Bias-variance tradeoff

- Posebna teorija
- Daće nam drugi ugao posmatranja generalizacije

Nagodba aproksimacije i generalizacije

- Želimo malo E_{out}:
  - dobru aproksimaciju ciljne funkcije f
  - van uzorka

- Kompleksniji skupa hipoteza $\mathcal{H} \Rightarrow$ bolja aproksimacija f
- Manje kompleksno $\mathcal{H} \Rightarrow$ bolja generalizacija van uzorka
- Idealno, $\mathcal{H} = \{f\}$ ... ali imamo bolje šanse da dobijemo na lotou...

- Jedan pristup kvantifikaciji ove nagodbe jeste VC analiza:

$$E_{out} \leq E_{in} + \Omega$$

  - E_{in} se odnosi na deo gde pokušavamo dobro da aproksimiramo f, ali ograničenje je da to radimo na uzorku
  - Sa druge strane, Ω se u potpunosti odnosi na generalizaciju

Nagodba sistematskog odstupanja i varijanse

- Nagodba sistematskog odstupanja i varijanse (*bias-variance tradeoff*) je drugačiji pristup
- Takođe dekomponujemo E_{out}, ali na nešto drugačije komponente:

1. Koliko dobro može $\mathcal{H}$ da aproksimira f
   - Kao da imamo pristup ciljnoj funkciji, a ograničeni smo baš na ovaj skup hipoteza
   - Tražimo onu hipotezu g koja najbolje aproksimira f
   - Potom kvantifikujemo kakve su performanse te najbolje hipoteze i to je naša mera sposobnosti aproksimacije
2. Koliko dobro možemo da zumiramo na dobro $h \in \mathcal{H}$
   - Moramo da zumiramo u skup hipoteza i pronađemo baš g
   - Na raspolaganju za to nam je samo skup podataka
   - Dakle, da li možemo da je pronađemo ili ćemo dobiti nešto što je loša aproksimacija aproksimacije

Nagodba sistematskog odstupanja i varijanse

- Analiza sistematsko odstupanje-varijansa će se odnositi na $y \in \mathbb{R}$
  - U VC analizi smo se ograničili na $y \in \{+1, -1\}$ - nije nemoguće ovo proširiti, ali bismo se morali upustiti u mnoge nepotrebne tehničke detalje koji ne bi doprineli našem razumevanju osnovnih koncepata
  - Sada ćemo imati uvid da se uvid da se ista nagodba i ista generalizaciona pitanja mogu primeniti i u problemima regresije

- I koristi kvadratnu grešku
  - Ograničenje analize – da bismo mogli analitički da dekomponujemo E_{out} na željene komponente (postoje načini da se proširi, ali su dosta složeniji)

Dekomponovanje E_{out}

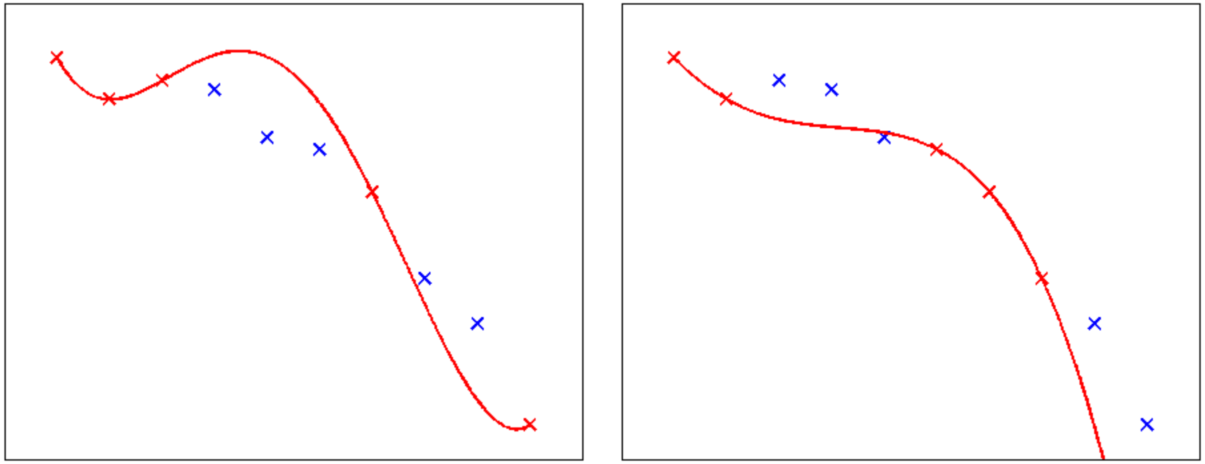
$$E_{out}\left(g^{(T)}\right) = E_{x}\left[\left(g^{(T)}(x) - f(x)\right)^{2}\right]$$

