\hat{f}(x)\neq y_m) \tag{1.17}$$

$$N(x) = \mathbb{E}_x[\mathbb{1}_{(Y\neq y_*)}] = \mathbb{P}_x(Y\neq y_*) \tag{1.18}$$

Teorem 1.2.2. *Neka je $|S| = 2$. Za 0-1 funkciju gubitka, rastav* (1.7) *vrijedi za $c_1 = 2\mathbb{P}_T(\hat{f}(x) = y_*) - 1$, te $c_2 = 1$ ako je $y_* = y_m$, $c_2 = -1$ inače.*

Važno je primjetiti da u ovom slučaju c_1 i c_2 mogu poprimiti negativne vrijednosti. S time se odražava činjenica da varijanca smanjuje grešku u pristranome modelu, te da je u podacima s puno šuma ne-optimalna predikcija točna. Obje te tvrdnje vrijede zbog postojanja samo dvije klase u S, pa ne pripadnost predikcije jednoj klasi znači pripadnost drugoj. S porastom varijance raste vjerojatnost da predikcija modela ne će biti jednaka glavnoj, pa će u tim slučajevima u pristranom modelu predikcija biti jednaka optimalnoj. Isto tako,

veći šum povećava vjerojatnost da stvarna vrijednost odziva ne će biti jednaka optimalnoj predikciji, pa samim time raste točnost ne-optimalnih predikcija.

Dokaz. Prvo pokazujemo da za proizvoljan $\tau \in T$ vrijedi

$$L_x(\hat{f}) = \mathbb{E}_x[L(Y, \hat{f}(x))] = L(y_*, \hat{f}(x)) + c_0 \mathbb{E}_x[L(Y, y_*)] \tag{1.19}$$

za

$$c_0 = \begin{cases} 1 & , \hat{f}(x) = y_* \\ -1 & , \hat{f}(x) \neq y_*. \end{cases}$$

Jednakost trivijalno vrijedi kada je $\hat{f}(x) = y_*$, te preostaje provjeriti slučaj kada $\hat{f}(x) \neq y_*$.

$$\begin{aligned} \mathbb{E}_x[L(Y, \hat{f}(x))] &= \mathbb{P}_x(Y \neq \hat{f}(x)) \\ &= 1 - \mathbb{P}_x(Y = \hat{f}(x)) \\ &= 1 - \mathbb{P}_x(Y \neq y_*) \\ &= L(y_*, \hat{f}(x)) - \mathbb{E}_x[L(Y, y_*)], \end{aligned}$$

gdje treća jednakost slijedi iz činjenice da imamo samo dvije klase, pa je $y = \hat{f}(x)$ ekvivalentno s $y \neq y_*$.
Analogno pokazujemo i

$$\mathbb{E}_T[L(y_*, \hat{f}(x))] = L(y_*, y_m) + c_2 \mathbb{E}_T[L(y_m, \hat{f}(x))] \tag{1.20}$$

za

$$c_2 = \begin{cases} 1 & , y_m = y_* \\ -1 & , y_m \neq y_*. \end{cases}$$

Ponovno, jednakost trivijalno vrijedi za $y_m = y_*$, a za slučaj $y_m \neq y_*$ imamo

$$\begin{aligned} \mathbb{E}_T[L(y_*, \hat{f}(x))] &= \mathbb{P}_T(y_* \neq \hat{f}(x)) \\ &= 1 - \mathbb{P}_T(y_* = \hat{f}(x)) \\ &= 1 - \mathbb{P}_T(y_m \neq \hat{f}(x)) \\ &= L(y_*, y_m) - \mathbb{E}_T[L(y_m, \hat{f}(x))]. \end{aligned}$$

Koristeći (1.19) dobivamo

$$\begin{aligned}\mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + \mathbb{E}_T[c_0\mathbb{E}_x[L(Y, y_*)]] \\ &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + \mathbb{E}_T[c_0]\mathbb{E}_x[L(Y, y_*)] \\ &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + c_1\mathbb{E}_x[L(Y, y_*)] \\ &\overset{(1.20)}{=} (1.7),\end{aligned}$$

gdje druga jednakost vrijedi jer je $\mathbb{E}_x[L(Y, y_*)]$ konstanta, a treća zbog

$$\mathbb{E}_T[c_0] = \mathbb{P}_T(\hat{f}(x) = y_*) - \mathbb{P}_T(\hat{f}(x) \neq y_*) = 2\mathbb{P}_T(\hat{f}(x) = y_*) - 1 = c_1.$$

□

U gornjem dokazu vrijednost c_2 možemo raspisati kao

$$c_2 = 1 - 2L(y_*, y_m) = 1 - 2B(x),$$

pa očekivana ukupna testna greška za model postaje

$$\mathbb{E}_T[L(\hat{f})] = \mathbb{E}[c_1 N(X)] + \mathbb{E}[B(X)] + \mathbb{E}[(1 - 2B(X))V(X)].$$

Drugim riječima, očekivana ukupna testna greška za 0-1 funkciju gubitka u problemima s dvije klase je suma šuma, prosjeka pristranosti i izraza $(1 - 2B(X))V(X)$, kojeg nazivamo *neto varijanca*.
Ista dekompozicija, uz odgovarajuće vrijednosti c_1 i c_2 vrijedi i u slučajevima kada postoji više od dvije klase.

Teorem 1.2.3. *Neka je S konačan skup takav da je $|S| \geq 2$. Za 0-1 funkciju gubitka, rastav* (1.7) *vrijedi za $c_1 = \mathbb{P}_T(\hat{f}(x) = y_*) - \mathbb{P}_T(\hat{f}(x) \neq y_*)\mathbb{P}_{T,x}(\hat{f}(x) = Y | y_* \neq Y, \hat{f}(x) \neq y_*)$, te $c_2 = 1$ ako je $y_m = y_*$, $c_2 = -\mathbb{P}_T(\hat{f}(x) = y_* | \hat{f}(x) \neq y_m)$ inače.*

Dokaz. Postupak je isti kao za teorem 1.2.2, s ključnom razlikom u činjenici da zbog postojanja više od dvije klase, $\hat{f}(x) \neq y_*$ i $y = \hat{f}(x)$ ne znači nužno da vrijedi $y = y_*$. Isto tako, $y_m \neq y_*$ i $y = y_*$ ne povlači $y = y_m$.
Kao i prije, jednakosti (1.19) i (1.20) trivijalno vrijede kada je $\hat{f}(x) = y_*$, odnosno $y_m = y_*$, za $c_0 = 1$, odnosno $c_2 = 1$.
Pretpostavimo da vrijedi $\hat{f}(x) \neq y_*$. Tada slijedi

$$\begin{aligned}\mathbb{P}_x(Y = \hat{f}(x)) &= \mathbb{P}_x(Y = \hat{f}(x)|Y \neq y_*)\mathbb{P}_x(Y \neq y_*) + \mathbb{P}_x(Y = \hat{f}(x)|Y = y_*)\mathbb{P}_x(Y = y_*) \\ &= \mathbb{P}_x(Y = \hat{f}(x)|Y \neq y_*)\mathbb{P}_x(Y \neq y_*). \end{aligned} \tag{1.21}$$

Sada imamo za proizvoljan $\tau \in T$

$$\begin{aligned}L_x(\hat{f}) = \mathbb{E}_x[L(Y, \hat{f}(x))] &= 1 - \mathbb{P}_x(Y = \hat{f}(x)) \\ &\overset{(1.21)}{=} 1 - \mathbb{P}_x(Y = \hat{f}(x)|Y \neq y_*)\mathbb{P}_x(Y \neq y_*) \\ &= (1.19),\end{aligned}$$

za $c_0 = -\mathbb{P}_x(Y = \hat{f}(x)|Y \neq y_*)$.
Isto tako, ako pretpostavimo da vrijedi $y_m \neq y_*$, imamo

$$\begin{aligned}\mathbb{P}_T(y_* = \hat{f}(x)) &= \mathbb{P}_T(y_* = \hat{f}(x)|y_m \neq \hat{f}(x))\mathbb{P}_T(y_m \neq \hat{f}(x)) \\ &\quad + \mathbb{P}_T(y_* = \hat{f}(x)|y_m = \hat{f}(x))\mathbb{P}_T(y_m = \hat{f}(x)) \\ &= \mathbb{P}_T(y_* = \hat{f}(x)|y_m \neq \hat{f}(x))\mathbb{P}_T(y_m \neq \hat{f}(x)),\end{aligned} \tag{1.22}$$

