Vrednovanje i odabir modela

Kranželić, Tin

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tin Kranželić

VREDNOVANJE I ODABIR MODELA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, studeni 2021.

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. _____

Mojoj obitelji, prijateljima, kolegama i svima ostalima koji su dali neki doprinos da dospijem do ovdje.

Posebno hvala kolegi i prijatelju Davoru Iljkiću na podršci i abnormalnoj količini smijeha tijekom ovih godina.

Sadržaj

Uvod

Napredak tehnologije današnje civilizacije doveo je do ogromnih količina podataka koje se generiraju svakodnevno u mnogim područjima ljudskog djelovanja. Sadašnjost nalaže da će te količine u budućnosti samo rasti. Tamo gdje postoje podaci, postoji i skriveno znanje i uvidi koji iza tih podataka stoje. Upravo je to razlog zbog kojeg će zanimanje statističara dobivati na sve većem značaju u budućnosti. U ovome radu pažnju ćemo posvetiti ključnim metodama bez kojih bi bilo kakvo predviđanje ili razumijevanje na temelju dosadašnjih podataka bilo nemoguće sa stanovišta teorije statističkog učenja, mlade teorije u čijim temeljima počiva statistika. Teorija statističkog učenja proučava fenomene koji su puno fundamentalniji i bliži nama ljudima, nego što bi se na prvu moglo reći. Primjera radi, svaki čovjek posjeduje sebi svojstvenu interpretaciju svijeta u kojem živi. Upravo ljudske interpretacije možemo promatrati kao svojevrsne modele. Naše interpretacije su oblikovane kroz iskustva. Ilustrativno, recimo da čovjek posjeti prvi put neki restoran i naruči neko jelo, te mu se svidi. Poučen tim iskustvom, za tog čovjeka je sasvim prirodno za očekivati da ako naruči neko drugo jelo u tom restoranu, da će mu se i ono svidjeti. Jednako kao što su ljudske interpretacije različite i niti jedna nije uvijek savršeno precizna i istinita, takvi su i generalno razni modeli koji se mogu na matematički precizan način opisati, a koje ćemo mi proučavati u ovom radu.
Glavna literatura koju ovaj rad prati jest [3]. Ovaj rad sadržava četiri poglavlja.
U prvom poglavlju navodimo neke osnovne pojmove i rezultate korištene u preostalim poglavljima, te pružamo motivaciju u vidu jednostavnih primjera.
U drugom poglavlju opisujemo važna svojstva modela, pristranost i varijancu, te analiziramo odnos između njih i način na koji porast kompleksnosti modela utječe na njih. Također pružamo konkretne primjere.
U trećem poglavlju analiziramo metode koje procjenjuju unutar-uzoračku pogrešku sa svrhom rješavanja problema odabira modela. Konkretno, opisujemo informacijske kriterije AIC i BIC, te pružamo konkretan primjer uporabe kriterija.
U četvrtom i zadnjem poglavlju opisujemo opću i jednostavnu metodu koja direktno procjenjuje izvan-uzoračku pogrešku, metodu unakrsnog vrednovanja. Na kraju ponovno dajemo konkretan primjer upotrebe metode.

Poglavlje 1

Preliminarni rezultati

1.1 Osnovni pojmovi i modeli

Na početku navedimo osnovne pojmove i koncepte koji će biti neophodni za ostatak ovog rada.

Vrednovanje i odabir modela predstavlja neizostavnu komponentu puno šire teorije zvane statističko učenje. Grubo rečeno, statističko učenje se odnosi na širok skup tehnika namijenjenih razumijevanju podataka. Osnovna podjela tih tehnika je na tzv. nadzirane i nenadzirane. Razlika između nadziranih i nenadziranih tehnika leži u tome što nadzirane tehnike neposredno iskorištavaju oznaku podataka za treniranje modela, dok su nenadzirane tehnike korištene nad neoznačenim podacima. U primjerima ovog rada, koristiti ćemo se isključivo nadziranim tehnikama. Tehnike nadziranog učenja za cilj imaju procjeniti statistički model iz podataka, u svrhu buduće procjene ili predikcije, ili pak dubljeg razumijevanja samog mehanizma koji podatke generira.

Nadalje, problemi u nadziranom učenju se mogu podjeliti u kvantitativne i kvalitativne. Sama podjela ovisi o tipu zavisne varijable. U slučaju kvantitativnog problema, ona je numerička, tj. neprekidna i poprima vrijednosti najčeše u skupu realnih brojeva $\mathbb{R}$. S druge strane, ako je zavisna varijabla diskretna ili kategorijska, tj. poprima konačno mnogo vrijednosti, tada je problem kvalitativan.
Podaci za treniranje u nadziranom učenju dolaze u obliku skupa:

$$\tau = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}, \tag{1.1}$$

pri čemu uređeni par $(x_i, y_i), i = 1, \cdots, N$ predstavlja jedno konkretno opažanje u obliku p-dimenzionalnog vektora nezavisnih varijabli - $x_i, p \geq 1$ i zavisne varijable - y_i.

Preostaje nam još samo ukratko opisati dva opća i često korištena modela koje ćemo i mi koristiti u primjerima ovog rada. Za kvantitativan problem, to je linearna regresija, a za kvalitativan problem, to je K - najbližih susjeda.

Linearna regresija pretpostavlja sljedeći oblik veze između zavisne varijable Y i nezavisnih varijabli, tj. p prediktora $X_1, X_2, \cdots, X_p$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{1.2}$$

pri čemu su $\beta_0, \beta_1, \cdots, \beta_p$ koeficijenti ili parametri modela, a ε slučajna varijabla s očekivanjem nula koja odgovara nepredvidivom dijelu, tzv. slučajna pogreška.
Linearna regresija pripada tzv. parametarskim modelima, te time njena upotreba mora zadovoljavati nekoliko jakih pretpostavki. Neke od najvažnijih pretpostavki su linearni odnos između varijabli poticaja i odaziva, nezavisnost grešaka, homogenost grešaka, te normalna distribuiranost grešaka. Ukoliko neka od pretpostavki nije opravdana, naša predviđanja mogu biti nevaljana.
Procjenitelje parametara $\beta_0, \beta_1, \cdots, \beta_p$, u oznaci $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, dobivamo iz skupa podataka za treniranje τ često korištenom metodom najmanjih kvadrata. Njome se minimizira srednje kvadratna greška, tj.

$$\hat{\beta} = \min_{\beta} RSS(\beta), \tag{1.3}$$

pri čemu su $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$, $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$, te

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{1.4}$$

Sada slijedi da je za novu točku $x_0 = (x_{01}, x_{02}, \cdots, x_{0p})$ predikcija $\hat{y}$ dana u obliku:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \tag{1.5}$$

Vrijedi napomenuti kako postoji puno ekstenzija općeg linearnog regresijskog modela ovdje prezentiranog, poput npr. polinomijalne regresije kojom se relaksira pretpostavka linearnosti.

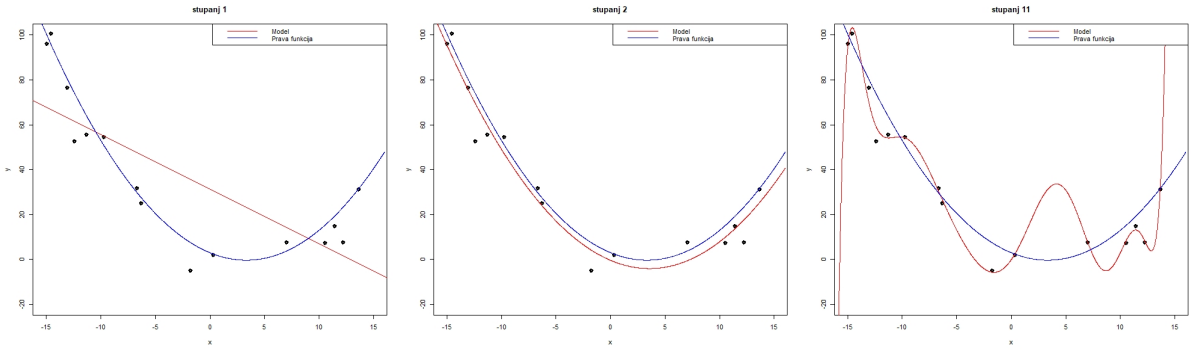
Općenito u kvalitativnim, tj. klasifikacijskim problemima, najpreciznija tehnika koja pravi najmanje pogrešnih klasifikacija, jest tzv. Bayesov klasifikator. Jednostavno rečeno, on pridružuje opažanje onoj klasi koja je najvjerojatnija obzirom na vrijednosti prediktora. U pozadini te tehnike leže poznati Bayesov teorem i uvjetna vjerojatnost. Primjenom te tehnike dobivamo i tzv. Bayesovu granicu odluke. Problem u praksi sa stvarnim podacima je u tome što mi ne znamo uvjetnu vjerojatnost zavisne varijable obzirom na vrijednosti nezavisnih varijabli, te je time nemoguće izračunati Bayesov klasifikator. Mnoge metode

pokušavaju procijeniti tu uvjetnu vjerojatnost, a jedna od njih je i metoda K - najbližih susjeda.
K - najbližih susjeda (eng. KNN) je neparametarska klasifikacijska tehnika, dakle ona ne pravi nikakvu pretpostavku o funkcijskom obliku veze između zavisne varijable i nezavisnih varijabli. Time nema nikakvih pretpostavki koje moraju biti zadovoljene. Za pozitivan cijeli broj K i novo opažanje x_0, KNN klasifikator pronalazi K najbližih točaka unutar skupa τ, u oznaci N_0. Potom procjenjuje uvjetnu vjerojatnost za određenu klasu kao udio točaka unutar N_0 čija je vrijednost zavisne varijable upravo ta klasa. Konačno, KNN klasifikator klasificira točku x_0 onoj klasi čija je uvjetna vjerojatnost najveća.

1.2 Motivacija i cilj

Pogledajmo neke jednostavne primjere koji će čitatelja uvesti u problematiku ovog rada.

Primjer 1.2.1. *Simuliramo situaciju na sljedeći način:*
Imamo jednu numeričku zavisnu varijablu y, te jednu numeričku nezavisnu varijablu x. Veza između varijable odziva i varijable poticaja je dana jednadžbom $y = 3-2*x+0.3*x^2$. *Također, da bi čitava stvar bila što realnija, dodajemo šum koji predstavlja neizostavnu komponentu svakidašnjeg života, a to je slučajnost. Jasno, u stvarnosti nemamo informaciju o pravoj naravi takve veze.*



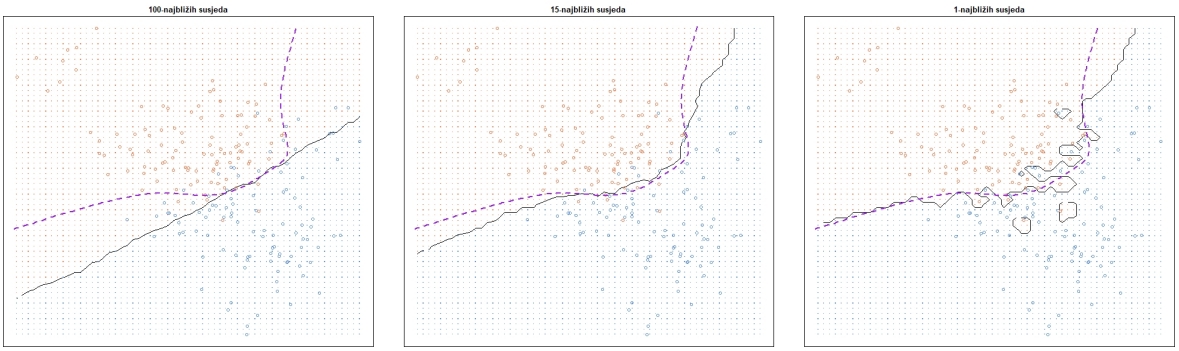
Slika 1.1: Plavom bojom označena je prava funkcija, dok su crvenom bojom označeni modeli različitih kompleksnosti.

Iako u ovom trenutku još nije definirano što bi za model točno značilo da je "najbolji", na intuitivnoj razini je jasno da bismo htjeli da je naš model što bliži pravoj funkciji, stoga

se možemo zapitati koji od tri modela sa Slike 1.1 bi bio "najbolji". Kao što vidimo, jednostavniji lijevi model dane podatke opisuje lošije nego preostala dva modela. S druge strane, kompliciraniji desni model dane podatke opisuje najbolje, gotovo pa savršeno. Međutim, što bi se dogodilo kada bi generirali još jedan podatak? U tom slučaju jasno vidimo kako bi srednji model gotovo sigurno bio bliži od preostala dva. O čemu se ovdje radi i što se točno dogodilo, vidjeti ćemo u nastavku ovog rada.

Pogledajmo još jedan primjer u kojemu je zavisna varijabla ovog puta binarna kategorijska. Shodno tome, tehnika koju koristimo više nije linearna regresija, već K - najbližih susjeda.

