Procena tačnosti modela

Zašto je potrebno uvoditi toliko različitih statističkih pristupa učenju, a ne samo jedan najbolji metod? U statistici važi pravilo da nema „besplatnog ručka“ (eng. *no free lunch*): nijedna metoda ne dominira nad svim ostalim metodama, nad svim mogućim skupovima podataka. Na određenom skupu podataka, jedna određena metoda može najbolje raditi, ali neka druga metoda može bolje raditi na sličnom, ali različitom skupu podataka. Stoga je važan zadatak odlučiti koja metoda daje najbolje rezultate za dati skup podataka. Izbor najboljeg pristupa može biti jedan od najizazovnijih delova izvođenja statističkog učenja u praksi.

Da bismo procenili učinak statističke metode učenja na datom skupu podataka, potreban nam je način da izmerimo koliko se njegova predviđanja zapravo podudaraju sa posmatranim podacima.

Za regresione probleme najčešće korišćena matrika je *mean squared error (MSE)* koja je data formulom:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

gde je $\hat{f}(x_i)$ predikcija koju $\hat{f}$ daje za i-tu observaciju.

Ali generalno ne interesuje koliko dobro metoda funkcioniše na trening podacima. *Umesto toga, nas zanima tačnost predikcija koje dobijamo kada primenimo našu metodu na prethodno nepoznatim (eng. unseen) podacima iz test skupa.* Zašto je to tako?

Primer 1: Pretpostavimo da smo zainteresovani za razvoj algoritma koji može da predvidi cene akcija na osnovu cena prethodnih akcija. Metodu možemo obučiti koristeći cene akcija iz proteklih 6 meseci. Ali nas zapravo ne zanima koliko dobro naša metoda predviđa prošlonedeljne cene akcija. *Umesto toga, brinemo o tome koliko će dobro predvideti sutrašnju ili sledeću mesečnu cenu.*

Primer 2: Na sličan način, pretpostavimo da imamo klinička merenja (npr. težinu, krvni pritisak, visinu, starost, porodičnu istoriju bolesti) za određeni broj pacijenata, kao i informacije o tome da li je svaki od pacijenata ima dijabetes. Ove pacijente možemo koristiti za obuku statističke metode učenja za predviđanje rizika od dijabetesa na osnovu kliničkih merenja. U praksi želimo da ovaj metod tačno predvidi rizik od dijabetesa za buduće pacijente na osnovu njihovih kliničkih merenja. Nismo zainteresovani da li metoda tačno predviđa rizik od dijabetesa za pacijente koji se koriste za obuku modela, jer već znamo koji od tih pacijenata ima dijabetes.

Neka je neki statistički metod obučavan nad trening podacima $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, gde je $x_i = (x_{i1}, x_{i2,} \dots, x_{ip})^T$, pri čemu smo dobili aproksimaciju $\hat{f}$, takvu da je $y_i \approx \hat{y}_i = \hat{f}(x_i)$ za bilo koju observaciju (x_i, y_i). Tada možemo da izračunamo $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$. Ako su

dobijene vrednosti približno jednake $y_1, y_2, \dots, y_n$, tada je vrednost MSE za trening podatke mala. Međutim, nas ne zanima da li je $\hat{f}(x_i) \approx y_i$, umesto toga želimo da znamo da li je $\hat{f}(x_0)$ približno jednako y_0, gde je (x_0, y_0) prethodno neviđeno observacija iz test skupa, observacija koja se nije koristila za obuku statističke metode učenja.