što povlači

$$\begin{aligned}\mathbb{E}_T[L(y_*, \hat{f}(x))] &= 1 - \mathbb{P}_T(y_* = \hat{f}(x)) \\ &= 1 - \mathbb{P}_T(y_* = \hat{f}(x)|y_m \neq \hat{f}(x))\mathbb{P}_T(y_m \neq \hat{f}(x)) \\ &= (1.20)\end{aligned}$$

za $c_2 = -\mathbb{P}_T(y_* = \hat{f}(x)|y_m \neq \hat{f}(x))$.
Tvrdnja slijedi na isti način kao u teoremu 1.2.2, sa razlikom da ovaj put vrijedi

$$\mathbb{E}_T[c_0] = \mathbb{P}_T(\hat{f}(x) = y_*) - \mathbb{P}_T(\hat{f}(x) \neq y_*)\mathbb{P}_{T,x}(Y = \hat{f}(x)|Y \neq y_*, \hat{f}(x) \neq y_*) = c_1.$$

□

Rastav za problem sa dvije klase je samo poseban slučaj teorema 1.2.3. Naime, tada vrijedi $\mathbb{P}_{T,x}(\hat{f}(x) = Y|y_* \neq Y, \hat{f}(x) \neq y_*) = 1$, te $\mathbb{P}_T(\hat{f}(x) = y_*|\hat{f}(x) \neq y_m) = 1$, pa se c_1 i c_2 svedu na vrijednosti iz teorema 1.2.2.
Po rastavu je vidljivo da, u slučajevima kada imamo više od dvije klase, varijanca ne doprinosi uvijek smanjenju greške u pristranim primjerima, budući da $\hat{f}(x) \neq y_m$ više ne povlači

nužno $\hat{f}(x) = y_*$. To znači da tolerancija na varijancu modela opada s brojem klasa, iz čega slijedi zaključak da su u slučaju 0-1 funkcije gubitka modeli s visokom varijancom prikladniji u slučajevima s manje klasa.
Postoje i slučajevi u kojima 0-1 funkcija gubitka nije dostatna, budući da nisu nužno sve greške jednake važnosti. Tipičan primjer toga je klasificiranje pacijenata sa rakom, gdje je puno veća greška osobu sa rakom klasificirati kao zdravu, nego obrnuto. U takvim slučajevima koriste se modificirane verzije 0-1 funkcije greške.

Teorem 1.2.4. *Neka je* $|S| = 2$. *Za funkcije gubitka za koje vrijedi* $L(y, y) = 0, \forall y \in S$, *te* $L(y_1, y_2) > 0, \forall y_1, y_2$ *takve da* $y_1 \neq y_2$, *rastav 1.7 vrijedi za* $c_1 = \mathbb{P}_T(\hat{f}(x) = y_*) - \frac{L(y_*,z)}{L(z,y_*)}\mathbb{P}_T(\hat{f}(x) \neq y_*), z \in S, z \neq y_*$, *te* $c_2 = 1$ *kada* $y_* = y_m$, $c_2 = -\frac{L(y_*,y_m)}{L(y_m,y_*)}$ *inače.*

Dokaz. Neka je (x, y) realizacija od (X, Y). Prvo pokazujemo da za proizvoljan $\tau \in T$ vrijedi

$$L(y, \hat{f}(x)) = L(y_*, \hat{f}(x)) + c_0 L(y, y_*), \tag{1.23}$$

za

$$c_0 = \begin{cases} 1 & , \hat{f}(x) = y_* \\ -\frac{L(y_*, \hat{f}(x))}{L(\hat{f}(x), y_*)} & , \hat{f}(x) \neq y_*. \end{cases}$$

Jednakost (1.23) vrijedi trivijalno kada je $\hat{f}(x) = y_*$, te preostaje provjeriti drugi slučaj. Budući da postoje samo dvije klase, $\hat{f}(x) \neq y_*$ povlači ili $\hat{f}(x) = y$, ili $y_* = y$. Ako je $\hat{f}(x) = y$, (1.23) se svede na

$$0 = L(y_*, \hat{f}(x)) - \frac{L(y_*, \hat{f}(x))}{L(\hat{f}(x), y_*)} L(y, y_*) = L(y_*, \hat{f}(x)) - L(y_*, \hat{f}(x)) = 0,$$

a ako $y_* = y$, imamo

$$L(y, \hat{f}(x)) = L(y, \hat{f}(x)) + 0.$$

Analogno pokazujemo da vrijedi i

$$L(y_*, \hat{f}(x)) = L(y_*, y_m) + c_2 L(y_m, \hat{f}(x)), \tag{1.24}$$

za

$$c_2 = \begin{cases} 1 & , y_m = y_* \\ -\frac{L(y_*, y_m)}{L(y_m, y_*)} & , y_m \neq y_*. \end{cases}$$

Ponovno, jednakost trivijalno vrijedi kada $y_m = y_*$, te preostaje provjeriti drugi slučaj. Za slučaj kada je $\hat{f}(x) = y_m$ imamo

$$L(y_*, y_m) = L(y_*, y_m) + 0,$$

a za $\hat{f}(x) = y_*$

$$0 = L(y_*, y_m) - \frac{L(y_*, y_m)}{L(y_m, y_*)} L(y_m, y_*) = L(y_*, y_m) - L(y_*, y_m) = 0.$$

Sada slijedi

$$\begin{aligned} \mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[\mathbb{E}_x[L(Y, \hat{f}(x))]] \\ &\overset{(1.23)}{=} \mathbb{E}_T[\mathbb{E}_x[L(y_*, \hat{f}(x)) + c_0 L(Y, y_*)]] \\ &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + \mathbb{E}_T[c_0]\mathbb{E}_x[L(Y, y_*)] \\ &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + c_1 \mathbb{E}_x[L(Y, y_*)] \\ &\overset{(1.24)}{=} (1.7), \end{aligned}$$

gdje treća jednakost vrijedi zbog činjenice da je c_0 konstanta za dani x, te zbog neovisnosti $\mathbb{E}_x[L(Y, y_*)]$ o T, a četvrta zbog

$$\mathbb{E}_T[c_0] = \mathbb{P}_T(\hat{f}(x) = y_*) - \mathbb{P}_T(\hat{f}(x) \neq y_*) \frac{L(y_*, z)}{L(z, y_*)} = c_1.$$

□

Po teoremu 1.2.4, slijedi da varijanca utječe više, odnosno manje, kako raste asimetrija u cijeni grešaka, ovisno o tome koji je tip greške skuplji.
Otvoreno je pitanje postoji li rastav (1.7) za asimetrične modifikacije 0-1 funkcije gubitka kada je $|S| > 2$.
Drugi način na koji se 0-1 funkcija gubitka može modificirati je dodavanjem troška za točne predikcije, odnosno postavljanjem $L(y, y) = c$, za neki $c \neq 0$. U tom slučaju dekompozicija greške sadrži dodatan član.

Primjer 1.2.5. *(Dekompozicija 0-1 greške s troškom za ispravne predikcije) Neka je $|S| = 2$ i L funkcija gubitka takva da za svaki $y_1, y_2 \in S$ vrijedi*

$$L(y_1, y_2) = \begin{cases} c & , y_1 = y_2 \\ 1 & , y_1 \neq y_2 \end{cases},$$

za neki skalar $c \geq 0$. Tada je, za datu realizaciju (x, y) od (X, Y) i $\tau in T$

$$L(y, \hat{f}(x)) = L(y_*, \hat{f}(x)) + L(y, y_*) + C_1, \tag{1.25}$$

gdje je $C_1 = \mathbb{1}_{(y_* \neq \hat{f}(x), \hat{f}(x) = y)}(2c - 2) - c$
Naime, u slučaju da je $\hat{f}(x) = y_*$, jednakost se svede na

$$L(y, y_*) = c + L(y, y_*) - c,$$

dok u suprotnom slučaju, kada je $\hat{f}(x) = y$ imamo

$$c = 1 + 1 - c + 2c - 2$$

a kada je $y_* = y$

$$L(y, \hat{f}(x)) = L(y, \hat{f}(x)) + c - c$$

.
Analogno, vrijedi i

$$L(y_*, \hat{f}(x)) = L(y_*, y_m) + L(y_m, \hat{f}(x)) + C_2 \tag{1.26}$$

za $C_2 = \mathbb{1}_{(y_m \neq y_*, \hat{f}(x) = y_*)}(2c - 2) - c$.
Sada se izraz očekivane testne greške od L može zapisati kao

$$\begin{aligned} \mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[\mathbb{E}_x[L(Y, \hat{f}(x))]] \\ &\overset{(1.25)}{=} \mathbb{E}_T[L(y_*, \hat{f}(x)) + \mathbb{E}_x[L(Y, y_*)] + \mathbb{E}_x[C_1]] \\ &= \mathbb{E}_T[L(y_*, \hat{f}(x))] + \mathbb{E}_x[L(Y, y_*)] + \mathbb{E}_{T,x}[C_1] \\ &\overset{(1.26)}{=} L(y_*, y_m) + \mathbb{E}_T[L(y_m, \hat{f}(x))] + \mathbb{E}_T[C_2] + \mathbb{E}_x[L(Y, y_*)] + \mathbb{E}_T[\mathbb{E}_x[C_1]] \\ &= N(x) + B(x) + V(x) + C(x), \end{aligned} \tag{1.27}$$