- f – ciljna funkcija
- g – konačna hipoteza – zavisi od skupa podataka T (drugi obučavajući skup bi za posledicu imao odabir druge konačne hipoteze g)
- E_x - očekivana vrednost kvadratne razlike g i f na celom ulaznom prostoru x

- Cilj nam je da razložimo E_{out} na dve konceptualne komponente (aproksimacija i generalizacija)

Očekivana greška generalizacije

- Trening skup je slučajno odabran uzorak od *N* primera
- Šta da je drugačijih *N* primera bilo izabrano? Kako bi se performanse algoritma promenile?



Primer: Fitovanje istog modela (4. stepena) za različite trening skupove (uočene opservacije su prikazane crvenom bojom). U zavisnosti od izbora trening skupa možemo dobiti *različite modele* koji imaju *različitu grešku generalizacije*

- Kakve su naše performanse *u proseku* ako imamo *N* opservacija?
  - Za ovo bi trebalo uprosečiti performanse modela preko svakog mogućeg modela (tj. svakog mogućeg trening skupa veličine *N*)

Očekivana greška predikcije

- $E_{out}\left(g^{(T)}\right)$ – zavisi od konkretnog skupa podataka T

- Da bismo se „rešili“ zavisnosti od konkretnog T u jednakosti, postupićemo na sledeći način:
  - Imamo „budžet“ od N trening primera za učenje
  - Možemo da generišemo različite skupove podataka T uz jedino ograničenje da svaki ima N primera
  - Svaki skup podataka će da rezultuje različitom hipotezom $g^{(T)}$, pri čemu će svaka imati različitu grešku van uzorka $E_{out}\left(g^{(T)}\right)$
  - Računaćemo očekivanu vrednost greške $E_{out}\left(g^{(T)}\right)$ za sve moguće skupove T od N tačaka

$$E_T\left[E_{out}\left(g^{(T)}\right)\right] = E_T\left[E_x\left[\left(g^{(T)}(x) - f(x)\right)^2\right]\right]$$
$$= E_x\left[E_T\left[\left(g^{(T)}(x) - f(x)\right)^2\right]\right]$$

- Fokusiraćemo se na $E_T\left[\left(g^{(T)}(x) - f(x)\right)^2\right]$

Prosečna hipoteza

- Definisaćemo „prosečnu“ hipotezu $\bar{g}(x)$ kao:

$$\bar{g}(x) = E_T\left[g^{(T)}(x)\right]$$

- Zamislimo da imamo *mnogo* skupova podataka $T_1, T_2, \ldots T_K$:

$$\bar{g}(x) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(T_k)}(x)$$

  - Hipoteza $\bar{g}$ bi trebala biti veoma dobra hipoteza
  - Za konkretno *T* imamo određene fluktuacije šta će biti hipoteza *g*
  - Ali, ako bismo uzeli sve moguće skupove *T* i uprosečili dobijene hipoteze, dobili bismo jako dobru hipotezu jer bismo posredno iskoristili sve moguće primere
  - Ovo u praksi, naravno, ne radimo – obično imamo jedan određen skup podataka i sve tačke u njemu koristimo za treniranje

Očekivana greška predikcije

$$E_T\left[\left(g^{(T)}(x)-f(x)\right)^2\right]=E_T\left[\left(\overbrace{g^{(T)}(x)-\bar{g}(x)}^{a}+\overbrace{\bar{g}(x)-f(x)}^{b}\right)^2\right]$$

$$E[(a+b)^2]=E[a^2+2ab+b^2]=E[a^2]+2E[ab]+E[b^2]$$

$2E[ab]$:

Ne zavisi od T, sa aspekta očekivanja je konstanta

$$2E_T\left[\left(g^{(T)}(x)-\bar{g}(x)\right)\boxed{\left(\bar{g}(x)-f(x)\right)}\right]=$$

konstanta

$$2\left(\bar{g}(x)-f(x)\right)E_T\left[g^{(T)}(x)-\boxed{\bar{g}(x)}\right]=$$

$$2\left(\bar{g}(x)-f(x)\right)\left(\boxed{E_T\left[g^{(T)}(x)\right]}-\bar{g}(x)\right)=0$$