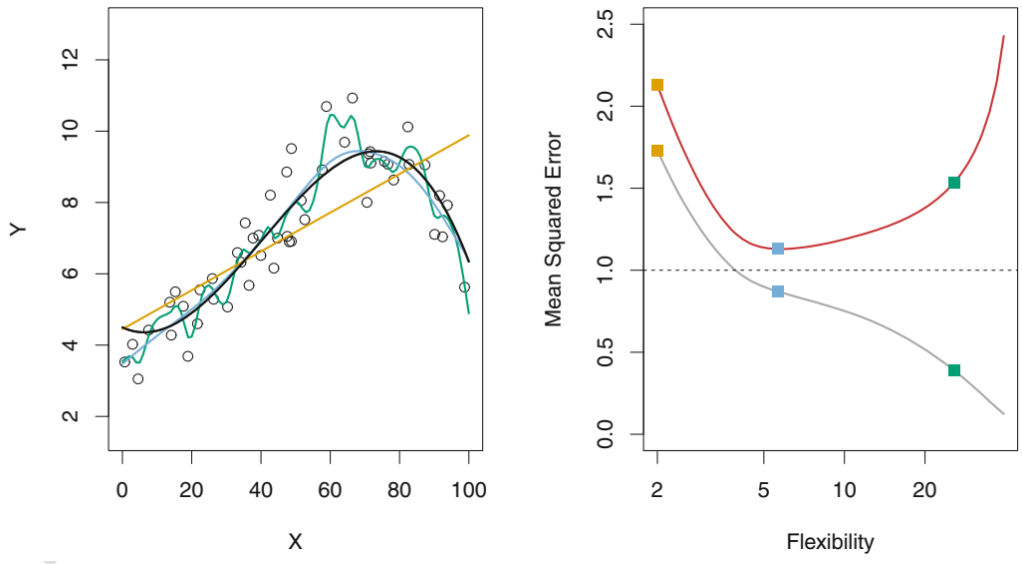
Cilj je da izaberemo metodu koja daje najniži MSE za test skup, umesto da tražimo najnižu vrednost za MSE nad trening podacima. Drugim rečima, ako bismo imali veliki broj testnih opservacija, tada bismo mogli da izračunamo

$$Avg(y_0 - \hat{f}(x_0))^2,$$

prosečnu kvadratnu grešku predikcije za sve test observacije (x_0, y_0). Izabrati model za koji je ova greška najmanja.

Ako ne postoji test skup iz nekog razloga, neko bi mogao da jednostavno odabere statističku metodu učenja koja minimalizuje trening MSE vrednost. Čini se da bi to mogao biti razuman pristup, jer se čini da su MSE za obuku i test MSE usko povezani. Međutim, postoji fundamentalni problem sa ovom strategijom: **Ne postoji garancija da će metoda sa najnižom trening MSE vrednošću imati najnižu test MSE vrednost.**

Kako to izgleda u praksi...



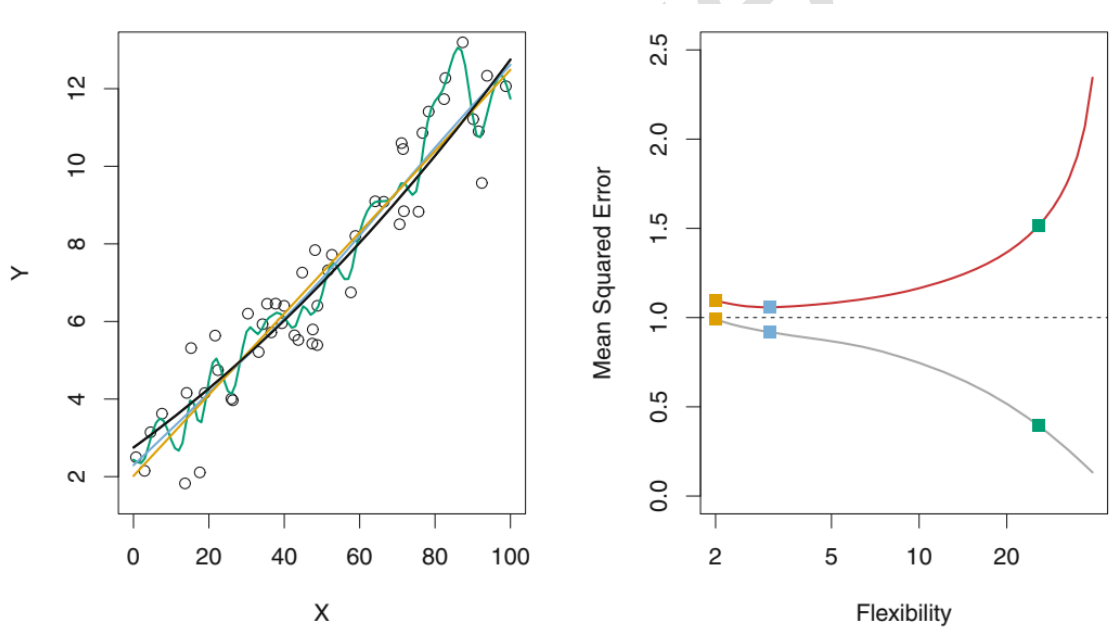
Slika 1. *Levo:* Podaci simulirani iz f, prikazani crnom bojom. Tri procene za f su prikazane: linearna regresija (narandžasta kriva) i dva *smoothing spline* (plava i zelena kriva). Npr. $f(x) = x^2 + 3x + log_2(\sqrt{X}) + \aleph(0,1)$ "e"