za

$$
\begin{aligned}
C(x) &= \mathbb{E}_T[\mathbb{E}_x[C_1]] + \mathbb{E}_T[C_2] \\
&= \mathbb{E}_T[\mathbb{1}_{(\hat{f}(x)\neq y_*)}\mathbb{P}_x(\hat{f}(x) = Y)](2c-2) - c \\
&+ \mathbb{1}_{(y_m\neq y_*)}\mathbb{P}_T(\hat{f}(x) = y_*)(2c-2) - c \qquad (1.28) \\
&= (\mathbb{P}_T(\hat{f}(x) \neq y_*)\mathbb{P}_x(y_* \neq Y) + \mathbb{1}_{(y_m\neq y_*)}\mathbb{P}_T(\hat{f}(x) \neq y_m))(2c-2) - 2c \\
&= (\mathbb{P}_T(\hat{f}(x) \neq y_*)(N(x) - c\mathbb{P}_x(y_* = Y)) + \mathbb{1}_{(y_m\neq y_*)}(V(x) - c\mathbb{P}_T(\hat{f}(x) = y_m))(2c-2) - 2c \\
&= N(x)(\mathbb{P}_T(\hat{f}(x) \neq y_*)(2c-2) + V(x)\mathbb{1}_{(y_m\neq y_*)}(2c-2) \\
&- c((\mathbb{P}_T(\hat{f}(x) \neq y_*)\mathbb{P}_x(y_* = Y) + \mathbb{1}_{(y_m\neq y_*)}\mathbb{P}_T(\hat{f}(x) = y_m))(2c-2) - 2). \qquad (1.29)
\end{aligned}
$$

Uvrštavajući (1.28) *u* (1.27)*, dobivamo konačnu dekompoziciju*

$$\mathbb{E}_T[L_x(\hat{f})] = c_1 N(x) + B(x) + c_2 V(x) + cO(x), \tag{1.30}$$

gdje je $c_1 = 1 + (\mathbb{P}_T(\hat{f}(x) \neq y_*)(2c-2)$*,* $c_2 = 1 + \mathbb{1}_{(y_m\neq y_*)}(2c-2)$ *i* $O(x) = ((\mathbb{P}_T(\hat{f}(x) \neq y_*)\mathbb{P}_x(y_* = Y) + \mathbb{1}_{(y_m\neq y_*)}\mathbb{P}_T(\hat{f}(x) = y_m))(2c-2) - 2$*.*
U slučajevima kada je $c = 0$*, dekompozicija* (1.30) *je, očekivano, jednaka onoj dobivenoj u teoremu 1.2.2. Kada je* $c > 0$*, dekompozicija ima dodatan član koji kompenzira trošak ispravne predikcije.*

1.2.3 Bregmanove divergencije

Bregmanove divergencije su skupine funkcija koje mjere razliku između dvije točke definiranu kroz pojmove strogo konveksne funkcije. U [4] je pokazano da za takve funkcije postoji dekompozicija pristranosti i varijance.

Definicija 1.2.6. *Neka je* S *interval u* $\mathbb{R}$*. Funkcija gubitka* L *je* Bregmanova divergencija *ako postoji strogo konveksna funckija* $F : S \to \mathbb{R}$ *takva da za svaki* $y_1, y_2 \in S$ *vrijedi*

$$L(y_1, y_2) = F(y_1) - F(y_2) - F'(y_2)(y_1 - y_2). \tag{1.31}$$

Klasičan primjer Bregmanove divergencije je sama srednjekvadratna funkcija gubitka. Naime, za $F(x) = x^2$ imamo za dani $y_1, y_2 \in S$

$$\begin{aligned} F(y_1) - F(y_2) - F'(y_2)(y_1 - y_2) &= y_1^2 - y_2^2 - 2y_2(y_1 - y_2) \\ &= y_1^2 - y_2^2 - 2y_1y_2 + 2y_2^2 \\ &= (y_1 - y_2)^2 \\ &= L(y_1, y_2). \end{aligned}$$

Lema 1.2.7. *Ako je funkcija gubitka L Bregmanova divergencija strogo konveksne funkcije F, tada za dani $x \in \mathbb{R}^p$ vrijede sljedeće jednakosti:*

- $y_* = \mathbb{E}_x[Y]$ (1.32)
- $F'(y_m) = \mathbb{E}_T[F'(\hat{f}(x))]$ (1.33)

Dokaz. Za dokazivanje prve jednakosti promatramo derivaciju funkcije $z \to \mathbb{E}_x[L(Y, z)]$, kako bi pronašli njenu točku minimuma.

$$\begin{aligned} \frac{\partial}{\partial z}\mathbb{E}_x[L(Y,z)] &= \frac{\partial}{\partial z}\mathbb{E}_x[F(Y) - F(z) - F'(z)(Y - z)] \\ &= \frac{\partial}{\partial z}(\mathbb{E}_x[F(Y)] - F(z) - F'(z)(\mathbb{E}_x[Y] - z)) \\ &= -F'(z) - F''(z)(\mathbb{E}_x[Y] - z) + F'(z) \\ &= -F''(z)(\mathbb{E}_x[Y] - z). \end{aligned}$$

Zbog svojstva stroge konveksnosti je $F''(z) > 0$, pa je jedina točka minimuma funkcije u $z = \mathbb{E}_x[Y]$. Tvrdnja slijedi iz definicije y_*.
Sličnim postupkom dokazujemo i drugu jednakost. Sada promatramo derivaciju funkcije $z \to \mathbb{E}_T[L(z, \hat{f}(x))]$.

$$\begin{aligned} \frac{\partial}{\partial z}\mathbb{E}_T[L(z,\hat{f}(x))] &= \frac{\partial}{\partial z}\mathbb{E}_T[F(z) - F(\hat{f}(x)) - F'(\hat{f}(x))(z - \hat{f}(x))] \\ &= F'(z) - \mathbb{E}_T[F'(\hat{f}(x))], \end{aligned}$$

iz čega zbog stroge konveksnosti od F i definicije od y_m slijedi tvrdnja. □

Teorem 1.2.8. *Ako je funkcija gubitka L Bregmanova divergencija za neku strogo konveksnu funkciju F, tada rastav* (1.7) *vrijedi za $c_1 = c_2 = 1$.*

Dokaz. Neka je $x \in \mathbb{R}^p$ i $\tau \in T$. Tada

$$\begin{aligned} L_x(\hat{f}) = \mathbb{E}_x[L(Y,\hat{f}(x))] &= \mathbb{E}_x[F(Y) - F(\hat{f}(x)) - F'(\hat{f}(x))(Y-\hat{f}(x))] \\ &= \mathbb{E}_x[F(Y)] - F(\hat{f}(x)) - F'(\hat{f}(x))(\mathbb{E}_x[Y] - \hat{f}(x)) \\ &\overset{(1.32)}{=} \mathbb{E}_x[F(Y)] - F(y_*) + F(y_*) - F(\hat{f}(x)) \\ &- F'(y_*)(\mathbb{E}_x[Y] - y_*) - F'(\hat{f}(x))(y_* - \hat{f}(x)) \\ &= \mathbb{E}_x[L(Y,y_*)] + L(y_*,\hat{f}(x)), \end{aligned} \tag{1.34}$$

te

$$\begin{aligned} \mathbb{E}_T[L(y_*,\hat{f}(x))] &= \mathbb{E}_T[F(y_*) - F(\hat{f}(x)) - F'(\hat{f}(x))(y_* - \hat{f}(x))] \\ &= F(y_*) - \mathbb{E}_T[F(\hat{f}(x))] - \mathbb{E}_T[F'(\hat{f}(x))(y_* - \hat{f}(x))] \\ &= F(y_*) - F(y_m) + F(y_m) - \mathbb{E}_T[F(\hat{f}(x))] \\ &- \mathbb{E}_T[F'(\hat{f}(x))](y_* - y_m) - \mathbb{E}_T[F'(\hat{f}(x))(y_m - \hat{f}(x))] \\ &\overset{(1.33)}{=} F(y_*) - F(y_m) + F(y_m) - \mathbb{E}_T[F(\hat{f}(x))] \\ &- F'(y_m)(y_* - y_m) - \mathbb{E}_T[F'(\hat{f}(x))(y_m - \hat{f}(x))] \\ &= L(y_*,y_m) + \mathbb{E}_T[L(y_m,\hat{f}(x))]. \end{aligned} \tag{1.35}$$