Primjer 1.2.2. *Zavisna varijabla y poprima jednu od dvije moguće vrijednosti, plava ili narančasta. Imamo i dvije nezavisne numeričke varijable. U ovom trenutku nisu važni detalji same simulacije, pa ih izostavljamo.*



Slika 1.2: Slijeva nadesno raste kompleksnost modela. Isctrkanom ljubičastog linijom je naznačena Bayesova granica odluke koja predstavlja optimalnu granicu odluke. Budući je ovo simulirana situacija, moguće ju je konkretno odrediti, dok u stvarnosti to nije slučaj.

Možemo vidjeti da je ishod jednak kao u prošlom primjeru. Laički rečeno, opet smo dobili da je srednje kompleksan model nekako najbliži pravoj istini.

Dva važna aspekta teorije statističkog učenja koja ćemo mi obraditi su odabir modela i vrednovanje modela. Odabir modela obuhvaća procjenjivanje performansi raznih modela u cilju odabira najboljeg i u skladu s određenim kriterijem. Vrednovanje modela se najčešće odnosi na procjenu pogreške predikcije, tj. predikcijske sposobnosti modela na neviđenom skupu podataka. Postoji puno metoda za procjenu, a mi ćemo obraditi neke od najvažnijih.

Za kraj ovog uvodnog poglavlja, korisno bi bilo istaknuti izreku britanskog statističara George E. P. Box-a: "Svi su modeli pogrešni, ali neki su korisni."

Poglavlje 2

Odnos varijance i pristranosti

2.1 Uvod

Kada imamo neki model koji je istreniran na skupu podataka za treniranje τ, nas ustvari zanima koliko je on "dobar", tj. koliko precizne predikcije će davati za nove, neviđene podatke. Nezavisan skup koji će sadržavati takve podatke zvati ćemo testni skup podataka, a opisanu sposobnost precizne predikcije modela zvati ćemo *generalizacija* modela. Kada kažemo da neki model dobro generalizira, mislimo na to da daje precizne predikcije na novim podacima. U tekstu koji slijedi, vidjeti ćemo da važnu ulogu u generalizaciji modela igraju *pristranost* te *varijanca* tog modela.

2.2 Pristranost, varijanca i kompleksnost

Pretpostavimo da imamo zavisnu varijablu Y, vektor nezavisnih varijabli X i model $\hat{f}$ procijenjen iz skupa podataka za treniranje τ. U svrhu mjerenja kvalitete predikcije i procijenjenog modela $\hat{f}$ potrebna nam je neka metrika. Zato uvodimo *funkciju gubitka* $L(Y, \hat{f}(X))$.

Sada pretpostavimo prvo da je varijabla odziva Y kvantitivna. Standardni odabiri za funkciju gubitka $L(Y, \hat{f}(X))$ u kvantitivnom slučaju su kvadratna greška:

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2 \tag{2.1}$$

ili apsolutna greška:

$$L(Y, \hat{f}(X)) = |Y - \hat{f}(X)| \tag{2.2}$$

Testna pogreška ili *generalizacijska pogreška* odnosi se na predikcijsku pogrešku nad nezavisnim testnim skupom podataka pri čemu je τ neki konkretan, fiksan skup podataka

za treniranje o kojemu sama pogreška ovisi. Označujemo ju sa:

$$Err_\tau = \mathbb{E}[L(Y, \hat{f}(X)) \mid \tau] \tag{2.3}$$

Međutim, ukoliko se želimo rješiti i slučajnosti unutar samog τ koja utječe na $\hat{f}$, tada dolazimo do *očekivane testne pogreške* ili *očekivane predikcijske pogreške*, u oznaci Err. Po zakonu potpunog očekivanja imamo:

$$\begin{aligned} Err &= \mathbb{E}[Err_\tau] \\ &= \mathbb{E}[\mathbb{E}[L(Y, \hat{f}(X)) \mid \tau]] \\ &= \mathbb{E}[L(Y, \hat{f}(X))] \end{aligned} \tag{2.4}$$

Nadalje, *pogrešku treniranja* definiramo kao:

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i)) \,. \tag{2.5}$$

Dakle, pogreška treniranja je prosječna greška na uzorku za treniranje.

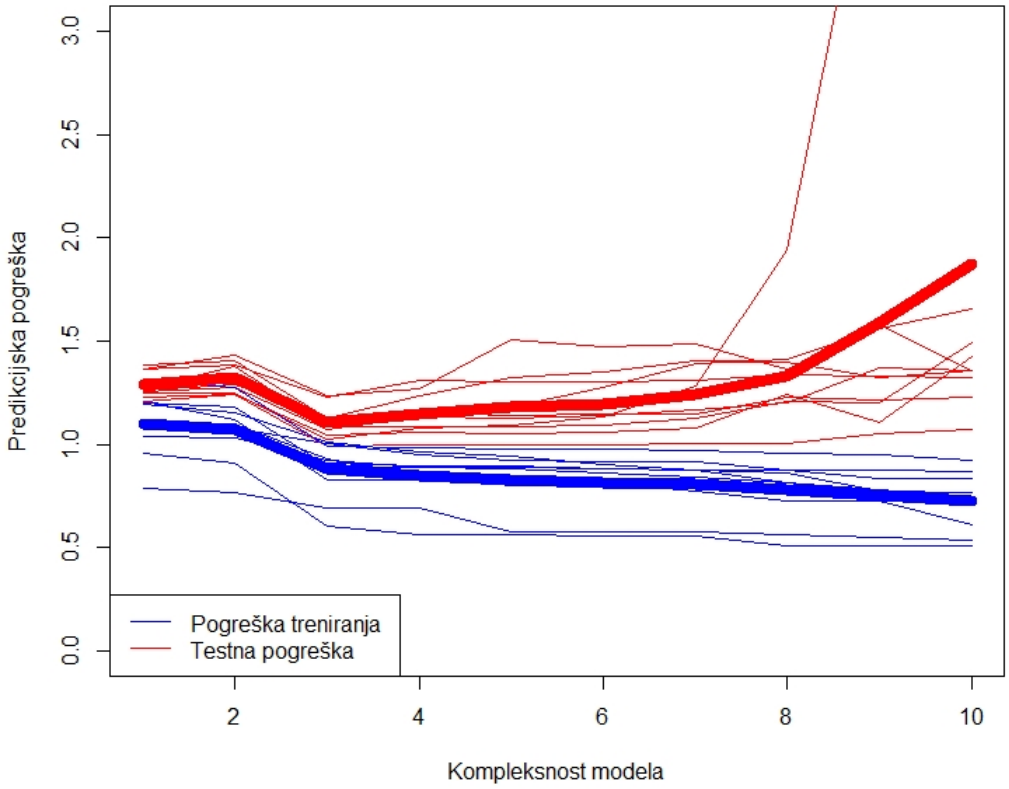
Upravo definirana pogreška treniranja nije dobar procjenitelj veličine koja će nas zanimati, a to je testna pogreška Err_τ. Razlog leži u tome što se pogreška treniranja konstantno smanjuje s povećanjem kompleksnosti modela, dok se testna pogreška smanjuje do određenog momenta nakon kojeg počinje rasti. Naime, što je model kompleksniji to je u stanju bolje se prilagoditi danom skupu podataka za treniranje. Pretjeranom prilagodbom model će pokušavati objasniti strukturu u podacima koja je ustvari nepostojeća, odnosno koja može dolaziti od slučajne pogreške ili neobjašnjivih fenomena vezanih uz podatke. Takav model će loše generalizirati. Za takav model još kažemo da ima problem pretreniranosti (eng. overfitting), dok za model koji osim što loše generalizira, ima i visoku pogrešku treniranja, kažemo da ima problem podtreniranosti (eng. underfitting). Opisano ponašanje pogreške treniranja i testne pogreške može se vidjeti na Slici 2.1.

Analogna priča vrijedi i za slučaj kada je varijabla odziva Y kvalitativna, tj. poprima vrijednosti unutar skupa $\{1, 2, \ldots, K\}$. Budući da sada procijenjujemo uvjetne vjerojatnosti, potrebne su male modifikacije u definiciji funkcije gubitka, te posljedično tome su pogreška treniranja i testna pogreška drukčije.
Označimo sa $\hat{Y}(X) = arg\,max_k\, \hat{p}_k(X)$, pri čemu je $p_k(X) = \mathbb{P}(Y = k \mid X)$.
Standardan odabir za funkciju gubitka u klasifikacijskom slučaju jest $0 - 1$ gubitak:

$$L(Y, \hat{Y}(X)) = I(Y \neq \hat{Y}(X)) = \begin{cases} 0, & \text{za} \quad Y = \hat{Y}(X) \\ 1, & \text{za} \quad Y \neq \hat{Y}(X) \,. \end{cases} \tag{2.6}$$

Još jedan moguć izbor za funkciju gubitka u klasifikacijskom slučaju:

$$L(Y, \hat{p}(X)) = -2 \sum_{k=1}^{K} I(Y = k) \log \hat{p}_k(X) = -2 \log \hat{p}_Y(X) \,. \tag{2.7}$$



Slika 2.1: Pogreška treniranja i testna pogreška u ovisnosti o kompleksnosti modela koja raste slijeva nadesno. Podebljane linije su prosječne vrijednosti pogrešaka sto modela istreniranih na deset simuliranih skupova podataka. Prava simulirana funkcija f je sinusoida, te su istrenirani modeli linearne, odnosno polinomijalne regresije do desetog stupnja. Kao što vidimo, dobili smo tzv. *U - oblik* krivulje procijenjene očekivane testne pogreške. U ovom slučaju, model trećeg stupnja pokazao se kao optimalni model.

Prateći porast kompleksnosti krivulji sa Slike 2.1, kažemo da se njihova pristranost smanjuje dok se njihova varijanca povećava. Prisjetimo se matematičkih defincija ova dva pojma. Pristranost (eng. bias) procjenitelja $\hat{\theta}$ nekog općenitog parametra ili funckije θ jest:

$$Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta\,. \tag{2.8}$$

Varijanca (eng. variance) slučajne varijable $\hat{\theta}$ jest:

$$Var(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]\,. \tag{2.9}$$

Analizirajmo sada odnos ova dva pojma kod procijenjenih modela. Pretpostavimo da se veza između zavisne varijable Y i vektora nezavisnih varijabli X može zapisati kao $Y = f(X) + \varepsilon$, pri čemu je f nepoznata funkcija, a ε odgovara slučajnoj pogrešci koja je nezavisna od X te vrijedi $\mathbb{E}[\varepsilon] = 0$ i $\sigma_\varepsilon^2 = Var(\varepsilon)$. Koristeći kvadratnu grešku, imamo da je očekivana predikcijska pogreška za procjenjeni regresijski model $\hat{f}(X)$ u točki $X = x_0$:

$$\begin{aligned}
Err(x_0) &= \mathbb{E}[(Y - \hat{f}(x_o))^2 \mid X = x_0] && (2.10)\\
&= \mathbb{E}[(f(X) + \varepsilon - \hat{f}(x_0))^2 \mid X = x_0] && (2.11)\\
&= \mathbb{E}[(f(x_0) + \varepsilon - \hat{f}(x_0) + \mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)])^2] && (2.12)\\
&= \mathbb{E}[(f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 + 2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])\varepsilon + \varepsilon^2 && (2.13)\\
&\quad + (\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2\\
&\quad + 2(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))(f(x_0) - \mathbb{E}[\hat{f}(x_0)])\\
&\quad + 2(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))\varepsilon]\\
&= (f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 + 2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])\,\mathbb{E}[\varepsilon] + \mathbb{E}[\varepsilon^2] && (2.14)\\
&\quad + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] + 2(f(x_0) - \mathbb{E}[\hat{f}(x_0)])\,\mathbb{E}[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)]\\
&\quad + 2\,\mathbb{E}[\varepsilon]\,\mathbb{E}[\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)]\\
&= (f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 + \mathbb{E}[\varepsilon^2] + \mathbb{E}[(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] && (2.15)\\
&= Bias^2(\hat{f}(x_0)) + \sigma_\varepsilon^2 + Var(\hat{f}(x_0)) && (2.16)
\end{aligned}$$

pri čemu smo u (2.14) iskoristili međusobnu nezavisnot slučajne pogreške ε i $\hat{f}(x_0)$.
Prvi član u jednakosti (2.16) predstavlja kvadrat pristranosti[1] odabranog modela te govori za koliko se u prosjeku naša predikcija razlikuje od prave vrijednosti. Srednji član označava varijancu slučajne pogreške, odnosno varijancu zavisne varijable Y u točki x_0 oko njene srednje vrijednosti $f(x_0)$. Ta veličina je neizbježna i ne ovisi o našem odabiru modela, bez obzira koliko dobar procjenitelj on bio. Treći i zadnji član predstavlja varijancu naše predikcije u danoj točki. Dakle, kvadrat pristranosti i varijanca predstavljaju reducibilni dio predikcijske pogreške, dok varijanca slučajne pogreške, odnosno šum, predstavlja ireducibilan dio očekivane predikcijske pogreške.
Jasno je kako bi za model bilo idealno da ima nisku pristranost i nisku varijancu jer bi u tom

[1] Ponekad ćemo se u analizi odnosa varijanca-pristranosti na ovu veličinu referirati samo kao pristranost, ali ustvari mislimo na kvadrat pristranosti.

slučaju davao u prosjeku relativno precizne predikcije s niskom varijabilnosti. Međutim, općenito smanjenjem jedne vrijednosti dovodi do povećanja druge, i obrnuto. Povećanje kompleknosti modela $\hat{f}$ vodi k povećanju varijance i smanjenju pristranosti, dok smanjenje kompleknosti modela $\hat{f}$ vodi k smanjenju varijance i povećanju pristranosti. Dakle, postoji neka srednja kompleksnost modela sa uravnoteženim odnosom između varijance i pristranosti koji će davati najbolje rezultate.