$=\bar{g}(x)$ (po definiciji)

Očekivana greška predikcije

$$E_T\left[\left(g^{(T)}(x)-f(x)\right)^2\right]=E_T\left[\left(g^{(T)}(x)-\bar{g}(x)\right)^2\right]+E_T\left[\boxed{\left(\bar{g}(x)-f(x)\right)^2}\right]$$

Ne zavisi od T, sa aspekta očekivanja je konstanta

$$E_T\left[\left(g^{(T)}(x)-f(x)\right)^2\right]=E_T\left[\left(g^{(T)}(x)-\bar{g}(x)\right)^2\right]+\left(\bar{g}(x)-f(x)\right)^2$$

Govori nam koliko se konačna hipoteza g, dobijena pomoću skupa podataka T razlikuje od

ciljne funkcije f

Govori nam koliko se konačna hipoteza g, dobijena pomoću skupa podataka T razlikuje od

najbolje moguće hipoteze koje možemo dobiti korišćenjem skupa hipoteza $\mathcal{H}$

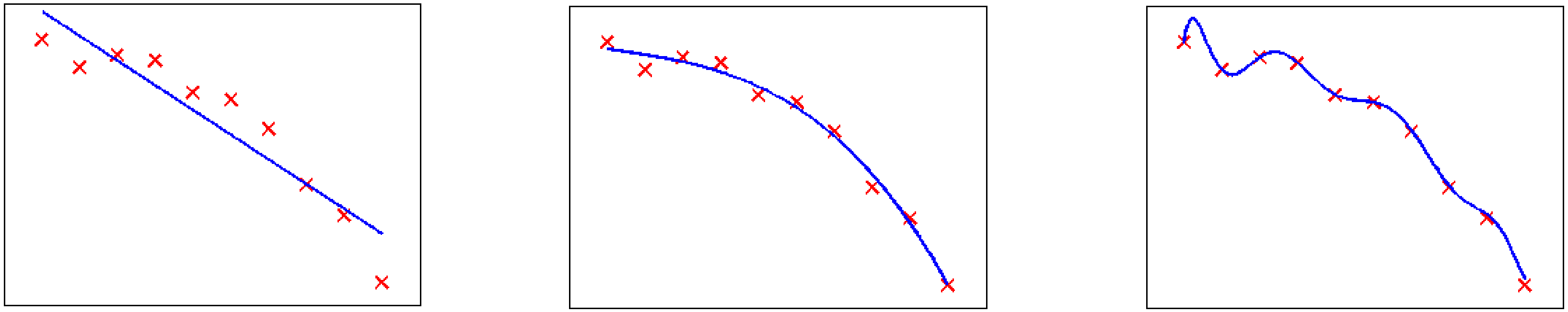
Varijansa

Govori nam koliko se najbolja moguća hipoteza iz skupa $\mathcal{H}$ razlikuje od