Tvrdnja teorema sada slijedi iz

$$\mathbb{E}_T[L_x(\hat{f})] = \mathbb{E}_T[\mathbb{E}_x[L(Y,\hat{f}(x))]] \overset{(1.34)}{=} \mathbb{E}_T[\mathbb{E}_x[L(Y,y_*)] + L(y_*,\hat{f}(x))] \overset{(1.35)}{=} (1.7)$$

za $c_1 = c_2 = 1$. □

U problemima binarne klasifikacije, odnosno klasifikacije kada je $Y \in \{0,1\}$, često se koriste metode čiji modeli $\hat{p} = \hat{p}(\tau)$ procjenjuju vjerojatnosti $p(x) := \mathbb{P}(Y = 1|X = x)$. Tada, iako je Y binarna varijabla, $\hat{p}(x)$ poprima vrijednosti na $[0,1]$, pa 0-1 funkcija gubitka i njene razne modifikacije više nisu dostatne. U tom slučaju se često koristi funkcija log-gubitka $L : \{0,1\} \times [0,1] \to [0,+\infty)$ definirana kao

$$L(y,\hat{p}(x)) = \begin{cases} -\log(\hat{p}(x)) & , y = 1 \\ -\log(1-\hat{p}(x)) & , y = 0 \end{cases}, \tag{1.36}$$

ili alternativno $L(y,\hat{p}(x)) = -y\log(\hat{p}(x)) - (1-y)\log(1-\hat{p}(x))$. Tu funkciju možemo proširiti na $L : [0,1]^2 \to [0,+\infty)$ na sljedeći način:

$$L(p_1, p_2) = p_1 \log(\frac{p_1}{p_2}) + (1 - p_1) \log(\frac{1 - p_1}{1 - p_2}). \tag{1.37}$$

Ako $p_1 \neq p_2$ i $p_2 \in \{0, 1\}$, vrijednost L je jednaka $+\infty$. S druge strane, ako je $p_1 = 0$ ili $1 - p_1 = 0$, pripadni član je jednak 0 zbog $\lim_{x \to 0_+} x \log(x) = 0$.

Napomena 1.2.9. *Funkcija definirana u* (1.37) *je u literaturi poznata kao KL divergencija, te se koristi za mjerenje statističke udaljenosti dviju vjerojatnosnih distribucija.*

Primjer 1.2.10. *(log-gubitak kao Bregmanova divergencija)*
Neka je $S = [0, 1]$ *i* $F : S \to \mathbb{R}$ *jednaka* $F(p) = -H(p)$*, gdje je* H *funkcija* entropije *definirana kao*

$$H(p) = -p \log(p) - (1 - p) \log(1 - p). \tag{1.38}$$

Ako funkciju log-gubitka L *proširimo kao u* (1.37)*, ona je tada Bregmanova divergencija od* F*. Naime, vrijedi*

$$\begin{aligned} F(p) &= p \log(p) + (1 - p) \log(1 - p) \\ F'(p) &= \log(p) + 1 - \log(1 - p) - 1 = \log(p) - \log(1 - p) \\ F''(p) &= \frac{1}{p} + \frac{1}{1 - p} > 0, \end{aligned}$$

pa je F *strogo konveksna. Preostaje još pokazati da vrijedi jednakost* (1.31).

$$\begin{aligned} & F(p_1) - F(p_2) - F'(p_2)(p_1 - p_2) \\ &= p_1 \log(p_1) + (1 - p_1) \log(1 - p_1) - p_2 \log(p_2) \\ &+ (1 - p_2) \log(1 - p_2) - (\log(p_2) - \log(1 - p_2))(p_1 - p_2) \\ &= p_1(\log(p_1) - \log(p_2)) + (1 - p_1)(\log(1 - p_1) - \log(1 - p_2)) \\ &= L(p_1, p_2). \end{aligned}$$

Budući da se proširena funkcija log-gubitka može izraziti kao Bregmanova divergencija, po teoremu 1.2.8 za nju vrijedi rastav (1.7) za $c_1 = c_2 = 1$. Također, po lemi 1.2.7 znamo da je $y_* = \mathbb{E}_x[Y]$, dok za glavnu predikciju vrijedi

$$\log(y_m) - \log(1 - y_m) = \mathbb{E}_T[\log(\hat{p}(x)) - \log(1 - \hat{p}(x))]. \tag{1.39}$$

Sređivanjem te jednakosti dobije se da je

$$y_m = \frac{\exp(c(x))}{1 + \exp(c(x))} \tag{1.40}$$

za $c(x) = \mathbb{E}_T[\log(\frac{\hat{p}(x)}{1-\hat{p}(x)})]$. Vrijedi i

$$B(x) = y_* \log(\frac{y_*}{y_m}) + (1 - y_*) \log(\frac{1 - y_*}{1 - y_m}), \tag{1.41}$$

$$\begin{aligned}
V(x) &= \mathbb{E}_T[y_m \log(\frac{y_m}{\hat{p}(x)}) + (1 - y_m) \log(\frac{1 - y_m}{1 - \hat{p}(x)})] \\
&= y_m(\log(y_m) - \mathbb{E}_T[\log(\hat{p}(x))]) + (1 - y_m)(\log(1 - y_m) - \mathbb{E}_T[\log(1 - \hat{p}(x))]) \\
&= y_m(\log(y_m) - \log(1 - y_m) - \mathbb{E}_T[\log(\hat{p}(x)) - \log(1 - \hat{p}(x))]) \\
&+ \log(1 - y_m) - \mathbb{E}_T[\log(1 - \hat{p}(x))] \\
&\overset{(1.39)}{=} \log(1 - y_m) - \mathbb{E}_T[\log(1 - \hat{p}(x))] \\
&\overset{(1.39)}{=} \log(y_m) - \mathbb{E}_T[\log(\hat{p}(x))],
\end{aligned} \tag{1.42}$$

$$\begin{aligned}
N(x) &= \mathbb{E}_x[Y \log(\frac{Y}{y_*}) + (1 - Y) \log(\frac{1 - Y}{1 - y_*})] \\
&= \mathbb{E}_x[Y \log(Y)] - \mathbb{E}_x[Y] \log(y_*) + \mathbb{E}_x[\log(1 - Y)] \\
&- \log(1 - y_*) - \mathbb{E}_x[Y \log(1 - Y)] + \mathbb{E}_x[Y] \log(1 - y_*) \\
&= \mathbb{E}_x[Y \log(Y) + (1 - Y) \log(1 - Y)] - y_* \log(y_*) - (1 - y_*) \log(1 - y_*) \\
&= \mathbb{E}_x[F(Y)] - F(y_*) \\
&= H(y_*) - \mathbb{E}_x[H(Y)].
\end{aligned} \tag{1.43}$$

Napomena 1.2.11. *Drugi naziv za funkciju log-gubitka je* gubitak unakrsne entropije*, te bi bilo prirodnije proširiti ju upravo na funkciju unakrsne entropije, definiranu kao* $H : [0, 1]^2 \to [0, +\infty)$

$$H(p, q) = -p \log(q) - (1 - p) \log(1 - q).$$

Ta se funkcija može iskazati i kao

$$H(p, q) = H(p) + L(p, q),$$

gdje je L označava KL divergenciju. $H(y) = 0$ *kada je* $y \in \{0, 1\}$*, pa su za svrhu mjerenja testne greške ta dva proširenja ekvivalentna. Vrijednost optimalne predikcije bi ostala ista kao i za KL divergenciju, budući da*

$$y_* = \arg\min_{i \in S} \mathbb{E}_x[H(Y, i)] = \arg\min_{i \in S} \mathbb{E}_x[H(Y) + L(Y, i)] = \arg\min_{i \in S} \mathbb{E}_x[L(Y, i)],$$

pa bi, ponovo zbog $H(Y) = 0$*, i vrijednost šuma ostala ista. Ipak, problem nastaje u glavnoj predikciji, iz razloga što*

$$y_m = \arg\min_{i \in S} \mathbb{E}_T[H(i, \hat{p}(x))] = \arg\min_{i \in S} \mathbb{E}_T[H(i) + L(i, \hat{p}(x))]$$

ne mora nužno biti jednako $\arg\min_{i \in S} \mathbb{E}_T[L(i, \hat{p}(x))]$*. Samim time, ni vrijednosti pristranosti i varijance ne moraju nužno biti jednake, te je upitno da li dekompozicija* (1.7) *vrijedi za log-gubitak proširen kao unakrsna entropija.*

1.3 Svojstva univerzalne dekompozicije

1.3.1 Utjecaj pristranosti na bagging

Za probleme klasifikacije često se koriste metode bazirane na stablima odluke, koje funkcioniraju na način da se set podataka dijeli na sve manje podskupove, te se na svakom od tih podskupova prilagođava jednostavan model. Iako je rezultate takvih modela lako interpretirati, prediktivna moć im je u praksi manjkava, pa se u svrhu poboljšanja iste koriste tzv. *bagging* metode, koje generiraju više stabala odluke, te agregiraju njihove rezultate.
Radi boljeg razumijevanja uspješnosti tih metoda s 0-1 funkcijom greške, uvodi se pojam *ispravne poredanosti*.