Za kraj ovog potpoglavlja, izračunajmo pristranost-varijanca dekompoziciju za konkretan primjer regresijske metode k - najbližih susjeda. Primjena k - najbližih susjeda u regresiji je analogan primjeni u klasifikaciji. Dakle, za danu točku x_0 pronađemo k najbližih točaka, u oznaci $x_{(1)}, x_{(2)}, \cdots, x_{(k)}$ te predikciju $\hat{f}_k(x_0)$ odredimo kao prosječnu vrijednost opažanja zavisnih varijabli koja pripadaju pronađenih k najbližih točaka, u oznaci $y_{(1)}, y_{(2)}, \cdots, y_{(k)}$. Odnosno, matematički zapisano: $\hat{f}_k(x_0) = \frac{1}{k}\sum_{i=1}^{k} y_{(i)}$.
Ovdje će nam važna pretpostavka biti ta da opažanja nezavisnih varijabli $x_1, x_2, \cdots, x_N$ držimo fiksnima, dok će slučajnost dolaziti od zavisnih varijabli $Y_1, Y_2, \cdots, Y_N$. Odnosno, ta pretpostavka nam kaže da svi mogući skupovi podataka za treniranje modela imaju jednake vrijednosti nezavisnih varijabli, dok varijacija dolazi samo od vrijednosti pripadajućih opažanja zavisnih varijabli.
Uvrštavanjem u jednakost (2.15) dobivamo:

$$Err(x_0) = \mathbb{E}[(Y - \hat{f}_k(x_0))^2 \mid X = x_0] \tag{2.17}$$

$$= \sigma_\varepsilon^2 + (\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}] - f(x_0))^2 + \mathbb{E}[(\frac{1}{k}\sum_{i=1}^{k} Y_{(i)} - \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}])^2] \tag{2.18}$$

Možemo prvo srediti izraz $\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}]$:

$$\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}] = \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k}(f(x_{(i)}) + \varepsilon_{(i)})] \tag{2.19}$$

$$= \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}) + \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\varepsilon_{(i)})] \tag{2.20}$$

$$= \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}) \tag{2.21}$$

Uvrštavanjem (2.21) u izraz za pristranost iz (2.18) dobivamo:

$$(\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}] - f(x_0))^2 = (f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}))^2 \tag{2.22}$$

Preostaje uvrstiti (2.21) u izraz za varijancu iz (2.18):

$$\mathbb{E}[(\frac{1}{k}\sum_{i=1}^{k} Y_{(i)} - \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} Y_{(i)}])^2] = \mathbb{E}[(\frac{1}{k}\sum_{i=1}^{k}(f(x_{(i)}) + \varepsilon_{(i)}) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}))^2] \tag{2.23}$$

$$= \mathbb{E}[(\frac{1}{k}\sum_{i=1}^{k} \varepsilon_{(i)})^2] \tag{2.24}$$

$$= \frac{1}{k^2}\,\mathbb{E}[(\sum_{i=1}^{k} \varepsilon_{(i)})^2] \tag{2.25}$$

$$= \frac{1}{k^2}\,\mathbb{E}[(\sum_{i=1}^{k} \varepsilon_{(i)} - \mathbb{E}[\sum_{i=1}^{k} \varepsilon_{(i)}])^2] \tag{2.26}$$

$$= \frac{1}{k^2} Var(\sum_{i=1}^{k} \varepsilon_{(i)}) \tag{2.27}$$

$$= \frac{1}{k^2}\sum_{i=1}^{k} Var(\varepsilon_{(i)}) \tag{2.28}$$

$$= \frac{\sigma_\varepsilon^2}{k} \tag{2.29}$$

pri čemu smo u (2.26) iskoristili činjenicu $\mathbb{E}[\varepsilon] = 0$, a u (2.28) nekoreliranost slučajnih pogrešaka.
Uvrštavanjem (2.22) i (2.29) nazad u (2.18), u konačnici dobivamo da je pristranost-varijanca dekompozicija za regresijsku metodu k - najbližih susjeda sljedećeg oblika:

$$Err(x_0) = \sigma_\varepsilon^2 + (f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}))^2 + \frac{\sigma_\varepsilon^2}{k} \tag{2.30}$$

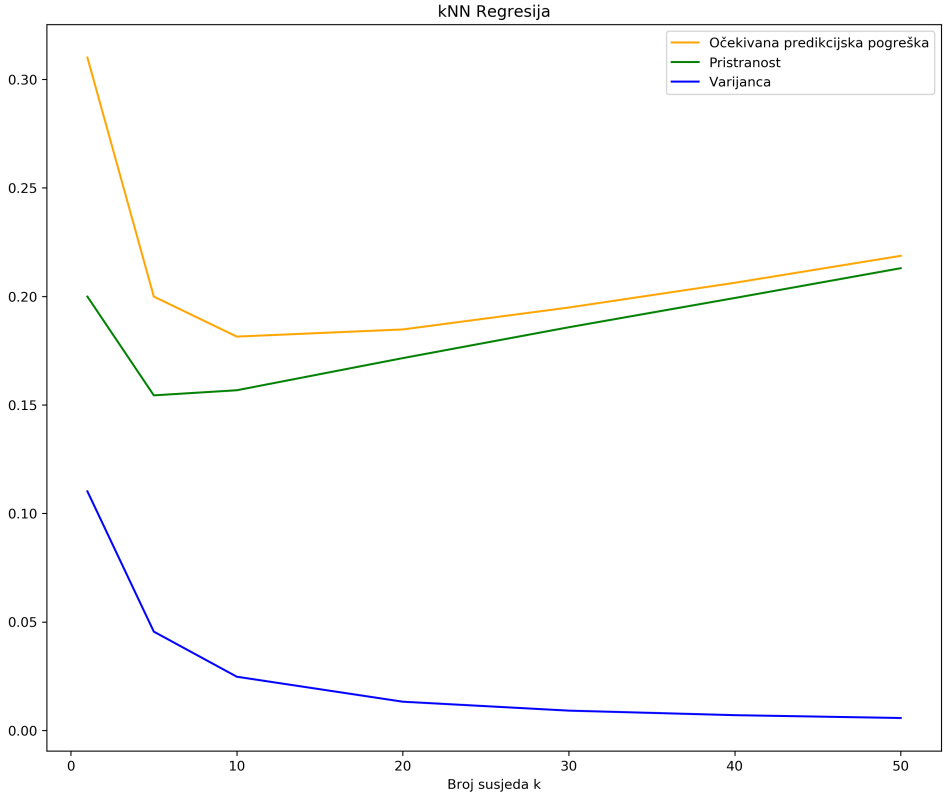
Kod metode k - najbližih susjeda, parametar k je povezan sa kompleksnošću modela na način da manji k označava veću kompleksnost. Po jednakosti (2.30) možemo opaziti nekoliko stvari:

- Za mali parametar k, model $\hat{f}_k$ u stanju je bolje adaptirati se pravoj funkciji f
- Povećanje parametra k vodi k povećanju pristranosti i smanjenju varijance modela
- Smanjenje parametra k vodi k povećanju varijance i smanjenju pristranosti modela

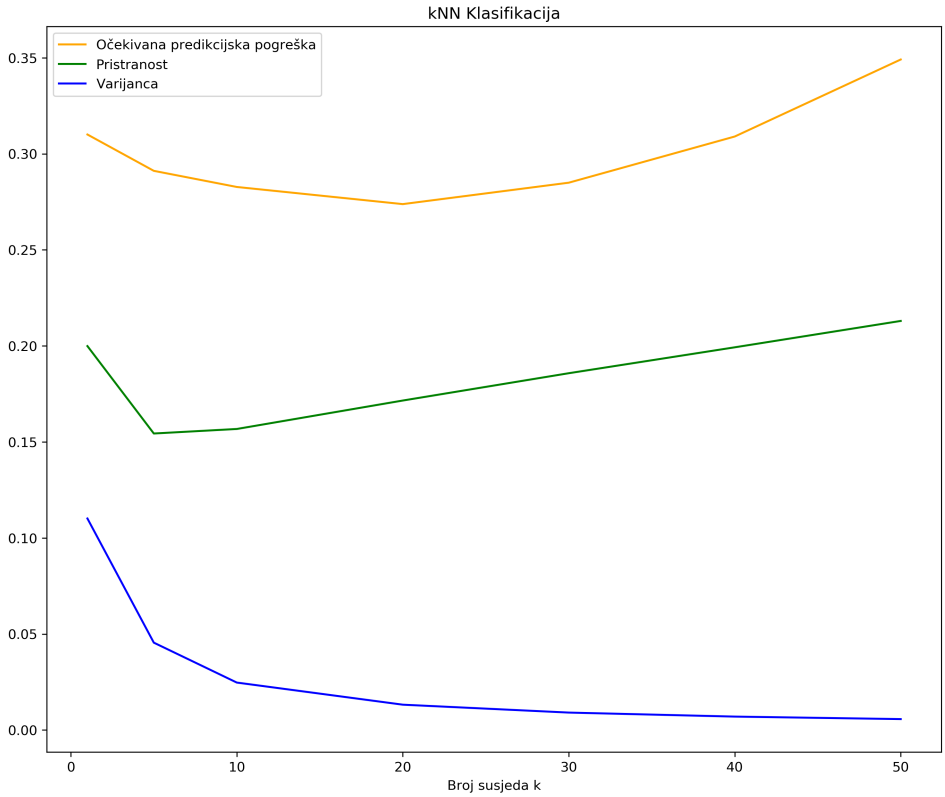
2.3 Primjeri

U ovom potpoglavlju analizirati ćemo odnos varijance-pristranosti u dva simulirana primjera.

Primjer 2.3.1. *Imamo da je* $N = 80$*, te* 20 *nezavisnih varijabli, u oznaci* $X_1, X_2, \cdots, X_{20}$*, pri čemu vrijedi* $X_i \sim U(0,1)$*,* $i = 1, 2, \cdots, 20$*. Zavisna varijabla* Y *poprima vrijednost* 0 *u slučaju* $X_1 \leq 1/2$*, te vrijednost* 1 *u slučaju* $X_1 > 1/2$*. Prva metoda koju koristimo nad podacima jest regresijka metoda k - najbližih susjeda pri čemu je tip funkcije gubitka jednak kvadratnoj grešci (2.1). Drugi slučaj je klasifikacijska metoda k - najbližih susjeda pri čemu je tip funkcije gubitka jednak* $0-1$ *gubitku (2.6). Prilikom ocjena modela različitih kompleksnosti, korišten je velik nezavisan testni skup podataka.*



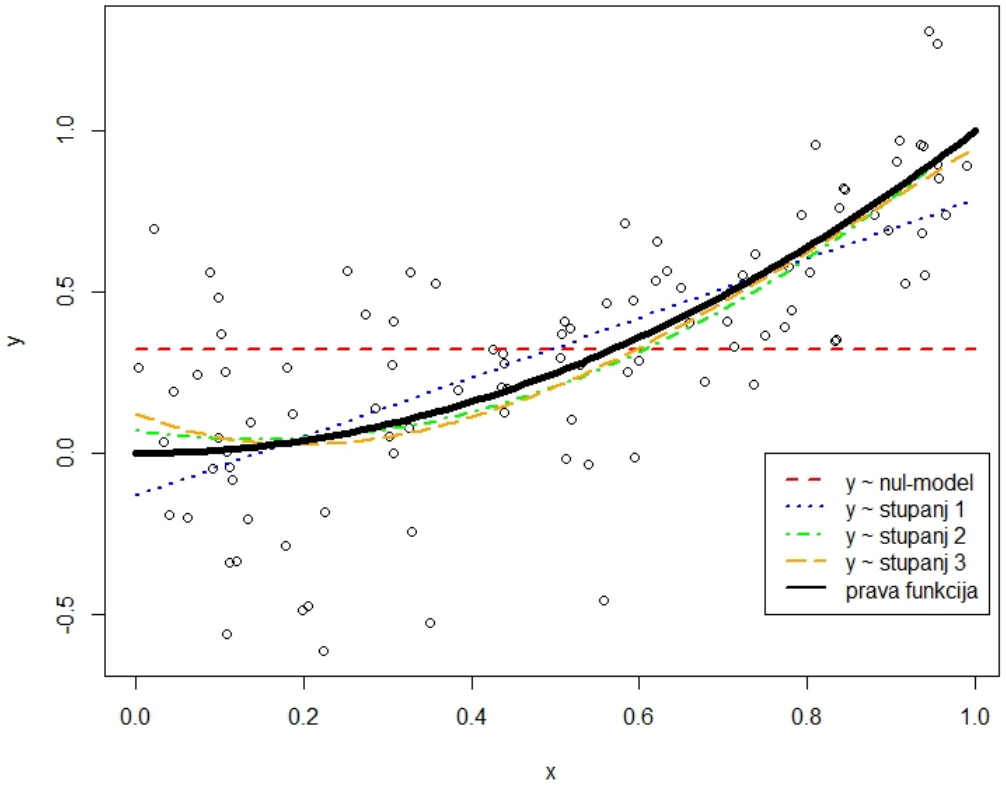
Slika 2.2: k - najbližih susjeda regresija. Vidimo kako je zbroj pristranosti i varijance približno jednak očekivanoj predikcijskoj pogrešci. Minimalna očekivana predikcijska pogreška postiže se za $k = 10$.