Desno: Trening MSE (siva kriva), test MSE (crvena kriva) i minimalno mogući test MSE za sve metode (isprekidana linija). Kvadratići predstavljaju trening i test MSE za tri aproksimacije prikazane na levoj strani.

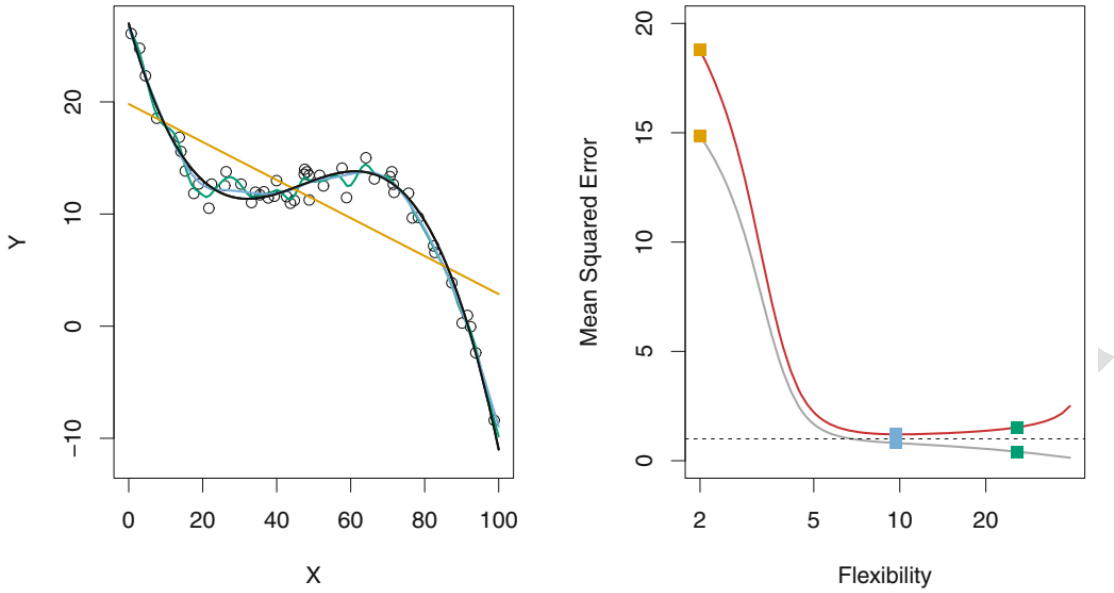
Trening MSE monotono opada kako se povećava fleksibilnost modela. U ovom primeru tačno f je nelinearno, pa narandžasti linearni fit nije dovoljno fleksibilan da bi dobro procenio f. Zelena kriva ima najniži MSE treninga od sve tri metode, jer odgovara najfleksibilnijoj od tri krive koje se uklapaju u levi panel.