Definicija 1.3.1. *U problemu klasifikacije, za dati* $x \in \mathbb{R}^p$ *metoda je* ispravno poredana *ako* $\forall y \in S, y \neq y_*$ *vrijedi* $\mathbb{P}_T(\hat{f}(x) = y) < \mathbb{P}_T(\hat{f}(x) = y_*)$

Budući da je glavna predikcija metode jednaka $\arg\max_{k \in S} \mathbb{P}_T(\hat{f}(x) = k)$ za 0-1 funkciju gubitka, vrijedi sljedeći teorem:

Teorem 1.3.2. *Za dati* $x \in \mathbb{R}^p$*, metoda je ispravno poredana akko je* $B(x) = 0$ *za 0-1 funkciju gubitka.*

u [1] je pokazano da bagging metoda pretvara ispravno poredanu metodu u skoro pa optimalnu, što sada možemo opravdati dosad pokazanim rezultatima. Naime, po teoremu 1.3.2 vrijedi da su ispravno poredane metode nepristrane, a po teoremu 1.2.2 znamo da će smanjenje varijance imati najviše utjecaja na smanjenje testne greške upravo u nepristranim metodama.

1.3.2 Testna greška metrika

Dekompozicija (1.7) i dalje nije primjenjiva za bilo koju funkciju gubitka, te nije definirano koji su nužni uvjeti za istu. Ipak, ako funkcija zadovoljava određene uvjete, moguće je njenu testnu grešku ograničiti pomoću funkcija pristranosti, varijance i šuma.

Definicija 1.3.3. *Funkcija gubitka L je metrika ako vrijede sljedeće tvrdnje:*

- $L(y_1, y_2) \geq 0, \forall y_1, y_2 \in S$ (1.44)
- $L(y, y) = 0, \forall y \in S$ (1.45)
- $L(y_1, y_2) = L(y_2, y_1), \forall y_1, y_2 \in S$ *(simetrija)* (1.46)
- $L(y_1, y_2) \leq L(y_1, y_3) + L(y_2, y_3), \forall y_1, y_2, y_3 \in S$ *(nejednakost trokuta).* (1.47)

Teorem 1.3.4. *Za danu funkciju gubitka L, ako je L metrika vrijedi*

$$\mathbb{E}_T[L_x(\hat{f})] \leq N(x) + B(x) + V(x) \tag{1.48}$$
$$\mathbb{E}_T[L_x(\hat{f})] \geq \max\{N(x) - B(x) - V(x), B(x) - N(x) - V(x), V(x) - N(x) - B(x)\}. \tag{1.49}$$

Dokaz. Koristeći nejednakost trokuta imamo

$$\begin{aligned}\mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[\mathbb{E}_x[L(Y, \hat{f}(x))]] \\ &\leq \mathbb{E}_T[\mathbb{E}_x[L(Y, y_*) + L(y_*, y_m) + L(y_m, \hat{f}(x))]] \\ &= N(x) + B(x) + V(x).\end{aligned}$$

Isto tako, koristeći nejednakost trokuta i svojstvo simetrije, za dani x

$$\begin{aligned}B(X) &= L(y_*, y_m) \\ &= \mathbb{E}_T[\mathbb{E}_x[L(y_*, y_m)]] \\ &\leq \mathbb{E}_T[\mathbb{E}_x[L(y_*, Y) + L(Y, \hat{f}(x)) + L(\hat{f}(x), y_m)]] \\ &= \mathbb{E}_T[\mathbb{E}_x[L(Y, y_*) + L(Y, \hat{f}(x)) + L(y_m, \hat{f}(x))]] \\ &= N(x) + \mathbb{E}_T[L_x(\hat{f})] + V(x),\end{aligned}$$

odnosno $\mathbb{E}_T[L_x(\hat{f})] \geq B(x) - N(x) - V(x)$. Ostale dvije komponente donje granice se dobivaju analognim postupkom, primjenjujući nejednakost trokuta na $N(x)$ i $V(x)$. □

Srednjekvadratna funkcija gubitka je primjer metrike čija je dekompozicija poznata. Ipak, činjenica da je funkcija gubitka metrika ne znači nužno da je dekompozicija testne greške te iste funkcije uopće moguća. U slučaju apsolutne funkcije gubitka, testna greška se nikako ne može iskazati u obliku (1.7).

Primjer 1.3.5. *(Rastav testne greške absolutne funkcije gubitka)*
Neka je L absolutna funkcija gubitka. Promatramo realizaciju (x, y) *od* (X, Y) *i* $\tau \in T$.

$$L(y, \hat{f}(x)) = \begin{cases} y - \hat{f}(x) & , y \geq \hat{f}(x) \\ \hat{f}(x) - y & , y < \hat{f}(x). \end{cases}$$

U prvom slučaju je $L(y, \hat{f}(x)) = y - \hat{f}(x) = y - y_* + y_* - \hat{f}(x) = c_{1,1}L(y, y_*) + c_{1,2}L(y_*, \hat{f}(x))$ *za*

$$c_{1,1} = \begin{cases} 1 & , y \geq y_* \\ -1 & , y < y_*, \end{cases}$$

$$c_{1,2} = \begin{cases} 1 & , y_* \geq \hat{f}(x) \\ -1 & , y_* < \hat{f}(x). \end{cases}$$

Za drugi slučaj imamo $L(y, \hat{f}(x)) = c_{2,1}L(y, y_*) + c_{2,2}L(y_*, \hat{f}(x))$ *za*

$$c_{2,1} = \begin{cases} 1 & , y < y_* \\ -1 & , y \geq y_*, \end{cases}$$

$$c_{1,2} = \begin{cases} 1 & , y_* < \hat{f}(x) \\ -1 & , y_* \geq \hat{f}(x), \end{cases}$$

pa oba slučaja možemo objediniti u jedan sa

$$L(y, \hat{f}(x)) = c_1 L(y, y_*) + c_2(y_*, \hat{f}(x)) \tag{1.50}$$

za

$$c_1 = \begin{cases} 1 & , (y \geq \hat{f}(x), y \geq y_*) \cup (y < \hat{f}(x), y < y_*) \\ -1 & , inače, \end{cases}$$

$$c_2 = \begin{cases} 1 & , (y \geq \hat{f}(x), y_* \geq \hat{f}(x)) \cup (y < \hat{f}(x), y_* < \hat{f}(x)) \\ -1 & , inače. \end{cases}$$

Analogno se pokaže i

$$L(y_*, \hat{f}(x)) = c_3 L(y_*, y_m) + c_4(y_m, \hat{f}(x)) \tag{1.51}$$

za

$$c_3 = \begin{cases} 1 & , (y_* \geq \hat{f}(x), y_* \geq y_m) \cup (y_* < \hat{f}(x), y_* < y_m) \\ -1 & , inače, \end{cases}$$

$$c_4 = \begin{cases} 1 & , (y_* \geq \hat{f}(x), y_m \geq \hat{f}(x)) \cup (y_* < \hat{f}(x), y_m < \hat{f}(x)) \\ -1 & , inače. \end{cases}$$

Sada je

$$\begin{aligned} \mathbb{E}_T[L_x(\hat{f})] &= \mathbb{E}_T[E_x[L(Y, \hat{f}(x))]] \\ &\overset{(1.50)}{=} \mathbb{E}_T[E_x[c_1 L(Y, y_*) + c_2 L(y_*, \hat{f}(x))]] \\ &= \mathbb{E}_{T,x}[c_1 L(Y, y_*)] + \mathbb{E}_T[E_x[c_2] L(y_*, \hat{f}(x))] \\ &\overset{(1.51)}{=} \mathbb{E}_{T,x}[c_1 L(Y, y_*)] + \mathbb{E}_T[E_x[c_2](c_3 L(y_*, y_m) + c_4 L(y_m, \hat{f}(x))] \\ &= \mathbb{E}_{T,x}[c_1 L(Y, y_*)] + E_T[E_x[c_2]c_3] B(x) + \mathbb{E}_T[E_x[c_2]c_4 L(y_m, \hat{f}(x))]. \end{aligned} \tag{1.52}$$

Razlike između (1.52) i (1.7) su očite. Osim što se sada izraz za pristranost množi sa skalarom, šum i varijanca su zamijenjeni novim, puno kompleksnijim izrazima, zbog kojih utjecaj svih triju komponenti na testnu grešku ovisi o međusobnom odnosu veličina Y, $\hat{f}(x)$, y_* i y_m.