ciljne funkcije f

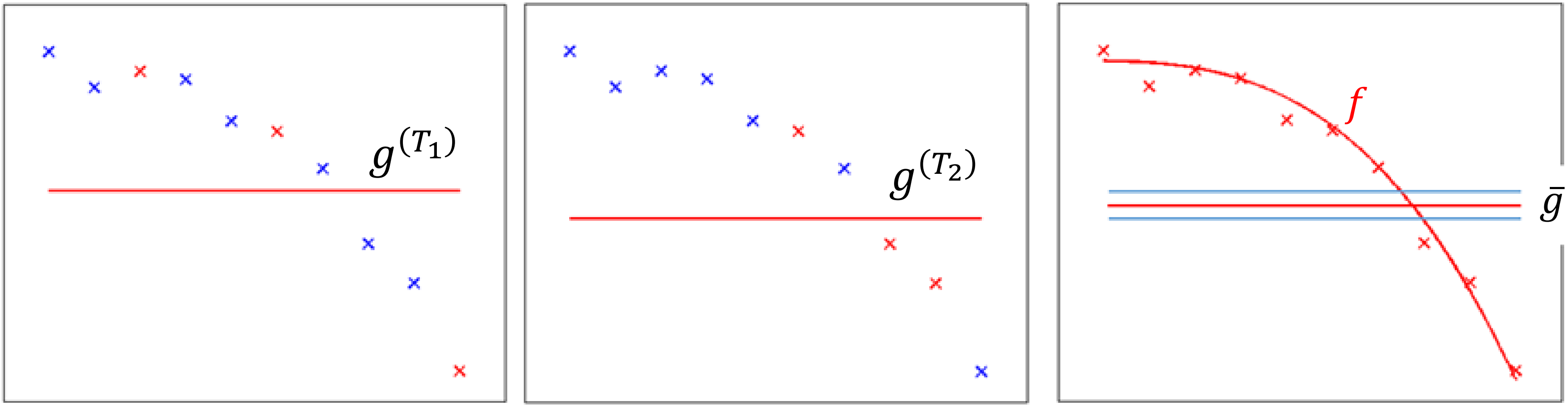
Sistematsko odstupanje

Izvori grešaka modela



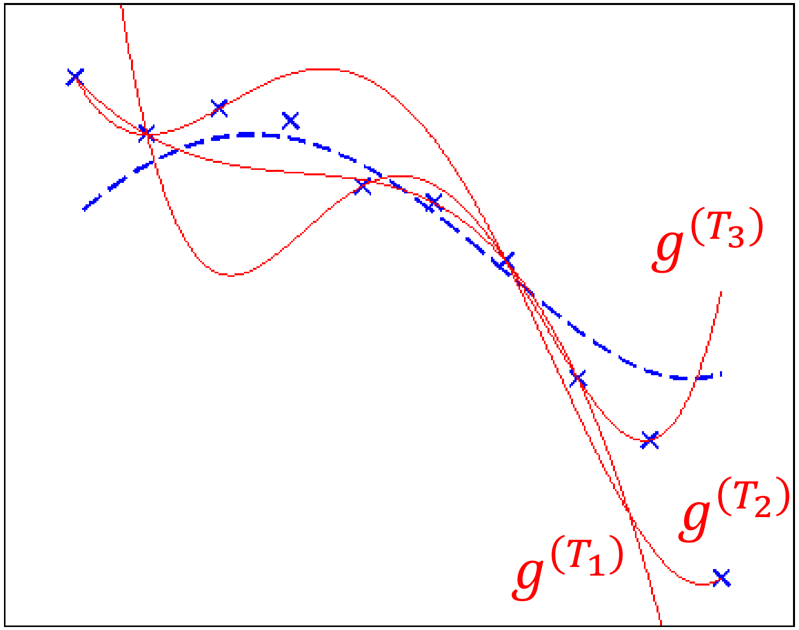
- I prvi i treći model imaju veliku generalizacionu grešku, ali su razlozi za to drugačiji
- Prvi model ima veliko *sistematsko odstupanje*
  - nije sposoban da obuhvati strukturu koju imaju podaci
- Treći model ima veliku *varijansu*
  - postoji veliki rizik da se prilagođavamo šablonima koji su se pojavili u našem malom ograničenom uzorku, a koji ne reflektuju stvaran šablon veze između x i y
  - Npr., u problemu predviđanja očekivanog životnog veka, naš uzorak (slučajno) sadrži atipične zemlje – sa nešto dužim/kraćim očekivanim životnim vekom u odnosu na trend koji postoji u podacima

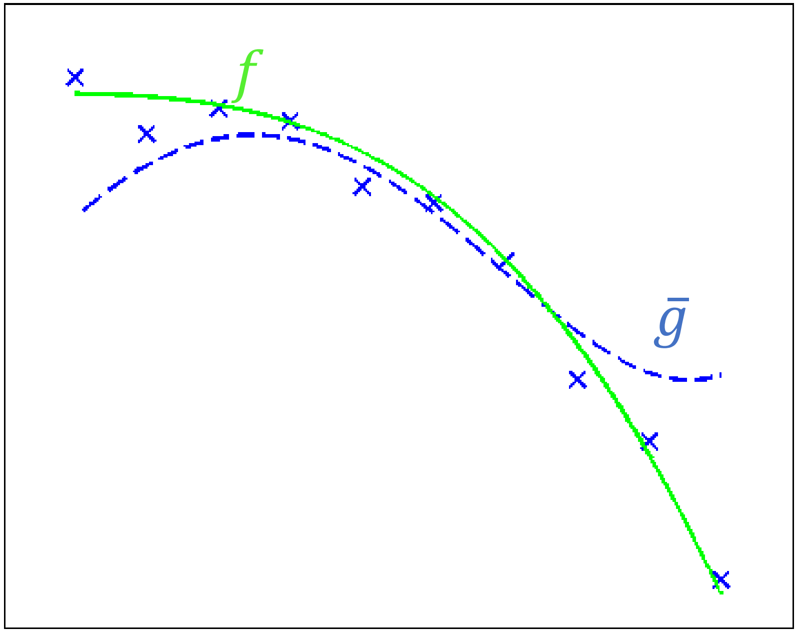
Sistematsko odstupanje modela niske kompleksnosti