Kada data metoda daje mali trening MSE, ali veliki test MSE, onda se kaže da je previše prilagođena podacima (eng. *overfitting the data*). *To se događa zato što naša statistička procedura učenja previše radi da bi pronašla obrasce u trening podacima, i možda uzima neke obrasce koji su samo uzrokovani pukom slučajnošću, a ne istinskim svojstvima nepoznate funkcije* f. Kada overfitujemo trening podatke, test MSE će imati veliku vrednost, jer „navodni obrasci" koje je metoda pronašla u trening podacima jednostavno ne postoje u test podacima.

Imajte na umu da, bez obzira da li se desio overfitting ili ne, gotovo uvek očekujemo da trening MSE bude manji od test MSE, jer većina statističkih metoda učenja bilo direktno ili indirektno nastoji da minimizira trening MSE. Overfitting se posebno odnosi na slučaj u kojem bi manje fleksibilan model dao manji test MSE.



Slika 2. Detalji su kao na slici 1, pre čemu se koristi drugačija unapred zadata f koja je mnogo bliža linearnom modelu. U ovom okruženju, linearna regresija pruža vrlo dobru aproksimaju (*fit*) podataka.



Slika 3. Detalji su kao na slici 1, pre čemu se koristi unapred zadata funkcija f koja je daleko od linearne. U ovom okruženju, linearna regresija pruža vrlo loše prilagođavanje (fit) podacima.

Bias-Variance kompromis (eng. *trade-off*)

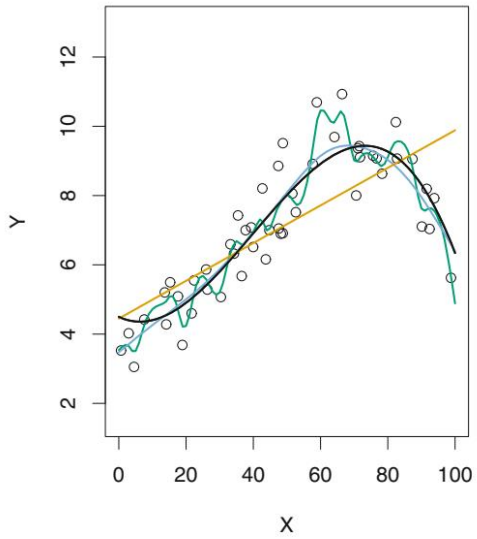
Malo matematike :) Očekivana test MSE vrednost, za datu vrednost x_0, može se napisati kao suma tri veličine:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + Var(e),$$

gde $E\left(y_0 - \hat{f}(x_0)\right)^2$ predstavlja očekivanu MSE na test skupu, i odnosi se na očekivani prosečni test MSE koji bismo dobili ako bismo više puta procenili f koristeći veliki broj trening skupova i testirali je za svako x_0. Ukupni očekivani test MSE može se izračunati kao prosek vrednosti $E\left(y_0 - \hat{f}(x_0)\right)^2$ za sve moguće vrednosti x_0 iz test skupa.

Jednačina nam govori sledeće: da bi se minimalizovala očekivana greška na test skupu potrebno je da izaberemo statističku metodu učenja koja istovremeno ima malu vrednost varijanse (eng. *variance*) i malu vrednost pristrasnosti (eng. *bias*).