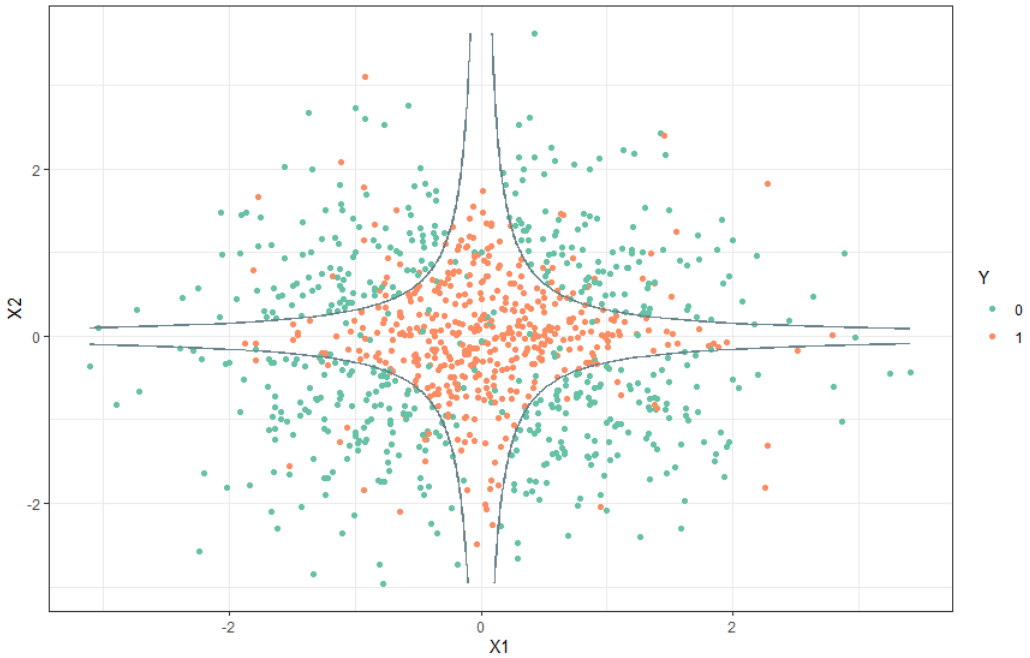
Poglavlje 2

Primjena univerzalne dekompozicije

U problemima klasifikacije često je korištena tzv. metoda k-najbližih susjeda, koja funkcionira na način da za datu obzervaciju x nađe k najbližih točaka u skupu za učenje, te za predikciju odabere najčešću klasifikaciju među tim točkama. Ona spada među najjednostavnije metode statističkog učenja, čija je glavna odlika jednostavna interpretabilnost modela, te u pojedinim slučajevima jaka prediktivna moć. Koristi se i u problemima regresije, kada se uzima prosjek vrijednosti odziva k najbližih točaka. Poznato je i jednostavno za pokazati da u slučaju srednjekvadratne greške pristranost raste s povećanjem parametra kompleksnosti k, dok varijanca pada. Na simuliranim podacima ćemo pokazati kako se taj odnos ponaša u klasifikacijskim problemima sa 0-1 funkcijom gubitka.
Podaci su generirani na sljedeći način. Dvije kovarijate X_1 i X_2 su uzorkovane nezavisno iz standardne normalne razdiobe, dok je odziv dobiven kao $Y = (1 - \epsilon)f(X_1, X_2) + \epsilon(1 - f(X_1, X_2))$, gdje je

$$f(x) = f(x_1, x_2) = \begin{cases} 1 & , |x_1 x_2| \leq \frac{\pi}{10} \\ 0 & , \text{inače,} \end{cases}$$

a ϵ je dobiven iz Bernoullijeve distribucije, za $p = 0.1$. Drugim riječima, vrijednost odziva za datu kovarijatu x će biti jednaka $f(x)$ sa vjerojatnošću 0.9
Kao što je vidljivo na slici 2.1, stvarna granica odluke između dvije klase odziva je nelinearna, s klasom 1 u sredini i klasom 0 na rubovima. Unatoč nelinearnosti, granica odluke je relativno jednostavna, pa zbog niske razine šuma očekujemo da će metoda kNN dati dobre rezultate.
Provjeravamo ponašanje kNN metode za vrijednost parametra kompleksnosti k u rasponu od 1 do 300.
Kako bi mogli procijeniti glavnu predikciju, generiramo uzorak slučajnog skupa za učenje T sa 100 skupova duljine 1000, te na svakom od tih skupova provodimo kNN metodu,



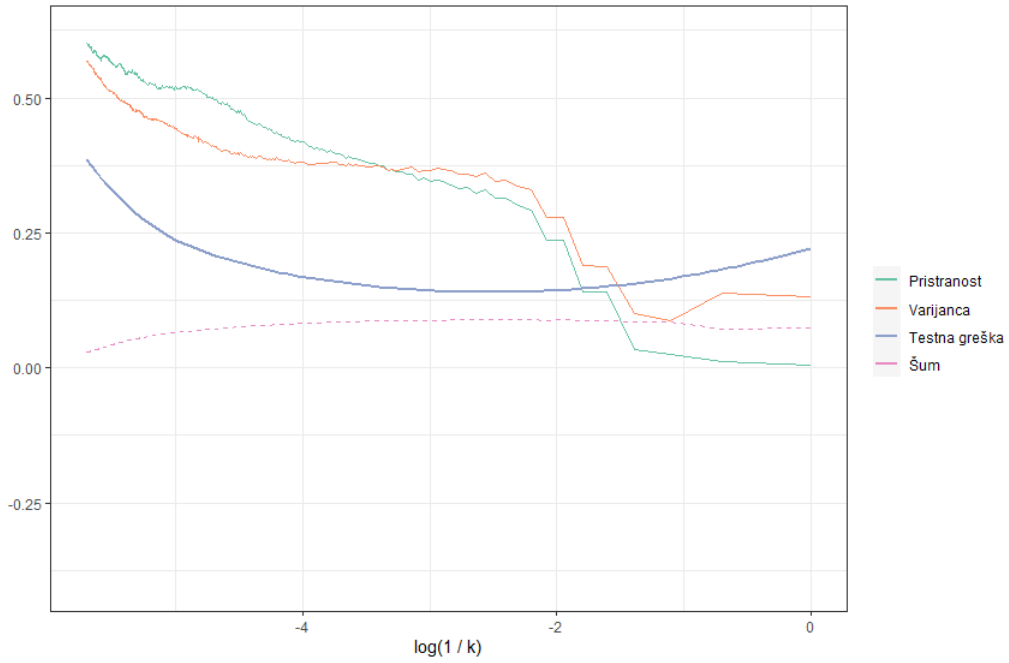
Slika 2.1: Primjer realizacije skupa za učenje

Stvarna granica odluke je ucrtana tamnoplavom bojom

koristeći poziv funkcije *knn* iz paketa *class* programskog jezika *R*. Na taj način dobivamo 100 modela naučenih na podacima dobivenima iz iste distribucije.
Za funkciju gubitka uzimamo 0-1 gubitak, pa je, u skladu s jednakostima (1.14) - (1.18), $y_* = f(x)$ i $N(x) = 0.1$. y_m za dati x iz testnog skupa tada procjenjujemo kao najčešću klasifikaciju od x među dobivenih 100 modela, pa se pristranost računa koristeći y_* i procijenjeni y_m. Varijanca u x se procjenjuje kao postotak modela čija se predikcija u x razlikuje od procijenjene glavne predikcije.
Kako bi mogli promatrati dekompoziciju testne greške, potrebno je procjeniti i vrijednosti c_1 i c_2 iz teorema 1.2.2. c_2 računamo kao $1 - 2B(x)$, a za procjenu optimalne predikcije y_* i skup 100 dobivenih modela M, c_1 procjenjujemo kao

$$\hat{c}_1 = 1 - 2\frac{\sum_{\hat{f}\in M} \mathbb{1}_{(\hat{f}(x)=y_*)}}{100}.$$

Za procjenu očekivane testne greške u x koristimo testni skup duljine 1000, tako da, koristeći vjerojatnosnu distribuciju od Y, za svaki model računamo testnu grešku kao



Slika 2.2: Odnos prosječne pristranosti i varijance

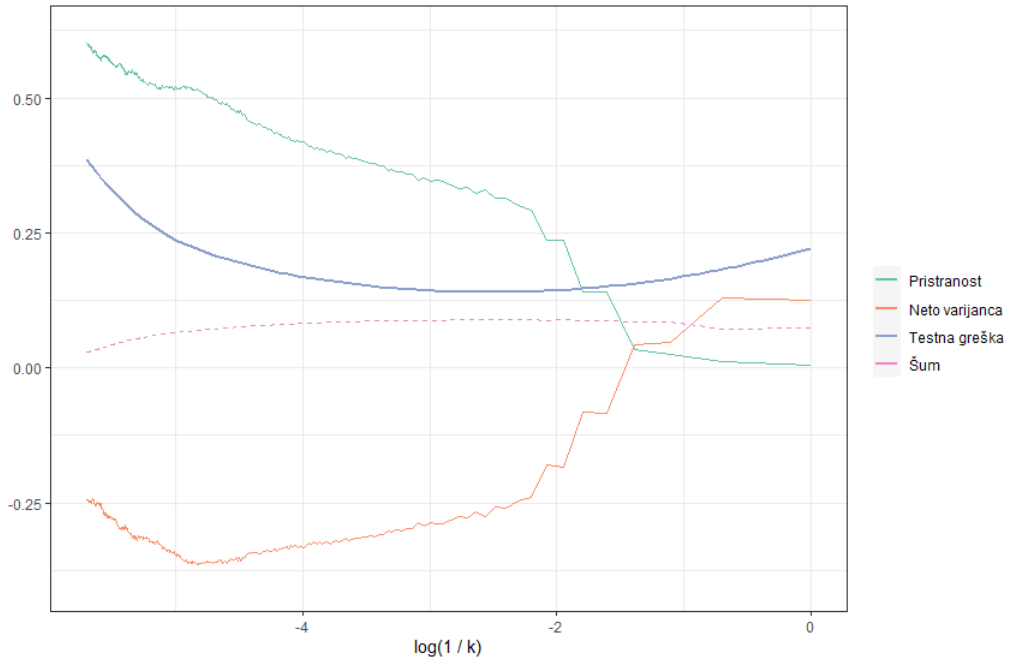
Graf prikazuje ponašanje prosječne pristranosti, varijance, šuma i testne greške ovisno o broju susjeda u kNN metodi, na simuliranom setu podataka. Šum je iskazan kao prosječna vrijednost $c_1 N(x)$.