Slika 2.3: k - najbližih susjeda klasifikacija. Minimalna očekivana predikcijska pogreška postiže se za $k = 20$.

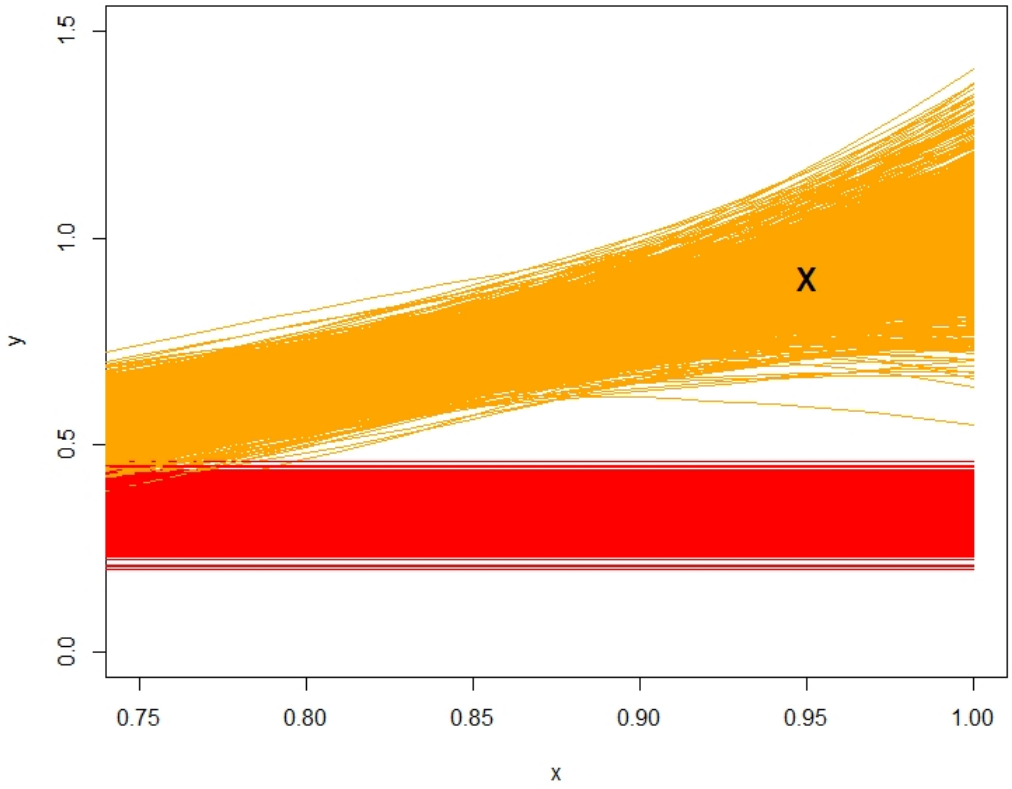
Ako usporedimo krivulje sa Slike 2.2 i Slike 2.3, vidimo da su krivulje pristranosti i varijance jednake. Međutim, kod klasifikacije i $0-1$ *gubitka, predikcijska pogreška se odnosi na udio pogrešno klasificiranih točaka. Time krivulje očekivanih predikcijskih pogrešaka nisu jednake u regresiji i klasifikaciji. Dakle, možemo zaključiti kako izbor drukčije funkcije gubitka dovodi do drukčijeg ponašanja odnosa varijance-pristranosti, te kao što smo vidjeli u ovom primjeru, do različitih procjena optimalnih vrijednosti za parametre od interesa, u ovom slučaju parametar k.*

Primjer 2.3.2. *Razmotrimo odnos varijance-pristranosti na jednostavnom slučaju polinomijalne regresije. Neka vrijedi* $Y = X^2 + \varepsilon$*, pri čemu je* $\varepsilon \sim N(\mu = 0, \sigma^2 = 0.3^2)$*. Generirane podatke uz* $N = 100$*, te četiri istrenirana modela do trećeg stupnja možemo vidjeti na idućoj Slici 2.4.*



Slika 2.4: Skup podataka za treniranje uz istrenirana četiri modela različith kompleksnosti označenih različitim bojama. Također, označena je i prava funkcija koja je podatke generirala.

Usporedimo sada odnos varijance-pristranosti za dva modela (nul-model bez nezavisnih varijabli i najkompleksniji model u našoj simulaciji, model stupnja tri) u točki $x_0 = 0.95$*. Na Slici 2.5 vidimo po tisuću istreniranih modela od oba tipa.*

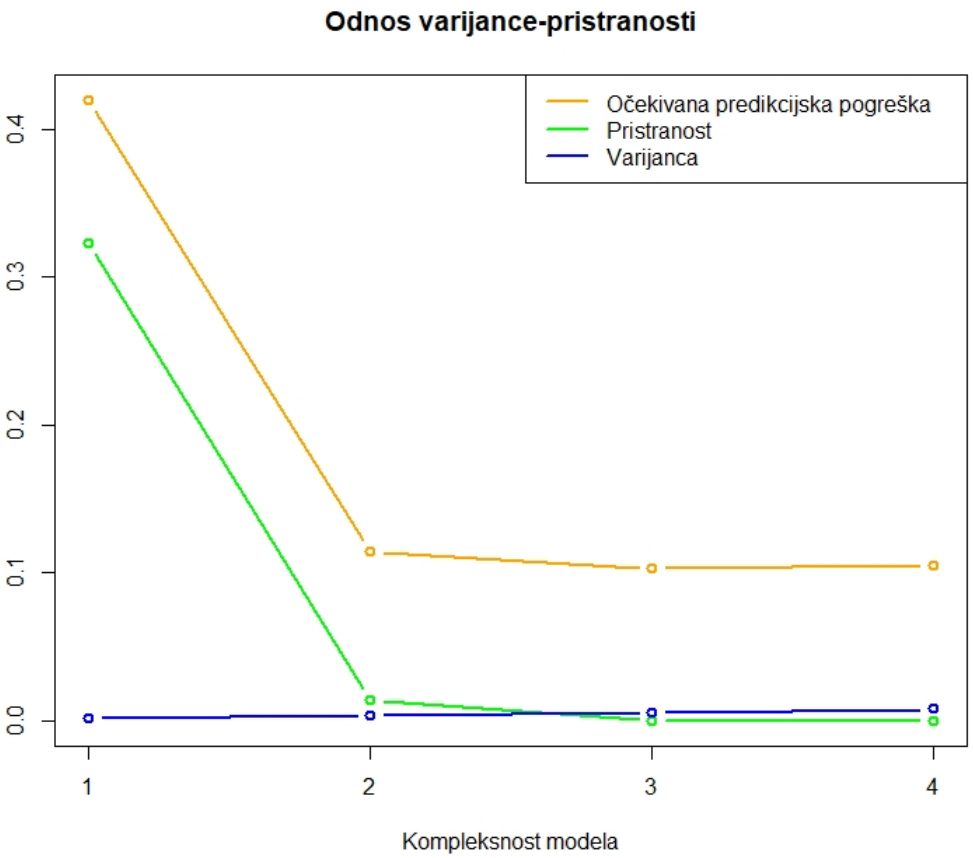


Slika 2.5: Crni "X" označava vrijednost prave funkcije u točki $x_0 = 0.95$. Vidimo da crvene linije istreniraih nul-modela u prosjeku daju krive predikcije sa određenom varijabilnošću. S druge strane, narančaste linije istreniranih modela stupnja 3 u prosjeku daju precizne predikcije, međutim sa većom varijabilnošću.

	Pristranost_kvadrat	Varijanca	Var_sl_pogr	suma	Err
1	0.322916033435613	0.00178403405874618	0.0962743701549573	0.420974437649317	0.420141073350847
2	0.013679376410943	0.00363549071189737	0.0962743701549573	0.113589237277798	0.11451590325724
3	3.63393721302144e-06	0.00581777261170364	0.0962743701549573	0.102095776703874	0.103129366204192
4	8.76603698610023e-07	0.00799060514554581	0.0962743701549573	0.104265851904202	0.10535988102352

Slika 2.6: Tablica vrijednosti za četiri modela u točki $x_0 = 0.95$

Sada možemo procijeniti varijancu i pristranost za sva četiri modela. Rezultati su prikazani na Slici 2.6 i Slici 2.7.



Slika 2.7: Prikaz odnosa varijance-pristranosti za sva četiri modela.

Po izloženom dobivamo rezultate kakve smo i očekivali. Porast kompleksnosti modela doveo je do smanjenja pristranosti i povećanja varijance. Vidimo da i nakon modela stupnja 2 dolazi do povećanja vrijednosti procijenjene očekivane predikcijske pogreške. Također suma triju komponenti: procijenjena pristranost, procijenjena varijanca te procijenjenja varijanca slučajne pogreške, je približno jednaka procijenjenoj očekivanoj predikcijskoj pogrešci.

Poglavlje 3

Analitičke metode za aproksimaciju unutar-uzoračke pogreške

3.1 Uvod

Kada imamo neki skup podataka pred nama i želimo pronaći statistički model koji dane podatke opisuje ili predviđa najbolje, tada nailazimo na dva problema: odabir modela i vrednovanje modela. Ako je naš skup podataka dovoljno velik, možemo ga na slučajan način podjeliti u tri dijela: skup za trening, validacijski skup i testni skup. Skup za trening služi za treniranje modela, validacijski skup koristimo za odabir modela, te naposljetku testni skup za vrednovanje konačno izabranog modela. Ne postoji općenito pravilo za veličinu triju spomenutih skupova, jer to ovisi o samoj prirodi podataka i o početnoj veličini skupa, no općenita podjela može biti: 50% skup za trening, 25% validacijski skup i 25% testni skup. Međutim, u praksi često nećemo posjedovati dovoljno veliki skup podataka da problem rješimo na gore opisani način. Razlozi mogu biti svakojaki, od nedostupnosti podataka do skupoće u dobavljanju novih podataka. U tom slučaju, problemu pristupamo na drukčiji način. Jedan od mogućih rješenja problema odabira modela leži u analitičkim metodama *AIC* i *BIC* kojima aproksimiramo validacijski korak.

3.2 Optimizam pogreške treniranja i unutar-uzoračka pogreška

Kao što smo spomenuli na stranici 7., pogreška treniranja $\overline{err}$ nije dobar procjenitelj testne pogreške Err_τ. Budući da se treniranjem model prilagodi skupu podataka za trening, pogreška treniranja $\overline{err}$ je *preoptimističan* procjenitelj testne pogreške Err_τ. Da je to uistinu tako, pokažimo na sljedećem primjeru.