- Slika: fitujemo konstantnu funkciju $y = \theta_0$ za različite skupove podataka veličine $N = 3$
  - $\bar{g}$– Za sve moguće skupove podataka veličine *N*, šta će *u proseku* biti model?
- Sistematsko odstupanje predstavlja razliku $\bar{g}$ i f
- Modeli niske kompleksnosti → veliko sistematsko odstupanje (razlika $\bar{g}$ i f je velika)

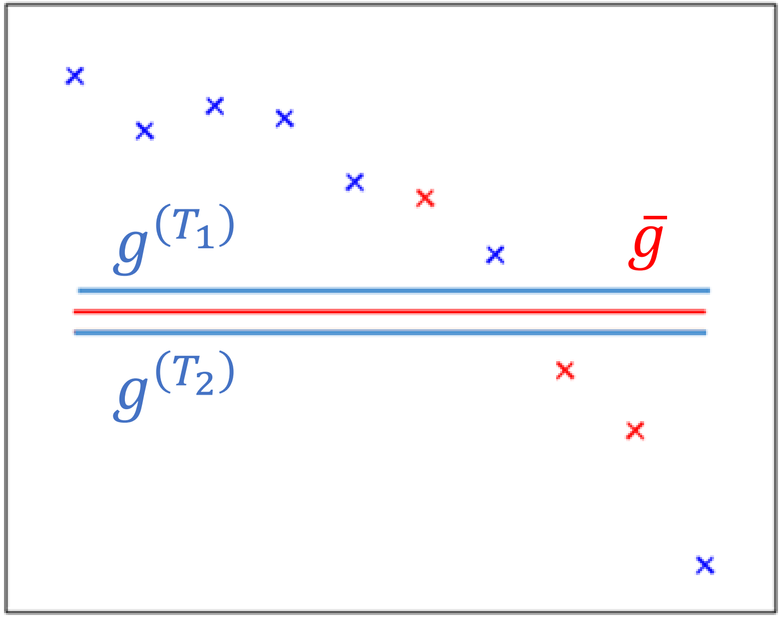
Sistematsko odstupanje kompleksnog modela





- Pretpostavimo da fitujemo polinom višeg stepena (slika: za različite skupove podataka veličine N = 5 fitujemo polinom 4. stepena)
- Model je prilično fleksibilan i u *proseku* dosta dobro odgovara stvarnoj vezi x i y
- Modeli visoke kompleksnosti → nisko sistematsko odstupanje

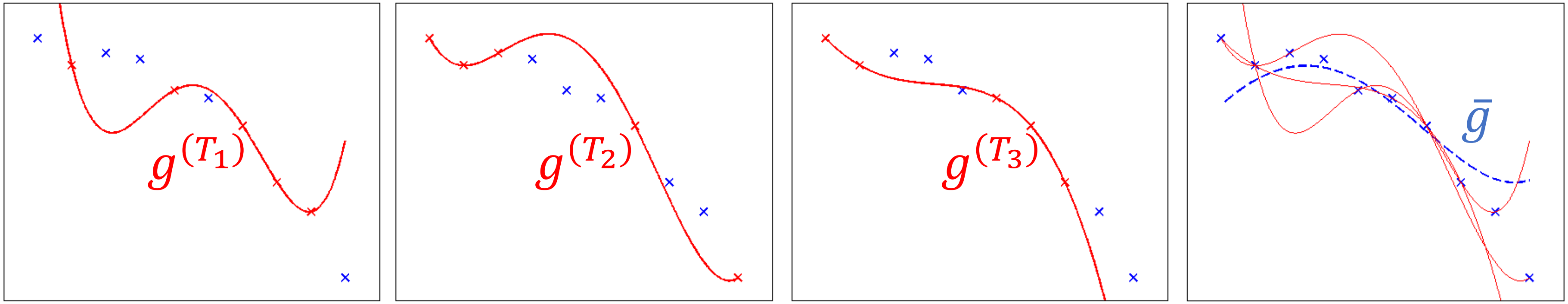
Varijansa modela niske kompleksnosti



Slika: fitujemo konstantnu funkciju $y = \theta_0$ za različite skupove podataka veličine N = 3