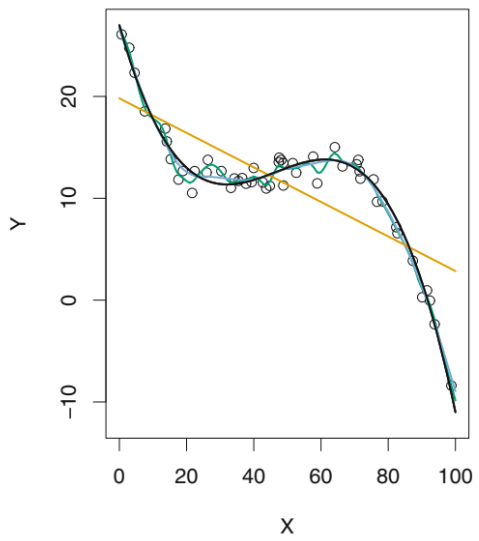
Varijansa se odnosi na količinu za koji bi se $\hat{f}$ promenila ako bismo je aproksimirali koristeći drugačiji trening skup. Pošto se trening podaci koriste za treniranje/učenje statističke metode, različiti trening skupovi rezultiraće različitim funkcijama $\hat{f}$. Ali u idealnom slučaju procena za f ne bi trebalo da variraju previše između trening skupova. Međutim, ako metoda ima veliku varijansu, male promene u trening podacima mogu rezultirati velikim promenama u $\hat{f}$. **Generalno, fleksibilnije statističke metode imaju veću varijansu.**



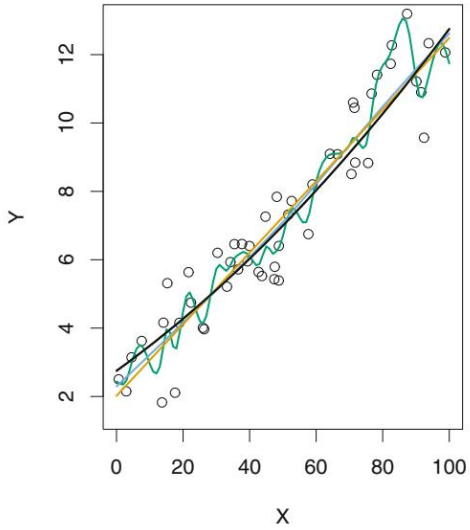
Interpretirati varijansu sa slike

(zelena vs narandžaste krive)

S druge strane, **bias** se odnosi na grešku koja nastaje aproksimacijom stvarnog/realnog problema (koji može biti izuzetno komplikovan) mnogo jednostavnijim modelom. Na primer, linearna regresija pretpostavlja da postoji linearni odnos između Y i $X_1, X_2, \ldots, X_p$. Malo je verovatno da bilo koji problem iz stvarnog života zaista ima tako jednostavan linearni odnos, pa će izvođenje linearne regresije nesumnjivo rezultirati nekim biasom u proceni f.



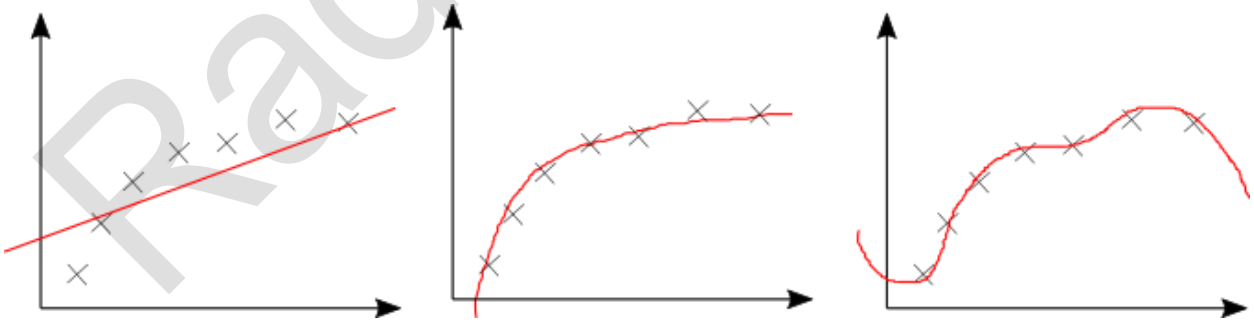
Linearna regresija rezultira velikim bias-om u ovom primeru!



Moguće je korišćenje linearne regresije za dobijanje tačne aproksim.

Generalno, fleksibilniji metode rezultiraju manjim biasom! Kao opšte pravilo, kada koristimo fleksibilnije metode, varijansa će se povećavati, a bias smanjivati.