Primjer 3.2.1. *Pretpostavimo da imamo model linearne regresije s p parametara, procijenjen metodom najmanjih kvadrata iz skupa podataka za trening* $(x_1, y_1), \cdots, (x_N, y_N)$*, pri čemu su podaci za trening dobiveni na slučajan način iz populacije. Neka* $\hat{\beta}$ *predstavlja procjenitelj dobiven metodom najmanjih kvadrata. Neka imamo i testni skup podataka* $(\widetilde{x}_1, \widetilde{y}_1), \cdots, (\widetilde{x}_M, \widetilde{y}_M)$ *dobiven na slučajan način iz iste populacije kao i skup podataka za trening. Pokažimo da vrijedi*

$$\mathbb{E}[R_{tr}(\hat{\beta})] \leq \mathbb{E}[R_{te}(\hat{\beta})]\,, \tag{3.1}$$

pri čemu su $R_{tr}(\beta) = \frac{1}{N}\sum_1^N (y_i - \beta^T x_i)^2$ *i* $R_{te}(\beta) = \frac{1}{M}\sum_1^M (\widetilde{y}_i - \beta^T \widetilde{x}_i)^2$ *.*

Neka $\hat{\beta}_{tr}$ *predstavlja procjenitelj najmanjih kvadrata iz skupa podataka za trening, te slično* $\hat{\beta}_{te}$ *procjenitelj najmanjih kvadrata iz testnog skupa podataka.*
Imamo da vrijedi:

$$R_{te}(\hat{\beta}_{tr}) = \frac{1}{M}\sum_{i=1}^{M} (\widetilde{y}_i - \hat{\beta}_{tr}{}^T \widetilde{x}_i)^2 \tag{3.2}$$

$$\geq \frac{1}{M}\sum_{i=1}^{M} (\widetilde{y}_i - \hat{\beta}_{te}{}^T \widetilde{x}_i)^2\,, \tag{3.3}$$

Dakle,

$$\mathbb{E}[R_{te}(\hat{\beta}_{tr})] \geq \mathbb{E}[\frac{1}{M}\sum_{i=1}^{M} (\widetilde{y}_i - \hat{\beta}_{te}{}^T \widetilde{x}_i)^2] \tag{3.4}$$

$$= \mathbb{E}[(\widetilde{y}_i - \hat{\beta}_{te}{}^T \widetilde{x}_i)^2] \tag{3.5}$$

$$= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} (\widetilde{y}_i - \hat{\beta}_{te}{}^T \widetilde{x}_i)^2]\,, \tag{3.6}$$

pri čemu jednakosti (3.5) i (3.6) vrijede jer su podaci iz testnog skupa dobiveni na slučajan način. Budući da N testnih podataka dolazi iz iste populacije kao i N podataka za trening, vrijedi:

$$\mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} (\widetilde{y}_i - \hat{\beta}_{te}{}^T \widetilde{x}_i)^2] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{\beta}_{tr}{}^T x_i)^2] \tag{3.7}$$

$$= \mathbb{E}[R_{tr}(\hat{\beta}_{tr})] \tag{3.8}$$

Dakle, dobili smo što smo i htjeli:

$$\mathbb{E}[R_{te}(\hat{\beta}_{tr})] \geq \mathbb{E}[R_{tr}(\hat{\beta}_{tr})] \tag{3.9}$$

Budući da se nezavisni vektori testnih podataka ne moraju poklapati s nezavisnim vektorima podataka za trening, vrijednost Err_τ možemo smatrati *izvan-uzoračkom* pogreškom. S druge strane, *unutar-uzoračku* pogrešku Err_{in} definiramo kao:

$$Err_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_Y[L(Y_i, \hat{f}(x_i)) \mid \tau]\,, \tag{3.10}$$

pri čemu slučajnost dolazi samo od Y_i, odnosno x_i su fiksne vrijednosti nezavisnih vektora jednake podacima za trening, a Y_i su nova opažanja varijabli odziva u tim točkama. Dakle, vrijednosti Y_i ne moraju biti jednake vrijednostima y_i iz skupa podataka za trening.

Sada definiramo *optimizam* kao:

$$op = Err_{in} - \overline{err}\,. \tag{3.11}$$

Kako je $\overline{err}$ pristran kao procjenitelj predikcijske pogreške, to je op najčešće pozitivan. Nadalje, prosječni optimizam, u oznaci ω, definiramo kao:

$$\omega = \mathbb{E}_y[op]\,, \tag{3.12}$$

gdje kao i prije imamo da su vrijednosti varijabli prediktora fiksne, a očekivanje je samo obzirom na vrijednosti varijabli odziva.
Sljedeća jednakost vrijedi za općeniti slučaj funkcije gubitka:

$$\omega = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)\,, \tag{3.13}$$

pri čemu Cov označava kovarijancu.

Pokažimo da jednakost (3.13) vrijedi u slučaju kada je funkcija gubitka jednaka kvadratnoj pogrešci.

$$\begin{aligned}
\omega &= \mathbb{E}_y[op] = \mathbb{E}_y[Err_{in} - \overline{err}] \\
&= \mathbb{E}_y[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_Y[(Y_i - f(x_i) + f(x_i) - \mathbb{E}[\hat{f}(x_i)] + \mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))^2] \\
&- \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i) + f(x_i) - \mathbb{E}[\hat{f}(x_i)] + \mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))^2]
\end{aligned} \tag{3.14}$$

U jednakosti (3.14) smo nakon uvrštavanja izraza Err_{in} i $\overline{err}$ dodali i oduzeli izraze $f(x_i)$ i $\mathbb{E}[\hat{f}(x_i)]$. Nakon grupiranja slijedi,

$$
\begin{aligned}
(3.14) = \mathbb{E}_y[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_Y[&\overbrace{(Y_i - f(x_i))^2}^{A_1} + \overbrace{(f(x_i) - \mathbb{E}[\hat{f}(x_i)])^2}^{B_1} + \overbrace{(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))^2}^{C_1} \\
&+ \overbrace{2(Y_i - f(x_i))(f(x_i) - \mathbb{E}[\hat{f}(x_i)])}^{D_1} + \overbrace{2(Y_i - f(x_i))(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))}^{E_1} \\
&+ \overbrace{2(f(x_i) - \mathbb{E}[\hat{f}(x_i)])(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))}^{F_1}] \\
&- \frac{1}{N}\sum_{i=1}^{N}(\overbrace{(y_i - f(x_i))^2}^{A_2} + \overbrace{(f(x_i) - \mathbb{E}[\hat{f}(x_i)])^2}^{B_2} + \overbrace{(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))^2}^{C_2} \\
&+ \overbrace{2(y_i - f(x_i))(f(x_i) - \mathbb{E}[\hat{f}(x_i)])}^{D_2} + \overbrace{2(y_i - f(x_i))(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))}^{E_2} \\
&+ \overbrace{2(f(x_i) - \mathbb{E}[\hat{f}(x_i)])(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))}^{F_2})] \qquad (3.15) \\
= \frac{1}{N}\sum_{i=1}^{N}&(\mathbb{E}_y[\mathbb{E}_Y[A_1] - A_2] + \mathbb{E}_y[\mathbb{E}_Y[B_1] - B_2] + \mathbb{E}_y[\mathbb{E}_Y[C_1] - C_2] \\
&+ \mathbb{E}_y[\mathbb{E}_Y[D_1] - D_2] + E_y[\mathbb{E}_Y[E_1] - E_2] + E_y[\mathbb{E}_Y[F_1] - F_2]) \qquad (3.16)
\end{aligned}
$$

Lako se može pokazati kako su svi izrazi unutar sume u (3.16) jednaki nuli, osim izraza $E_y[\mathbb{E}_Y[E_1] - E_2]$. Njega možemo dodatno raspisati:

$$
\begin{aligned}
E_y[\mathbb{E}_Y[E_1] - E_2] &= \mathbb{E}_y[\mathbb{E}_Y[2(Y_i - f(x_i)(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))] \\
&\quad - 2(y_i - f(x_i)(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))] && (3.17)\\
&= -2\,\mathbb{E}_y[y_i - f(x_i)(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))] && (3.18)\\
&= 2\,\mathbb{E}_y[(y_i - \mathbb{E}[y_i])(\hat{f}(x_i) - \mathbb{E}[\hat{f}(x_i)])] && (3.19)\\
&= 2Cov(y_i, \hat{y}_i)\,, && (3.20)
\end{aligned}
$$

pri čemu smo u (3.18) koristili nezavisnost izraza $(Y_i - f(x_i)$ i $(\mathbb{E}[\hat{f}(x_i)] - \hat{f}(x_i))$ uz činjenicu $\mathbb{E}_Y[Y_i] = f(x_i)$. Uvrštavajući (3.21) nazad u (3.16) konačno dobivamo:

$$\omega = \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)\,,$$

a to smo i htjeli pokazati.

Relacija (3.13) nam ustvari govori da što jače prilagodimo model skupu podataka za trening, to je $Cov(\hat{y}_i, y_i)$ veći, a zbog toga je posljedično i optimizam veći.

Uz (3.13), djelovanjem s matematičkim očekivanjem na jednakost (3.11) dolazimo do:

$$\mathbb{E}_y[Err_{in}] = \mathbb{E}_y[\overline{err}] + \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) \tag{3.21}$$

Jednakost (3.21) se može dodatno pojednostaviti uz određene pretpostavke. Na primjer, ako imamo kao i ranije da je $Y = f(x) + \varepsilon$, uz $\mathbb{E}[\varepsilon] = 0, \sigma_\varepsilon^2 = Var(\varepsilon)$ te pretpostavimo linearni regresijski model s $d-1$ nezavisnom varijablom, odnosno zapisano u matričnom obliku $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, pri čemu su:

$$\begin{aligned}
\mathbf{Y} &= (y_1, y_2, \cdots, y_N)^T \in \mathbb{R}^N \\
\mathbf{X} &= (x_1^T, x_2^T, \cdots, x_N^T)^T \in M_{N,d}(\mathbb{R}) \\
\varepsilon &= (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N)^T \in \mathbb{R}^N \\
\beta &= (\beta_0, \beta_1, \cdots, \beta_{d-1})^T \in \mathbb{R}^d,
\end{aligned}$$

Znamo da je tada $\hat{\mathbf{Y}} = \mathbf{HY}$, pri čemu su $\hat{\mathbf{Y}}$ vektor predikcija i $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ ortogonalni projektor. Tada slijedi:

$$Cov(\hat{\mathbf{Y}}, \mathbf{Y}) = Cov(\mathbf{HY}, \mathbf{Y}) = \mathbf{H}Cov(\mathbf{Y}, \mathbf{Y}) = \mathbf{H}\sigma_\varepsilon^2 I_{NxN} = \sigma_\varepsilon^2 \mathbf{H}$$

Pretpostavkom da je $\mathbf{X}$ punog ranga, odnosno $r(\mathbf{X}) = d$, te koristeći komutativnost operatora traga matrice *tr* dobivamo:

$$tr(\mathbf{H}) = tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = tr(\mathbf{X}(\mathbf{X}^T\mathbf{X}^T\mathbf{X})^{-1}) = tr(I_d) = d$$

Sada jednostavno slijedi:

$$\begin{aligned}
\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) &= tr(\mathbf{H})\sigma_\varepsilon^2 \\
&= d\sigma_\varepsilon^2
\end{aligned} \tag{3.22}$$

Dakle, u slučaju linearnog regresijskog modela, relacija (3.21) poprima oblik:

$$\mathbb{E}_y[Err_{in}] = \mathbb{E}_y[\overline{err}] + 2\frac{d}{N}\sigma_\varepsilon^2 \tag{3.23}$$

Iz relacije (3.23) možemo vidjeti da se optimizam povećava linearno sa brojem nezavisnih varijabli, odnosno što je model kompleksniji to je optimizam veći. Također, optimizam opada što je veličina skupa podataka za treniranje veća.

Jedan način na koji bismo mogli indirektno procijeniti predikcijsku pogrešku je preko procijenitelja unutar-uzoračke pogreške Err_{in} tako da procijenimo optimizam i zbrojimo ga s pogreškom treniranja $\overline{err}$, odnosno

$$\widehat{Err}_{in} = \overline{err} + \hat{\omega}\,. \tag{3.24}$$

Tehnike koje ćemo proučavati u idućem potpoglavlju upravo to i rade, no za nešto užu klasu procjenitelja koji su linearni u parametrima. Za razliku od njih, tehnike kojima ćemo se baviti u idućem poglavlju će neposredno procjenjivati izvan-uzoračku pogrešku Err. Te tehnike su primjenjive u svim situacijama, odnosno bez obzira o kojoj se funkciji gubitka radilo te bio li procjenitelj linearan ili nelinearan. Iako sami procjenitelj unutar-uzoračke pogreške nije od velikog interesa u vidu vrednovanja modela jer ne očekujemo da će se vrijednosti prediktora u testnom skupu podataka podudarati sa vrijednostima unutar skupa podataka za trening, pokazat će se da je ipak efektivan glede odabira modela.

3.3 AIC i BIC

Informacijski kriteriji predstavljaju skupinu statističkih alata za rješavanje problema odabira modela. Oni u mjerenju kakvoće modela u obzir uzimaju mjeru u kojoj model same podatke opisuje te njegovu kompleksnost. Postoji mnogo informacijskih kriterija, a mi ćemo obraditi dva najpoznatija. To su Akaike Informacijski Kriterij (skraćno AIC) i Bayesov Informacijski Kriterij (skraćeno BIC). Iako ovdje nećemo iznositi puni izvod formula za AIC i BIC, navest ćemo informacijski kriterij koji je temelj većini ostalih informacijskih kriterija, te obrazložit ćemo dovoljno tvrdnji za dobivanje opće intuicije iza kriterija AIC i BIC. Stoga krenimo redom.