- Modeli ne variraju mnogo oko očekivanog modela $\bar{g}$
- Mala kompleksnost → niska varijansa

Varijansa modela visoke kompleksnosti



- Pretpostavimo da fitujemo polinom višeg stepena (slika: za različite skupove podataka veličine *N* = 5 fitujemo polinom 4. stepena)
- Prosečan model $\bar{g}$ je prilično razuman ali predikcije pojedinačnih modela $g^{(T_i)}$ prilično variraju (predikcije su nestalne)
- Velika kompleksnost → visoka varijansa

Očekivana greška predikcije

- Dakle,

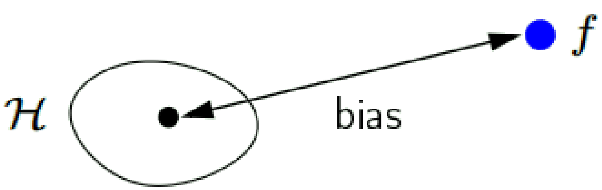
$$E_T\left[E_{out}\left(g^{(T)}\right)\right] = E_x\left[E_T\left[\left(g^{(T)}(x) - f(x)\right)^2\right]\right]$$
$$= E_x[\text{bias}(x) + \text{var}(x)] = \text{bias} + \text{var}$$

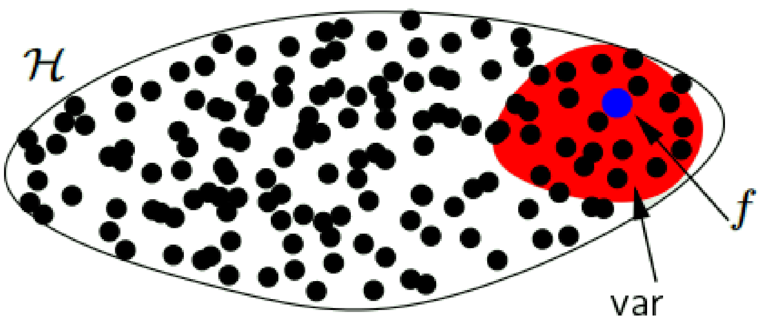
(očekivanu vrednost bias po x ćemo prosto zvati bias i, isto tako, za varijansu)

- Ovo je *bias-variance decomposition*

Nagodba sistematskog odstupanja i varijanse

$$\text{bias} = E_x\left[\left(\bar{g}(x) - f(x)\right)^2\right] \qquad \text{var} = E_x\left[E_D\left[\left(g^{(T)} - \bar{g}(x)\right)^2\right]\right]$$