$$L_x(\hat{f}) = 0.9 \cdot \mathbb{1}_{(\hat{f}(x)=y_*)} + 0.1 \cdot \mathbb{1}_{(\hat{f}(x)\neq y_*)}.$$

Prosječna očekivana testna greška je tada

$$\hat{\mathbb{E}}_T[L_x(\hat{f})] = \sum_{\hat{f}\in M} \frac{L_x(\hat{f})}{100}.$$

Tako dobivene vrijednosti očekivane testne greške, pristranosti, varijance i neto varijance $(1 - 2B(x))V(x)$ ponovo uprosječujemo po svim kovarijatama u testnom skupu kako bi dobili procjenu njihovih prosječnih očekivanih vrijednosti.
Na grafu 2.2 primjećujemo da se prosječna pristranost ponaša kao i očekivano, raste povećanjem vrijednosti parametra k, no prosječna varijanca također raste. Zanimljivo je i da se zbog prisutnosti multiplikativnog faktora c_1 utjecaj šuma smanjuje s rastom testne greške. Također,



Slika 2.3: Dekompozicija testne greške modela

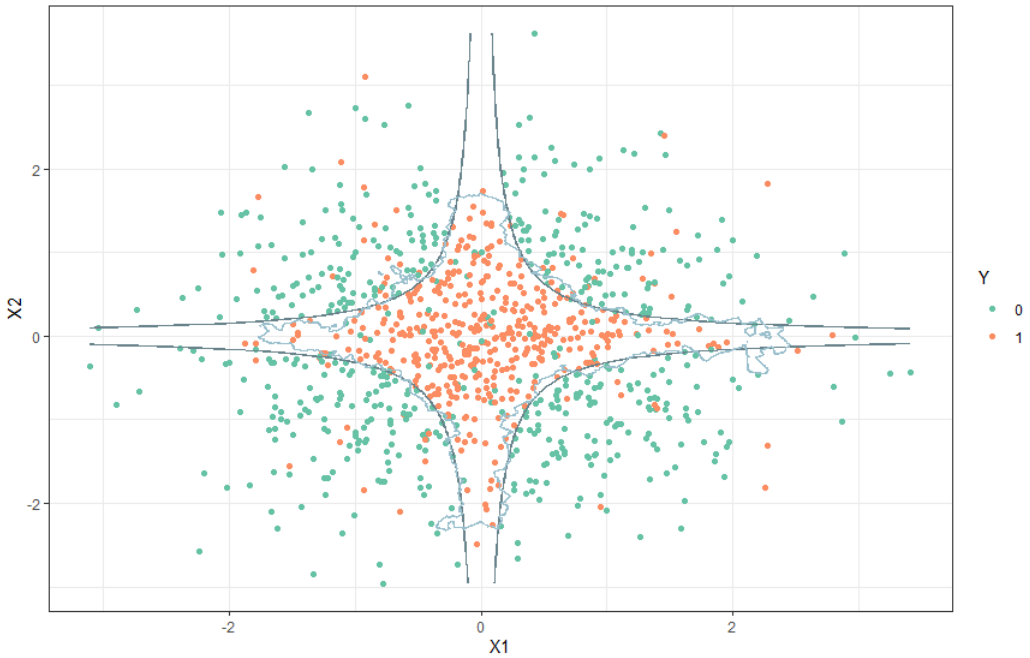
Uz pretpostavku da u podacima nema šuma, prosječna testna greška je jednaka sumi prosječne pristranosti, šuma i neto varijance.

prosječna testna greška se ne ponaša kao suma tih triju vrijednosti, što je najbolje vidljivo na dijelu grafa na kojemu su pristranost i varijanca veće od testne greške, pa ovaj primjer dobro ilustrira kako klasična dekompozicija za srednjekvadratnu grešku nije primjenjiva za 0-1 funkciju gubitka.

Uzimajući u obzir rezultate teorema 1.2.2 i prijašnju diskusiju, sama varijanca nije jedan od članova dekompozicije testne greške, već je zamijenjena *neto varijancom*, iskazanom kao $(1 - 2B(x))V(x)$.

Na grafu 2.3 vidljivo je da rastav iz teorema 1.2.2 vrijedi, te je sada testna greška jednaka sumi pristranosti, neto varijance i šuma, pri čemu neto varijanca ponekad poprima negativne vrijednosti. Vrijednost prosječne neto varijance se sada smanjuje sa porastom parametra k, osim za $k > 273$ gdje joj vrijednost raste. Testna greška je najmanja kada je $k = 9$, te u tom slučaju iznosi 0.142, što je jako dobar rezultat s obzirom da ireducibilna greška iznosi 0.1. Sada učimo model kNN za taj k na realizaciji skupa za učenje prikazanoj na slici 2.1.

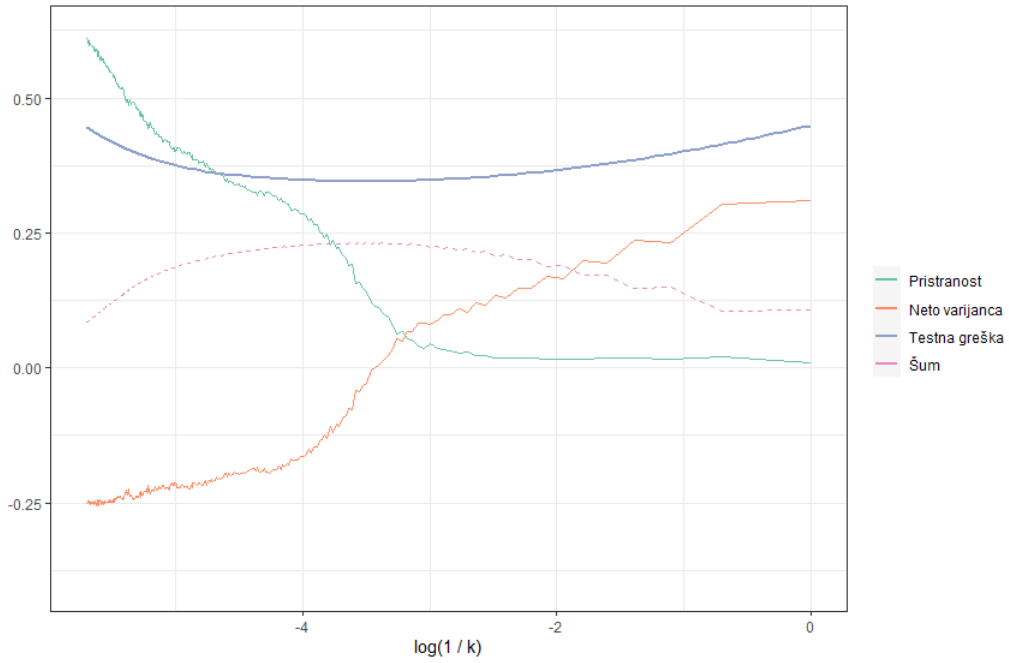
Na slici 2.4 se vidi da je tako dobiveni model uspio relativno dobro aproksimirati stvarnu granicu odluke, iako mu prediktivna moć opada na rubovima grafa zbog manjeg broja



Slika 2.4: Granica odluke kNN modela

Svjetloplave linije predstavljaju granicu odluke modela kNN metode s najmanjom testnom greškom, naučenog na skupu za učenje duljine 1000.

obzervacija.
Graf 2.5 prikazuje dekompoziciju očekivane testne greške modela na podacima distribuiranim kao ranije, s razlikom da je sada ϵ uzorkovan iz Bernoullijeve distribucije za $p = 0.3$. Sve vrijednosti su procijenjene istom metodom kao prije. Vidljivo je da su sada s rastom parametra k i porast pristranosti i smanjenje varijance sporiji, te se minimum očekivane testne greške postiže tek u $k = 33$, i iznosi 0.345. Ovakvi rezulati su u skladu s očekivanjima. Intuitivno, s većom razinom šuma, potrebne su veće vrijednosti k kako bi bili sigurni da ispravna klasa prevladava među najbližim susjedima.



Slika 2.5: Dekompozicija testne greške modela na podacima s puno šuma

Poglavlje 3

Dodatak

U ovome odjeljku su ukratko objašnjene metode stabala odluke i bagging. Opisi ne idu u detalj budući da sam princip po kojemu te metode funkcioniraju nije bitan za ovaj rad, pa se tako spominje njihova primjena isključivo u problemima klasifikacije. U [3] se može naći temeljitiji pregled svih spomenutih metoda.