Kullback-Leibler informacija (skraćeno KL informacija) mjeri količinu izgubljene informacije kada funkciju f aproksimiramo funkcijom g, ili drugim rječnikom, udaljenost funkcija f i g. Matematička definicija glasi:

$$I(f, g) = \int f(x) \log(\frac{f(x)}{g(x)})\, dx\,. \tag{3.25}$$

U kontekstu statističkog učenja, funkcija f reprezentira pravu funkciju koja je podatke generirala, dok funkcija g reprezentira aproksimirajući model. Jasno je kako bismo htjeli pronaći model g koji će dovesti do minimalnog gubitka informacija, odnosno koji će minimizirati $I(f, g)$.
Važan pojam kod obje metode, AIC i BIC, je pojam maksimalne vjerodostojnosti kojeg se ovdje prisjećamo. Neka imamo uzorak od N opažanja $y_1, y_2, \cdots, y_N$ slučajnih varijabli (u kontekstu statističkog učenja, zavisnih varijabli) $Y_1, Y_2, \cdots, Y_N$, i neka je $f(Y_1, Y_2, \cdots, Y_N \mid$

θ) funkcija gustoće koja ovisi o skupu parametara θ. Tada je funkcija vjerodostojnosti $\mathcal{L}$ definirana kao:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}(\theta \mid y_1, y_2, \cdots, y_N) \\ &= f(y_1, y_2, \cdots, y_N \mid \theta)\,. \end{aligned} \tag{3.26}$$

Metodom maksimalne vjerodostojnosti procijenjujemo skup parametara θ za koji je vrijednost s desne strane jednakosti (3.26) najveća, odnosno skup parametara θ koji je najvjerodostojniji obzirom na dane podatke.

Pogledajmo sada koja je veza između KL informacije i maksimalne vjerodostojnosti. Ako uzmemo da unutar (3.25) vrijedi $g(x) = g(x \mid \theta)$, pri čemu je θ skup parametara procijenjen iz podataka, tada (3.25) možemo dalje zapisati kao:

$$\int f(x) \log(\frac{f(x)}{g(x \mid \theta)})\, dx = \int f(x) \log(f(x))\, dx - \int f(x) \log(g(x \mid \theta))\, dx \tag{3.27}$$

$$= \mathbb{E}_f[\log(f(x))] - \mathbb{E}_f[\log(g(x \mid \theta))] \tag{3.28}$$

$$= \textit{konstanta} - \mathbb{E}_f[\log(g(x \mid \theta))] \tag{3.29}$$

Iz (3.29) vidimo da je gubitak informacije, odnosno KL udaljenost obrnuto proporcionalna s procjeniteljem maksimalne vjerodostojnosti, tj. što je model vjerojatniji obzirom na opažene podatke, to je KL udaljenost manja. To povlači da ako imamo dva ili više modela, mi možemo računati njihovu relativnu udaljenost obzirom na pravu funkciju na temelju njihovih procjenitelja maksimalne vjerodostojnosti.

Sada imamo sve potrebno za iznijeti forumulu za AIC nekog statističkog modela:

$$AIC = -2 \log(\mathcal{L}(\hat{\theta} \mid y)) + 2d \tag{3.30}$$

pri čemu je $\log(\mathcal{L}(\hat{\theta} \mid y))$ maksimalna vrijednost log-vjerodostojnosti, a d je broj parametara modela.

Razlog zbog kojeg ne možemo koristiti KL informaciju neposredno jest taj što u praksi ne znamo pravu funkciju f. Iako mi ne možemo izračunati apsolutnu udaljenost između modela g i prave funkcije f, AIC daje procjenjenu očekivanu relativnu udaljenost između modela g i prave funckije f, tj. AIC predstavlja procjenu KL informacije. Stoga, ako imamo više modela, najpoželjniji model jest onaj čiji je AIC najmanji. Kao što vidimo iz jednakosti (3.30), AIC kažnjava kompleksnije modele, te time nastoji spriječiti pretreniranost, odnosno modele s nepoželjnim svojstvima niske pristranosti i visoke varijance, kao što je opisano u drugom poglavlju.

Za razliku od AIC, BIC nije motiviran od strane KL informacije, već je motiviran kroz bayesovski pristup problemu odabira modela. Označimo sa $\mathbb{P}(M_i)$ *apriornu* vjerojatnost modela i, te sa $f(y)$ apriornu gustoću opaženih podataka y. Potom definiramo $f(y \mid M_i)$ kao gustoću podataka od y kao da su generirani od strane modela i. Koristeći Bayesov teorem dolazimo do *aposteriorne* vjerojatnosti $\mathbb{P}(M_i \mid y)$:

$$\mathbb{P}(M_i \mid y) = \frac{f(y \mid M_i) \cdot \mathbb{P}(M_i)}{f(y)}$$

Ako imamo dva potencijalna modela M_i i M_j pored podataka y, način na koji odabrati jednog od njih bi bio da analiziramo omjer njihovih aposteriornih vjerojatnosti. Općenito, kada imamo familiju više modela, uzimamo da su njihove apriorne vjerojatnosti jednake. Sada definiramo Bayesov faktor $BF(i, j)$:

$$\begin{aligned} BF(i, j) &= \frac{\mathbb{P}(M_i \mid y)}{\mathbb{P}(M_j \mid y)} \\ &= \frac{\frac{f(y|M_i)\cdot\mathbb{P}(M_i)}{f(y)}}{\frac{f(y|M_j)\cdot\mathbb{P}(M_j)}{f(y)}} \\ &= \frac{f(y \mid M_i)}{f(y \mid M_j)} \end{aligned} \tag{3.31}$$

Ako vrijedi $BF(i, j) > 1$, onda je model M_i vjerojatniji od modela M_j s obzirom na opažene podatke y, te u tom slučaju odabiremo model M_i umjesto modela M_j. Ako vrijedi $BF(i, j) < 1$, situacija je suprotna, te odabiremo model M_j.

Sada možemo iznijeti opći oblik formule za BIC:

$$BIC = -2\log(\mathcal{L}(\hat{\theta} \mid y)) + d\log(N)\,, \tag{3.32}$$

pri čemu je N veličina uzorka.

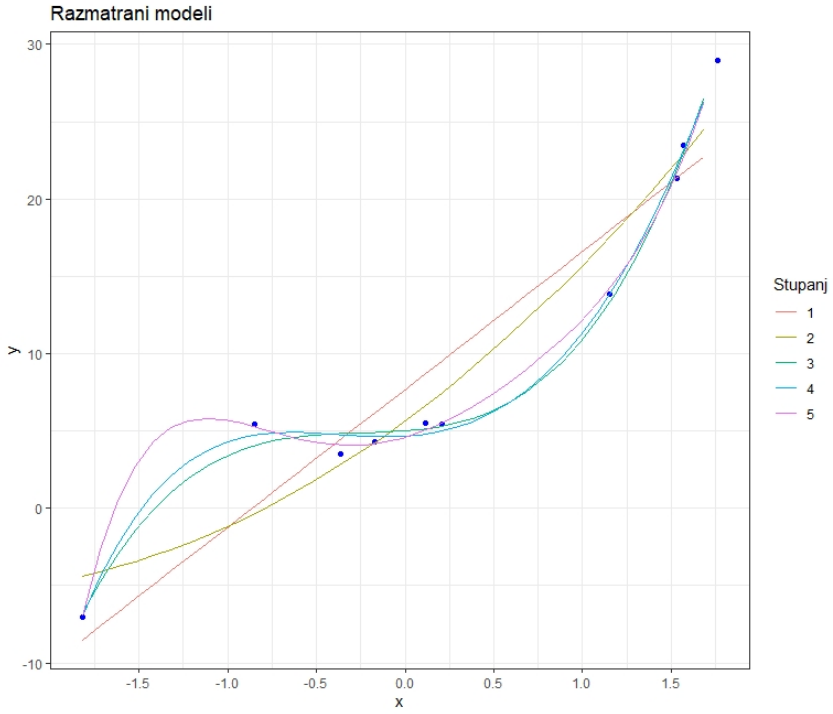
Iz relacije (3.32) vidimo da za $N \geq 8$, BIC strože kažnjava kompleksnije modele od AIC, te time daje prednost jednostavnijim modelima. Kao i u slučaju AIC, model s najmanjom vrijednosti BIC je najbolji izbor jer on ujedno ima i najveću aposteriornu vjerojatnost. Štoviše, ako uz podatke y imamo skup od K modela, te izračunamo BIC svakog od njih, u oznaci $BIC_k, k = 1, 2, \cdots, K$, slijedi da je aposteriorna vjerojatnost svakog od modela M_k:

$$\mathbb{P}(M_k \mid y) = \frac{e^{-\frac{1}{2}BIC_k}}{\sum_i^K e^{-\frac{1}{2}BIC_i}} \tag{3.33}$$

Spomenimo i neka svojstva kriterija AIC i BIC, te neke njihove međusobne razlike. Za razliku od AIC, BIC posjeduje svojstvo *konzistentnosti* koje govori da se vjerojatnost odabira pravog modela, ukoliko se takav nalazi u skupu razmatranih modela, približava jedan kada veličina uzorka teži k beskonačnosti. Za razliku od BIC, AIC naginje kompleksnijim modelima jer, kao što smo vidjeli, manje kažnjava kompleksnost modela od BIC. Dakle, sami performans kriterija AIC i BIC ovisi o konkretnoj situaciji i veličini uzorka. U idućem potpoglavlju iznijeti ćemo jedan primjer usporedbe kriterija AIC i BIC.

3.4 Primjeri

Primjer 3.4.1. *Neka je prava funkcija f dana formulom $f(x) = x + 2x^2 + 3x^3 + 5$, te neka su $X \sim U(-2, 2)$ i $\varepsilon \sim N(0, 1)$. Radimo simulaciju na sljedeći način: za različite veličine uzorka N, usporediti ćemo rezultate kriterija AIC i BIC primjenjenih na skup aproksimirajućih modela do petog stupnja. Uz vrijednosti samih kriterija AIC i BIC, ispisati ćemo i aposteriorne vjerojatnosti u slučaju BIC kriterija, koristeći relaciju (3.33).*



Slika 3.1: Prikaz podataka i svih 5 modela istreniranih na njima za $N = 10$. Uočavamo visoku pristranost modela stupnja jedan i stupnja dva, što je i za očekivati.

	kriterij	vrijednost	vjerojatnost	stupanj
1	AIC	58.61576	Neprimjenjivo	1
2	AIC	56.28307	Neprimjenjivo	2
3	AIC	31.96591	Neprimjenjivo	3
4	AIC	30.99038	Neprimjenjivo	4
5	AIC	24.13145	Neprimjenjivo	5
6	BIC	59.52351	5.59054689486579e-08	1
7	BIC	57.49341	1.54271796547036e-07	2
8	BIC	33.47883	0.0252920750033809	3
9	BIC	32.80589	0.0354089259115885	4
10	BIC	26.24955	0.939298788907765	5

Slika 3.2: Slučaj $N = 10$. Vidimo da je AIC izabrao model petog stupnja kao najbolji, a isti rezultat dao je i BIC. Očito je da uzorak nije dovoljno velik da informacijski kriteriji poluče dobre rezultate.

	kriterij	vrijednost	vjerojatnost	stupanj
1	AIC	585.4692	Neprimjenjivo	1
2	AIC	553.2624	Neprimjenjivo	2
3	AIC	283.3380	Neprimjenjivo	3
4	AIC	285.0087	Neprimjenjivo	4
5	AIC	283.0094	Neprimjenjivo	5
6	BIC	593.2848	2.77626474880562e-65	1
7	BIC	563.6831	7.43677580290143e-59	2
8	BIC	296.3638	0.829886603915529	3
9	BIC	300.6397	0.0978428229769431	4
10	BIC	301.2456	0.0722705731075278	5

Slika 3.3: Slučaj $N = 100$. AIC i dalje odabire model petog stupnja kao najbolji, no sa veoma malom prednošću nad modelom trećeg stupnja. BIC ovdje raspoznaje model trećeg stupnja kao najbolji.

Vidimo da je za već prilično malen uzorak BIC raspoznao pravi model kao najbolji. Daljnjim povećanjem uzorka, BIC nastavlja odabirati model stupnja tri kao uvjerljivo najbolji, dok AIC tek nakon nekog vremena odabire model trećeg stupnja, no ne tako uvjerljivo pred modelima četvrtog i petog stupnja.

Poglavlje 4

Unakrsno vrednovanje

4.1 Uvod

Metoda unakrsnog vrednovanja (eng. cross-validation) pripada skupini tzv. metoda ponovnog uzorkovanja. Metode ponovnog uzorkovanja predstavljaju važan i moćan alat u modernoj statistici zbog sve manje računarske ograničenosti zahvaljujući tehnološkom napretku. Kao što i sam naziv kaže, metode ponovnog uzorkovanja koriste početni skup podataka za trening na način da iz njega uzorkuju podatke više puta pomoću kojih potom treniraju modele koje dalje mogu koristiti u različite svrhe, često zbog ocjene preciznosti određenog procjenitelja ili u slučaju unakrsnog vrednovanja, procjene sposobnosti generalizacije modela. Dakle, metodu unakrsnog vrednovanja možemo koristiti u svrhu procedure odabira modela, ili pak za vrednovanje modela.