Iz prakse: *Kako povećavamo fleksibilnost klase metoda, bias se u početku smanjuje brže nego što se povećava varijansa. Shodno tome, očekivani test MSE opada. Međutim, u nekom trenutku povećanje fleksibilnosti ima malo uticaja na bias, ali počinje da značajno povećava varijansu. Kada se to dogodi, test MSE se povećava.*



Underfit (high bias)	Ok (just right)	Overfit (high variance)

Linearna regresija

Linearna regresija je vrlo jednostavan nadgledani pristup učenju. Linearna regresija je naročito koristan alat za predviđanje kvantitativnog odgovora.

Zadatak: Pretpostavimo da se u ulozi statističkih konsultanata od nas traži da na osnovu *Advertising* podataka predložimo marketinški plan za sledeću godinu koji će rezultirati velikom prodajom proizvoda.

Evo nekoliko važnih pitanja koja bi trebalo da damo odgovor:

1. **Da li postoji veza između budžeta za oglašavanje i prodaje?**

   Prvi cilj je da pružimo dokaz o povezanosti između troškova oglašavanja i prodaje. Ako su dokazi slabi, onda se može tvrditi da nikakav novac ne bi trebalo trošiti na oglašavanje!

2. **Koliko je jak odnos između budžeta za oglašavanje i prodaje?**
3. **Koji mediji doprinose prodaji?**

   Da bismo odgovorili na ovo pitanje, moramo pronaći način da razdvojimo pojedinačne efekte svakog medija kada smo potrošili novac na sva tri medija.

4. **Koliko tačno možemo proceniti efekat svakog medija na prodaju?**

   Za svaki dolar potrošen na oglašavanje u određenom mediju, za koji iznos će se povećati prodaja? Koliko tačno možemo ovo predvideti iznos povećanja?

5. **Koliko tačno možemo predvideti buduću prodaju?**
6. **Da li je odnos linearan?**
7. **Postoji li sinergija među reklamnim medijima?**

   Možda potrošiti 50.000 USD na televizijsko oglašavanje i 50.000 USD na radio oglašavanje rezultira većom prodajom od izdvajanja 100.000 američkih dolara bilo televizija ili radio pojedinačno. U marketingu je ovo poznato kao *efekat sinergije*, dok se u statistici naziva *efektom interakcije*.

Jednostavna linearna regresija opravdava svoje ime: to je vrlo neposredan, jednostavan linearni pristup za predviđanje kvantitativnog odgovora Y na osnovu jedne prediktorske promenljive X. Pretpostavlja se da postoji približno linearna veza između X i Y. Matematički ovaj linearni odnos možemo zapisati kao:

$$Y \approx \beta_0 + \beta_1 X.$$

Na primer, X može predstavljati TV oglašavanje, a Y prodaja. Tada možemo regresirati prodaju na TV oglašavanju postavljanjem modela:

$$sales \approx \beta_0 + \beta_1 \times TV.$$

U datoj jednačini, β_0 i β_1 su dve nepoznate konstante koje predstavljaju presek (eng. *intercept*) i nagiba (eng. *slope*) u linearnom modelu. Zajedno su β_0 i β_1 poznati kao koeficijenti ili parametri modela.

Nakon što smo koristili trening podatke za određivanje parametara modela, $\hat{\beta}_0$ i $\hat{\beta}_1$, može se predvideti buduća prodaja na osnovu određene vrednosti TV oglašavanja računanjem

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x,$$

gde $\hat{y}$ označava predviđanje za Y na osnovu $X = x$.

Određivanje koeficijenata

Podaci: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

Neka je $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ predikcija za Y na osnovu $X = x_i$. Tada, $e_i = y_i - \hat{y}_i$ predstavlja i-ti residual linearnog modela. Tada se definiše sledeća veličina *residual sum of squares (RSS)* kao:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2.$$

Cilj je minimizovati RSS vrednost. Pokazuje se da važi sledeće:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ove jednačine predstavljaju metodu najmanjih kvadrata (eng. *least squares coefficient estimates*) za koeficijente proste linearne regresije.