3.1 Stabla odluke

Metode bazirane na stablima odluke dijele prostor prediktora $D \subseteq \mathbb{R}^p$ na manje pravokutnike koristeći binarne rekurzije, te na svakom od tih pravokutnika prilagođavaju jednostavan model, najčešće konstantu. Model se tada iskaže kao

$$\hat{f}(x) = \sum_{t \in T} \hat{c}_t \mathbb{1}_{\{x \in t\}}, \tag{3.1}$$

gdje je T skup svih podskupova dobivenih particioniranjem, a $\hat{c}_t$ je iznos predikcije kada je obzervacija sadržana u podskupu t.
Neka je $n(t) = |t|$, te $n_k(t) = |\{x_i \in t : y_i = k\}|$. Definiramo

$$\hat{p}_k(t) = \frac{n_k(t)}{n_t}. \tag{3.2}$$

Za 0-1 funkciju gubitka, predikciju klasifikacijskog stabla $\hat{c}_t$ procjenjujemo kao $\arg\max_{k \in S} \hat{p}_k(t)$. Definiramo "trošak" stabla T kao

$$C(T) = \sum_{t \in T} \hat{p}(t) i(t) \tag{3.3}$$

za

$$\hat{p}(t) = \frac{n(t)}{n}, \tag{3.4}$$

gdje je n duljina uzorka. i je proizvoljna funkcija koja predstavlja tzv. mjeru nečistoće, za što se često koriste *Ginijev indeks*

$$i(t) = \sum_{k \neq k'} \hat{p}_k(t)\hat{p}_{k'}(t) \tag{3.5}$$

ili *entropija*

$$i(t) = -\sum_{k \in S} \hat{p}_k(t) \log(\hat{p}_k(t)). \tag{3.6}$$

Dijeljenje stabla se radi na način da se za dati $j \in \{1, .., p\}$ i $s \in \mathbb{R}$, za svaki $t \in T$ definiraju $t_L = \{x \in t : x_j \leq s\}$ i $t_D = \{x \in t : x_j > s\}$, s čime se dobije novo stablo $T(j, s)$. Tada za točku dijeljenja tražimo j_* i s_* koji maksimiziraju $C(T) - C(T(j, s))$. Može se pokazati da su za dani $t \in T$

$$(j_*, s_*) = \arg\min_{j,s} \{\hat{p}(t_L)i(t_L) + \hat{p}(t_D)i(t_D)\}. \tag{3.7}$$

Dijeljenje se odvija sve dok se ne zadovolji proizvoljno određen kriterij zaustavljanja, na primjer dok god dodavanje novog čvora smanjuje grešku za više od λ, za neki $\lambda > 0$.

3.2 Bagging

Mala stabla odluke najčešće nisu u stanju obuhvatiti svu kompleksnost modela, a velika imaju veliku varijancu zbog osjetljivosti odabira čvorova na ulazne podatke. U svrhu smanjivanja visoke varijance velikih stabala koristi se bagging, metoda u kojoj se prvo bootstrapom nad početnim skupom generiraju novi skupovi za učenje, te se nad svakim od tih skupova nauči stablo odluke. Za predikciju se tada uzima najčešća klasifikacija među svim tako dobivenim stablima. Varijanca se pritom tipično smanjuje, a pristranost ostaje nepromijenjena.

Bibliografija

[1] Leo Breiman, *Bagging predictors*, Machine learning **24** (1996), 123–140.

[2] Pedro Domingos, *A unified bias-variance decomposition*, Proceedings of 17th international conference on machine learning, Morgan Kaufmann Stanford, 2000, str. 231–238.

[3] Trevor Hastie, Robert Tibshirani i Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, sv. 2, Springer, 2009.

[4] David Pfau, *A generalized bias-variance decomposition for bregman divergences*, Unpublished Manuscript (2013).

Sažetak

Cilj ovog rada je istražiti odnos pristranosti i varijance u okvirima van srednjekvadratne greške, u svrhu boljeg razumijevanja uspješnosti pojedinih metoda.
Na početku rada se definiraju osnovni pojmovi vezani uz statističko učenje, nakon čega se navode generalizirane definicije optimalne i glavne predikcije, pristranosti, varijance i šuma. Koristeći te definicije oblikuje se generalni oblik dekompozicije testne greške, te se pokazuje da tako definirana dekompozicija vrijedi za srednjekvadratnu grešku. Nakon srednjekvadratne obrađuje se 0-1 funkcija gubitka u slučajevima klasifikacije. Pokaže se da dekompozicija vrijedi i tada, s tim da se iznosi šuma i varijance množe skalarima. Vrijednosti tih skalara otkrivaju da porast varijance smanjuje testnu grešku u pristranim primjerima. Sličan rastav s modificiranim vrijednostima skalara postoji i u slučajevima kada postoje više od dvije klase, te kada je funkcija gubitka asimetrična. U slučaju kada je 0-1 greška modificirana na način da se i točni pogodci ocjenjuju s pozitivnom vrijednošću, rastav sadrži dodatan član. Također se definira pojam Bregmanove divergencije, te se pokaže da dekompozicija vrijedi ako je funkcija gubitka Bregmanova divergencija.
U problemima klasifikacije često se koriste stabla odluke, metode koje skup za učenje iterativno dijele na manje podskupove, te na svakom podskupu prilagođavaju jednostavan model. Budući da takve metode često imaju visoku varijancu, u svrhu smanjivanja iste koristi se bagging, metoda bazirane na agregiranju rezultata brojnih stabala odluke. Uspješnost te metode je u prijašnjoj literaturi objašnjena kroz pojam ispravno poredane metode, te se pokaže da je metoda ispravno poredana u x ako je nepristrana u x.
Ako je funkcija gubitka metrika, za nju ne mora nužno postojati dekompozicija, kao što je to slučaj za apsolutnu grešku. Ipak, iznos njene testne greške se može ograničiti odozgo i odozdo koristeći šum, pristranost i varijancu.
Svojstva dekompozicije u slučajevima klasifikacije su prikazana na simuliranom skupu podataka, gdje se pokaže odnos pristranosti i varijance na primjeru modela k-najbližih susjeda sa 0-1 funkcijom gubitka.
Konačno, radi boljeg razumijevanja rada spomenut je proces kojim metode stabala odluke, te bagging generiraju predikcije.

Summary

The goal of this thesis is to explore the relationship between bias and variance outside of the constraints of squared-loss error, with the purpose of a better understanding of success of certain statistical learning methods.
At the beginning, basic terms related to statistical learning are defined, after which a generalized definition is given for the main and optimal predictions, noise, bias and variance. Using those definitions, a general form of test error decomposition is formed, and it is shown that a decomposition defined as such holds for squared error. Afterwards the topic shifts to 0-1 error in terms of classification problems. It is shown that the decomposition holds in that case as well, with noise and variance both multiplied by scalars. Those scalar values reveal that higher variance reduces test error in biased examples. A similar decomposition with adjusted scalar values is shown to exist when there are more than two classes, and also with an asymmetrical 0-1 error modification. If the 0-1 error is modified in a way that grades correct guesses with a positive value, the decomposition contains an additional member. Bregman divergences are also defined, and it is shown that a decomposition holds if the loss function is a Bregman divergence.
In classification methods, decision trees are often used, which are methods that repeatedly divide the training set into smaller subsets, and then apply a simple method to each subset separately. Since those methods often display high variance, bagging, a method that aggregates results of multiple decision trees is used. The success of bagging was explained in the literature using correctly ordered methods, and it is shown that a method is correctly ordered in x if it is unbiased in x.
If a loss function is a metric, there does not necessarily need to exist a valid decomposition for it, as is the case with absolute loss. Still, the value of its test error can be limited from above and below using noise, bias and variance.
Properties of the decomposition in classification problems are shown on a simulated data set, where the bias-variance trade-off is shown for a k-nearest neighbours model with a 0-1 loss function.
Finally, for better understanding of the thesis, an outline is given for the process through which decision tree methods, as well as bagging, generate predictions.

Životopis

Rođen sam 15. prosinca 1995. u Šibeniku. Pohađao sam osnovnu školu u Vodicama gdje sam i odrastao, te sam nakon toga upisao Prirodoslovno-matematičku gimnaziju Antuna Vrančića u Šibeniku. Kroz te periode obrazovanja sam redovno nastupao po brojnim natjecanjima iz matematike na županijskoj, regionalnoj i državnoj razini. Ipak, 2014. godine, po završetku srednje škole upisujem smjer medicinu na Medicinskom fakultetu u Splitu, da bi kasnije odlučio kako je matematika ispravniji izbor za mene, pa se 2017. godine prebacujem na Preddiplomski sveučilišni studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Po završetku preddiplomskog studija 2020. godine, nastavljam akademsko obrazovanje upisom diplomskog sveučilišnog studija Matematička statistika na istom fakultetu.