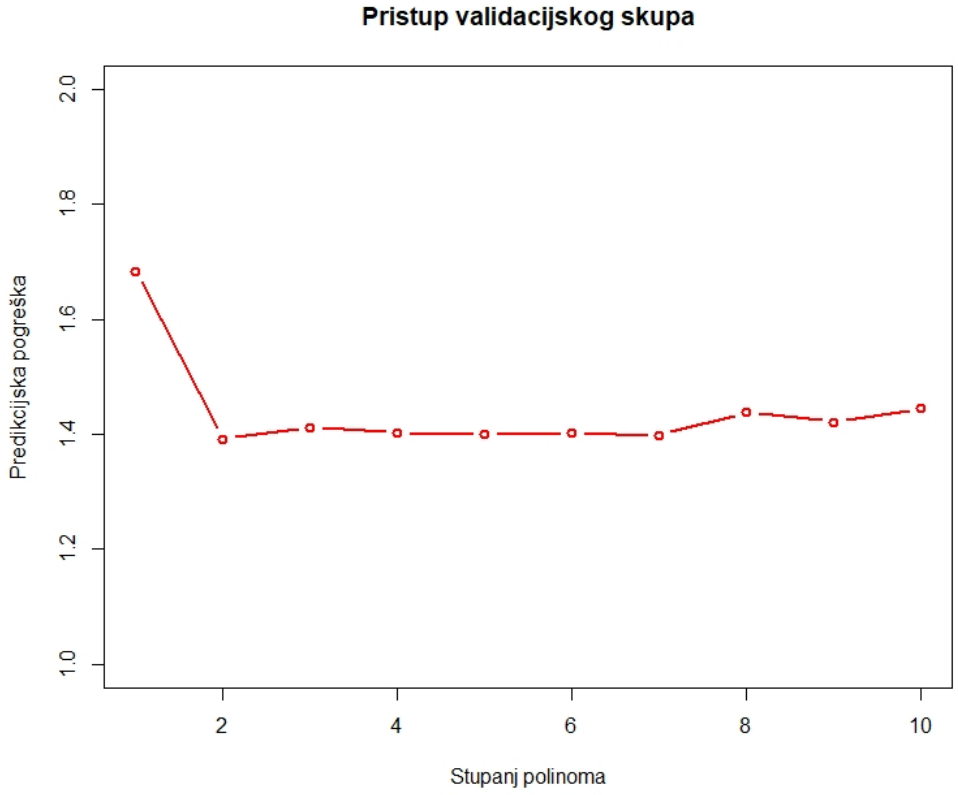
4.2 Opis metode

Na početku je važno napomenuti kako unakrsno vrednovanje možemo koristiti samo u odabiru modela ili u vrednovanju modela, ali ne i u obje procedure. U suprotnom, dolazi do pretreniranosti modela, odnosno do problema visoke varijance. Drugim riječima, konačna evaluacija modela bila bi preoptimistična, što je svakako nepoželjna pojava. Postoji nadogradnja metode unakrsnog vrednovanja, tzv. ugniježđeno unakrsno vrednovanje, kojom se mogu istovremeno riješiti problemi odabira i vrednovanja modela, no tom metodom se nećemo ovdje baviti.

Većina metoda efektivnije procijenjuje očekivanu testnu pogrešku Err, umjesto testne pogreške Err_τ koja je nama od interesa, pa tako i metoda unakrsnog vrednovanja (vidi [3], odjeljak 7.12). Krenimo sada s opisom same metode.

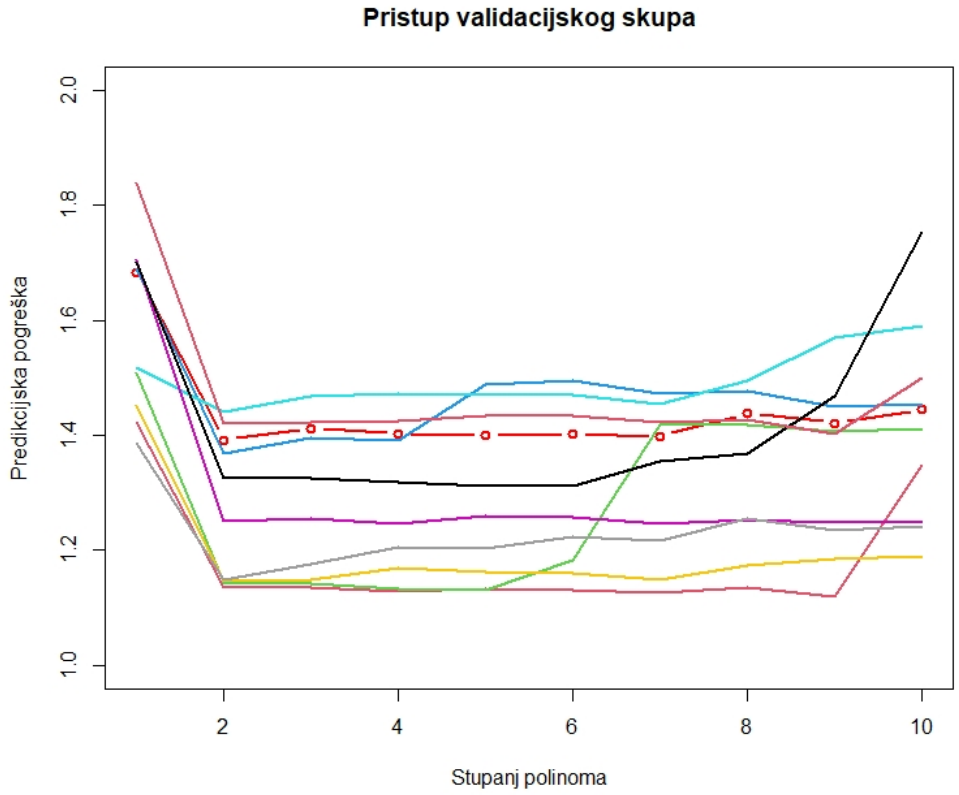
Pretpostavimo prvo da smo u situaciji u kojoj početni skup podataka na slučajan način podijelimo na dva dijela: skup podataka za trening i validacijski skup podataka. Kao i ranije, veličina samih skupova podataka ovisi o broju i prirodi podataka, Nakon podjele, istreniramo model pomoću skupa za trening i procijenimo predikcijsku pogrešku pomoću validacijskog skupa. Upravo opisani pristup je tzv. pristup validacijskog skupa. Međutim, iako jednostavna metoda, pogledajmo neke njene manjkavosti u idućem primjeru.

Primjer 4.2.1. *Opet simuliramo situaciju na sličan način kao i do sada. Neka je* $N = 200$ *i neka je prava funkcija dana formulom* $f(x) = 2x + 0.3x^2 + 4$*, pri čemu su* $X \sim U(-2, 2)$ *i* $\varepsilon \sim N(0, 1)$*. Potom dijelimo skup podataka na dva dijela jednake veličine: skup podataka za treniranje i validacijski skup podataka.*



Slika 4.1: Rezultat pristupa validacijskog skupa na temelju jedne podjele podataka. Vidimo kako je metoda prepoznala da je model polinoma drugog stupnja precizniji od običnog linearnog, te da polinomi višeg stupnja ne pridonose daljnjem povećanju preciznosti.

Možemo se zapitati očekujemo li istu krivulju za svaku moguću podjelu podataka. Rezultati su prikazani na sljedećoj slici.



Slika 4.2: Krivulje na ovoj slici predstavljaju procjenu predikcijske pogreške dobivene na temelju deset različitih podjela početnog skupa podataka. Kao što možemo opaziti, postoji određena varijabilnost između procjena.

Dakle, iako možemo zaključiti kako linearan model nije optimalan u pogledu predikcijske pogreške i kako modeli polinoma stupnja višeg od dva ne pridonose mnogo obzirom na model polinoma stupnja dva, vidimo da ne postoji suglasnost između procjenitelja koji su dobivenih različitom podjelom podataka o tome koji model rezultira najmanjom predikcijskog pogreškom na validacijskom skupu.

Iz dosad viđenog, možemo zaključiti kako pristup validacijskog skupa u situaciji kada imamo mali broj podataka pred sobom ima dva problema:

- Procjenitelj predikcijske pogreške ima visoku varijancu, te ovisi o samoj podjeli podataka
- Samom podjelom podataka smanjujemo broj podataka korištenih za treniranje modela, te budući da model koji je treniran nad manjim skupom podataka ima manju sposobnost generalizacije, konačni procjenitelj dobivem ovom metodom ima veću pristranost.

Metoda unakrsnog vredovanja koju ćemo izložiti u nastavku upravo rješava ova dva problema.

Metoda K-strukog unakrsnog vrednovanja (eng. K-Fold Cross-Validation) radi podjelu podataka u K manjih disjunktnih dijelova otprilike jednakih veličina. Zatim ponavlja isti proces za svaki od tih K dijelova. Konkretno, trenira model na preostalih $K-1$ dijelova, te potom istreniranim modelom računa predikcije na tom jednom izostavljenom dijelu iz procedure treniranja modela, te tako dolazi do jednog od K procjenitelja predikcijske pogreške. U konačnici, kombinira svih K izračunatih procjenitelja predikcijske pogreške.

Neka preslikavanje $\kappa : \{1, 2, \cdots, N\} \to \{1, 2, \cdots, K\}$ označava particiju početnog skupa podataka i neka $\hat{f}^{-k}(x)$ označava istrenirani model na skupu bez k-tog dijela, pri čemu $k = 1, 2, \cdots, K$. Tada imamo da je procjenitelj predikcijske pogreške dobiven metodom unakrsnog vrednovanja, u oznaci $CV(\hat{f})$, jednak:

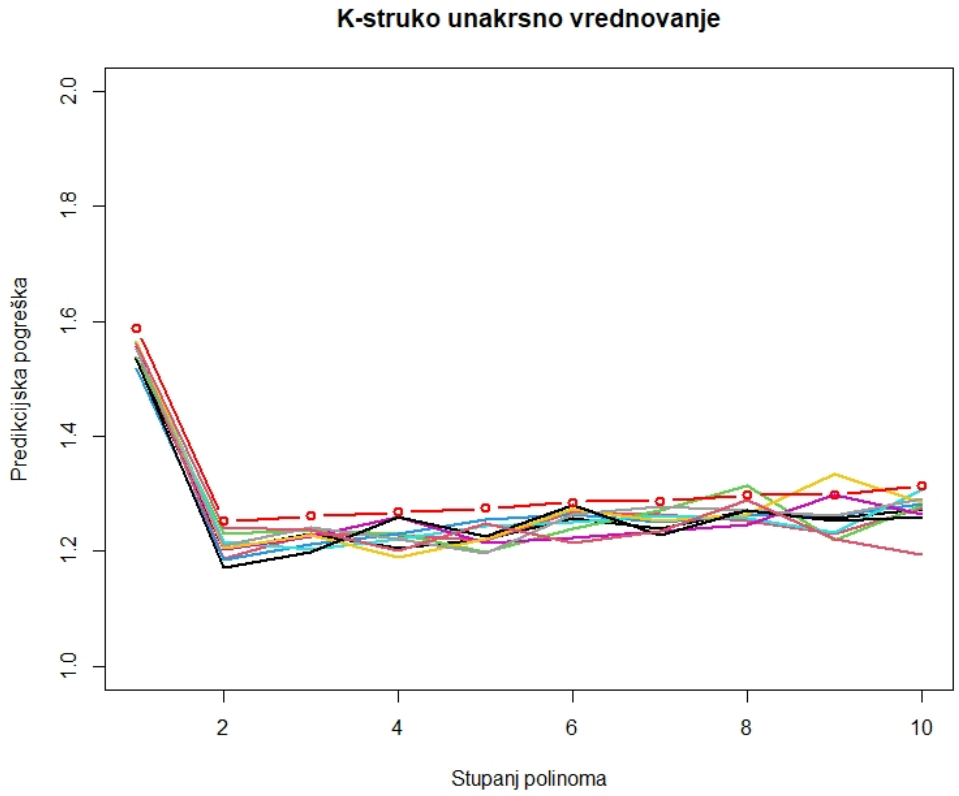
$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i)) \tag{4.1}$$

Standardni odabiri za parametar K u K-strukom unakrsnom vrednovanju su $K = 5$ ili $K = 10$. Specijalan slučaj kada je $K = N$ se naziva *pojedinačno unakrsno vrednovanje* (eng. Leave-One-Out Cross Validation, skraćeno LOOCV). Tada vrijedi $\kappa(i) = i$, za $i = 1, 2, \cdots, N$, te jednakost (4.1) prelazi u:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-i}(x_i)) \tag{4.2}$$

Dakle kada radimo LOOCV, modele treniramo onoliko puta koliko imamo podataka, što očigledno može biti potencijalno problematično ukoliko je N jako velik i/ili ako pojedinačno treniranje traje dugo, iako vrijedi napomenuti da u slučajevima nekih specijalnih metoda učenja postoje "kraći putevi" računanja procjenitelja predikcijske pogreške.

Pogledajmo sada upotrebu metoda K-strukog unakrsnog vrednovanja za slučaj $K = 10$ i LOOCV u situaciji kao u Primjeru 4.2.1.
Time dobivamo sljedeću sliku:



Slika 4.3: Crvena krivulja sa točkama označava LOOCV procjenitelj predikcijske pogreške za modele polinomijalne regresije do desetog stupnja. 10-struko unakrsno vrednovanje je provedeno za deset različitih podjela podataka, kao i u prošlom primjeru.