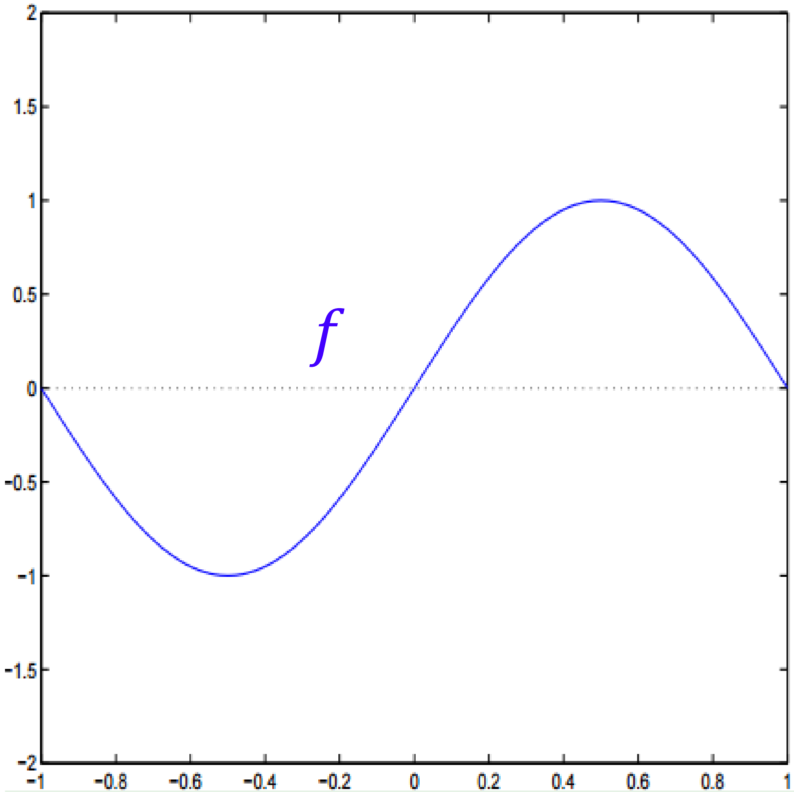




$\downarrow \qquad \mathcal{H}\uparrow \qquad \uparrow$

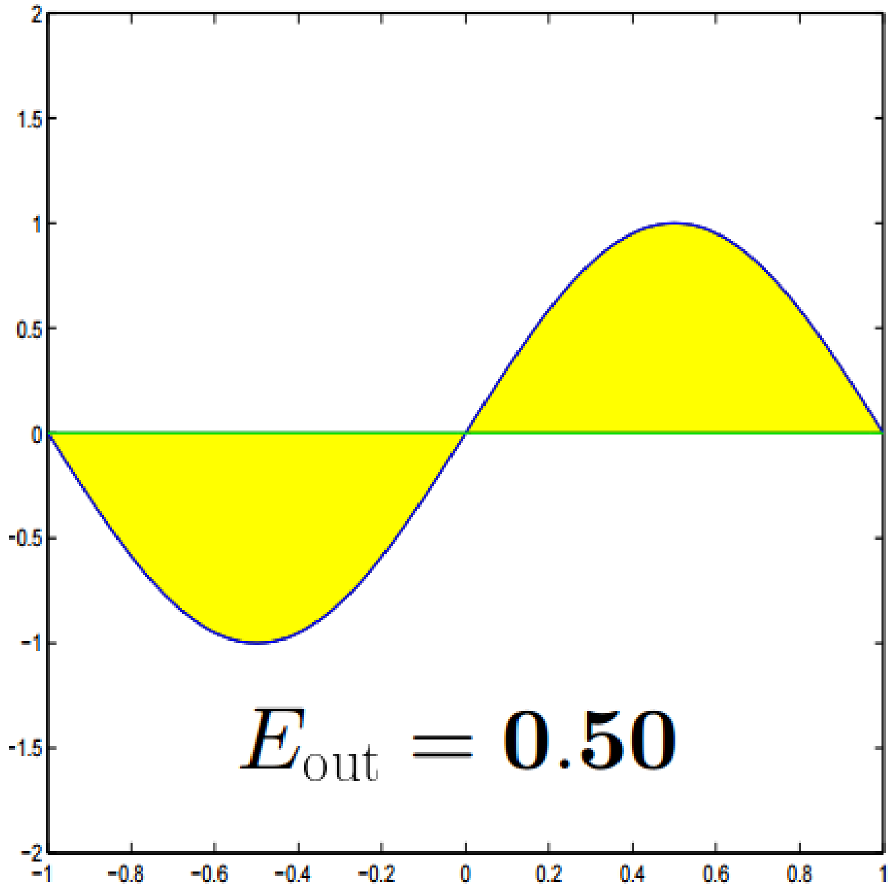
Primer

- Ciljna funkcija
$$f = \sin(\pi x)\,, f\colon [-1,1] \to \mathbb{R}$$
- Fiksiraćemo $N = 2$
- Pokušaćemo da učimo pomoću dva modela:
  - $\mathcal{H}_0\colon h(x) = b$
  - $\mathcal{H}_1\colon h(x) = ax + b$
- Koji je bolji, $\mathcal{H}_0$ ili $\mathcal{H}_1$?
- Ključno pitanje: bolji za šta? Ako pričamo o aproksimaciji, to se razlikuje od učenja
- Imajte na umu da je ovo pitanje različito od „šta bolje aproksimira sinusoid – konstanta ili linija“. Pitanje je „date su mi dve tačke iz *nepoznate* ciljne funkcije. Šta je bolje da koristim – konstantu ili liniju“