Uspoređujući Slike 4.2 i 4.3 možemo primjetiti kako je varijabilnost procjenitelja predikcijske pogreške u slučaju 10-strukog unakrsnog vrednovanja manja od varijabilnosti procjenitelja predikcijske pogreške u slučaju pristupa validacijskog skupa. Iduće pitanje koje se prirodno nameće jest pitanje odnosa varijance-pristranosti kod procjenitelja predikcijske pogreške u razmatranim metodama.

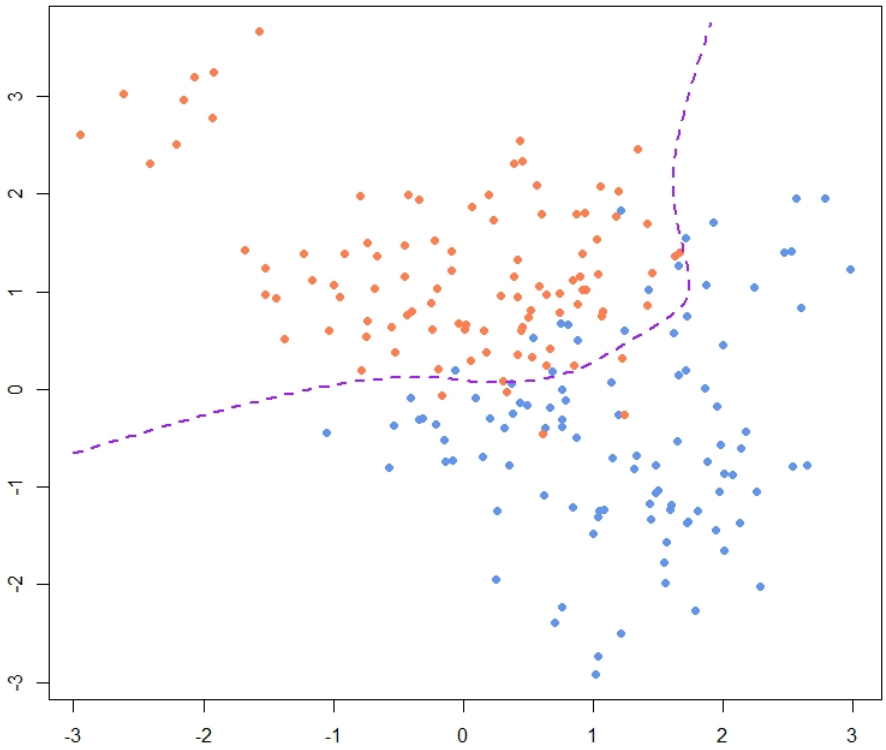
Budući da u pristupu validacijskog skupa koristimo značajno manji skup podataka za treniranje, očito je kako će procjenitelj imati veću pristranost nego procjenitelj dobiven pojedničnim unakrsnim vrednovanjem u kojem su modeli trenirani nad gotovo čitavim skupom podataka. Međutim, upravo jer je N modela trenirano nad N visoko koreliranih (sličnih, budući da se razlikuju u samo jednom podatku) skupova podataka, procjenitelj

dobiven metodom LOOCV ima visoku varijancu. S druge strane, uporabom K-strukog unakrsnog vrednovanja, koreliranost skupova podataka nad kojima je K modela trenirano se smanjuje, te time i varijanca procjenitelja. Međutim, sada dolazi do porasta pristranosti procjenitelja u odnosu na metodu pojedinačnog unakrsnog vrednovanja. Dakle, procjenitelj predikcijske pogreške dobiven metodom LOOCV ima nisku pristranost i visoku varijancu, dok procjenitelj dobiven metodom K-strukog unakrsnog vrednovanja ima nižu varijancu i višu pristranost, no i dalje nižu pristranost od procjenitelja dobivenog pristupom validacijskog skupa u slučaju kada imamo malo podataka. U praksi se pokazuje da je K-struko unakrsno vrednovanje za $K = 5$ ili $K = 10$ dobar odabir za metodu procjene predikcijske pogreške, budući da takav procjenitelj ne pati od visoke pristranosti ni visoke varijance.

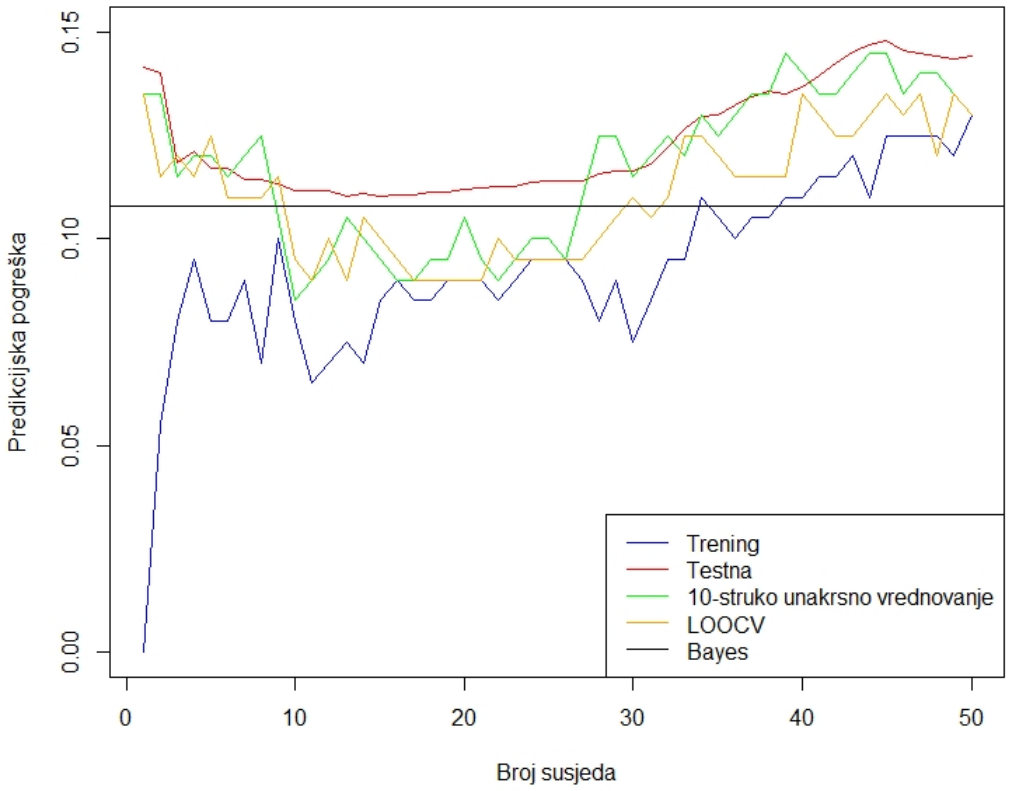
U slučaju malog skupa podataka, ovisno o brzini učenja primjenjivane metode, čak i procjenitelj dobiven K-strukim unakrsnim vrednovanjem za $K = 5$ ili $K = 10$ može patiti od visoke varijance, no budući da se metode unakrsnog vrednovanja, osim u problemu vrednovanja modela, mogu primjeniti i u problemu odabira modela, u tom slučaju takva pojava ne bi predstavljala značajan problem ukoliko samo želimo pronaći najbolji model unutar skupa razmatranih modela.

4.3 Primjeri

Primjer 4.3.1. *Vraćamo se na situaciju iz Primjera 1.2.2, te sada opisujemo pozadinu podataka. Prvo je generirano deset sredina za svaku od grupa. Za grupu plavih, deset sredina je generirano iz dvodimenzionalne normalne razdiobe* $N_2((1,0), \sum)$*, dok je za grupu narančastih deset sredina generirano iz dvodimenzionalne normalne razdiobe* $N_2((0,1), \sum)$*, pri čemu je u oba slučaja kovarijacijska matrica* $\sum = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$*. Potom je generirano sto uzoraka od svake od grupa, pri čemu je svaki podatak ponovno generiran iz dvodimenzionalne normalne razdiobe, pri čemu je očekivanje jedan od deset mogućih predefiniranih sredina. U ovom primjeru, računamo razne predikcijske pogreške, uključujući predikcijske pogreške dobivene metodama unakrsnog vrednovanja, pri čemu je korištena klasifikacijska tehnika K-najbližih susjeda.*



Slika 4.4: Prikaz skupa podataka. Ljubičasta isprekidana linija kao i ranije označava Bayesovu granicu odluke koja je optimalna granica odluke, te predikcijska pogreška vezana uz nju, Bayesova pogreška, predstavlja minimalnu moguću stvarnu predikcijsku pogrešku.



Slika 4.5: Kao što možemo vidjeti, optimalan broj susjeda je negdje oko 15, te vidimo da načelno obje predikcijske pogreške dobivene metodama unakrsnog vrednovanja dobro prate trend prave predikcijske pogreške, iako opažamo pojavu pristranosti, odnosno vidimo da na većem dijelu podcjenjuju pravu predikcijsku pogrešku. Štoviše, na određenom dijelu su manje čak i od Bayesove pogreške, što možemo pripisati čistoj slučajnosti i malom skupu trening podataka. Dakle, u ovom slučaju možemo zaključiti kako bi metode unakrsnog vrednovanja polučile solidne rezultate u problemu odabira modela.

Bibliografija

[1] David Dalpiaz, *R for Statistical Learning*, 2020.

[2] R. Durret, *Probability: Theory and Examples*, Wadsworth & Brooks, 1991.

[3] T. Hastie, R. Tibshirani i J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.*, Springer, 2011.

[4] G. James, D. Witten, T. Hastie i R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R, , 2nd ed.*, Springer, 2021.

[5] Alan Jeffares, *elements-of-statistical-learning*, GitHub, `https://github.com/alanjeffares/elements-of-statistical-learning`.

[6] S. Kullback i R. A. Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics, 22 (1), 79–86., 1951.

[7] Gabriel J. Odom, *Covariance beween value and prediction in linear regression*, Stack Exchange, `https://stats.stackexchange.com/questions/319578/covariance-beween-value-and-prediction-in-linear-regression`.

[8] Rodvi, *Derivation of bias-variance decomposition expression for K-nearest neighbor regression*, Stack Exchange, `https://stats.stackexchange.com/questions/189806/derivation-of-bias-variance-decomposition-expression-for-k-\nearest-neighbor-regr`.

[9] N. Sarapa, *Teorija vjerojatnosti, treće izdanje*, Školska knjiga, 2002.

[10] Y. Song, *The AIC-BIC dilemma: An in-depth look*, Delft University of Technology, 2020.

[11] Sethu Vijayakumar, *The Bias–Variance Tradeoff*, University of Edinburgh, 2007.

Sažetak

U ovom radu ukratko smo analizirali ključne ideje iza problema odabira i vrednovanja modela. Prilikom procjene statističkog modela iz opaženih podataka, izbor najprikladnijeg modela između više potencijalnih modela je neizbježan problem. Također, jednom kada imamo finalni model, željeli bismo znati i koliko precizne predikcije na neviđenim podacima možemo očekivati. U tu svrhu, analizirali smo odnos varijance i pristranosti modela, te način na koji porast kompleksnosti modela utječe na njih. Potom smo uveli definicije optimizma pogreške treniranja i objasnili zašto pogreška treniranja nije nužno vjerodostojan indikator konačne generalizacije modela. Također smo razmotrili analitičke metode AIC i BIC za aproksimaciju unutar-uzoračke pogreške u svrhu odabira modela. Na kraju smo dali kratak pregled tehnike unakrsnog vrednovanja koja pruža direktan procjenitelj izvan-uzoračke pogreške, te kao takva, osim u svrhu odabira modela može poslužiti i u svrhu vrednovanja modela. Sve navedene metode i tehnike potkrijepili smo jednostavnim, ali ilustrativnim primjerima.

Summary

In this thesis, we have briefly analyzed the key ideas of the model selection and evaluation problems. When estimating a statistical model from observed data, choosing the most appropriate model among multiple potential models is an inevitable problem. Once we have the final model, we would also like to know what to expect from the final model regarding the accuracy of predictions on unseen data. For this reason, we analyzed the relationship between model variance and bias, and the way in which the increase in model complexity affects them. We then introduced definitions of training error optimism and explained why training error is not necessarily a reliable indicator of the final generalization of the model. We also considered AIC and BIC, analytical methods for in-sample error approximation for model selection purposes. Finally, we have given a brief overview of the cross-validation method which provides us with the direct estimator of extra-sample error, and as such, in addition to the purpose of model selection, it can also serve the purpose of model evaluation. We have supported all the above methods and techniques with simple but illustrative examples.

Životopis

Rođen sam 3. prosinca 1996. u Zagrebu. Pohađao sam Osnovnu školu Brestje, te potom opću gimnaziju u Sesvetama koju sam završio 2015. godine. Iste godine upisao sam preddiplomski sveučilišni studij Matematike na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Iduće godine se prebacujem na preddiplomski sveučilišni studij Matematika, smjer nastavnički koji završavam 2019. godine. Zatim upisujem diplomski sveučilišni studij Matematičke statistike na istom fakultetu. Tijekom čitavog studiranja aktivno sam se bavio grapplingom u sklopu akademske sportske zajednice SUBOS.