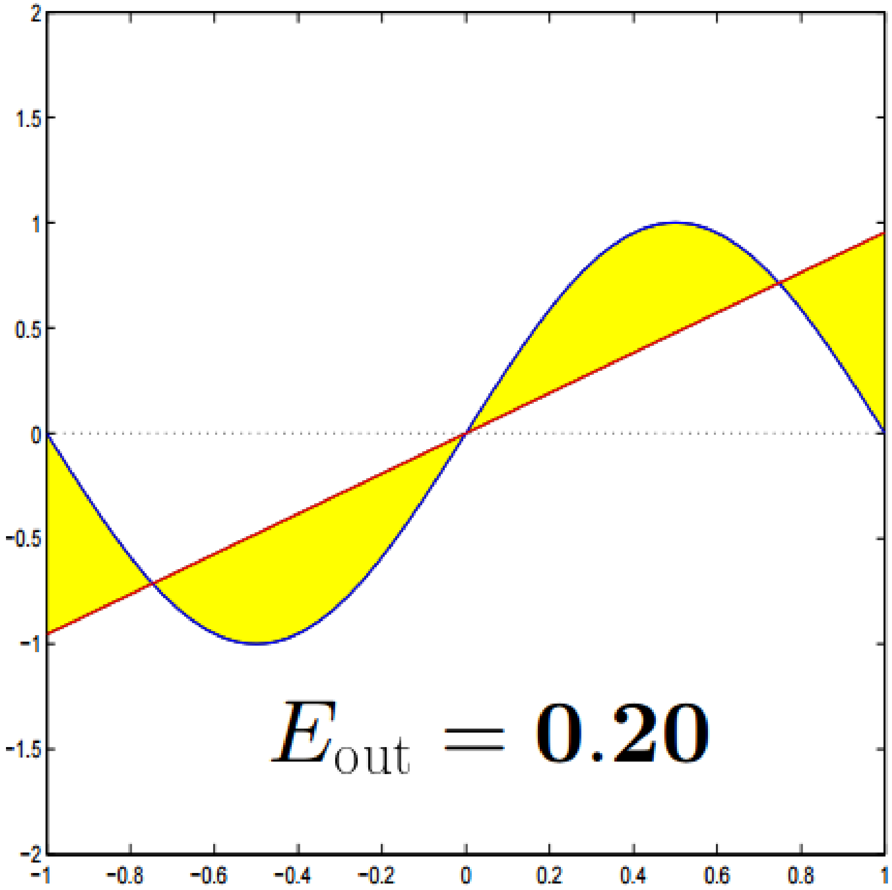


Aproksimacija

$\mathcal{H}_0$

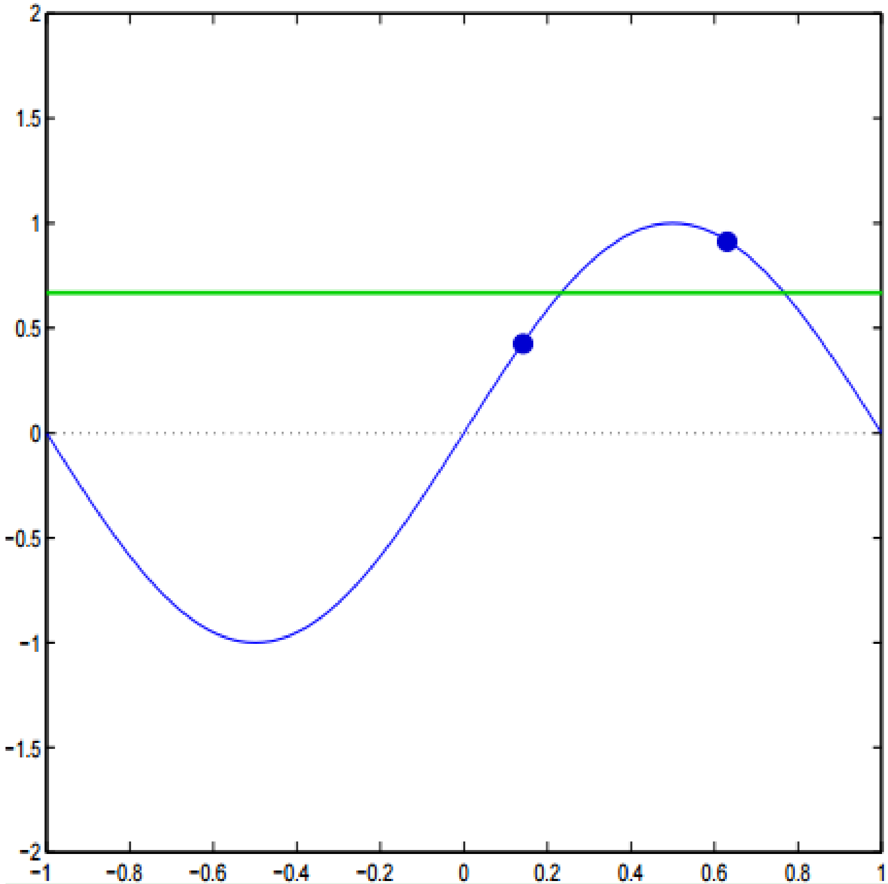


$\mathcal{H}_1$

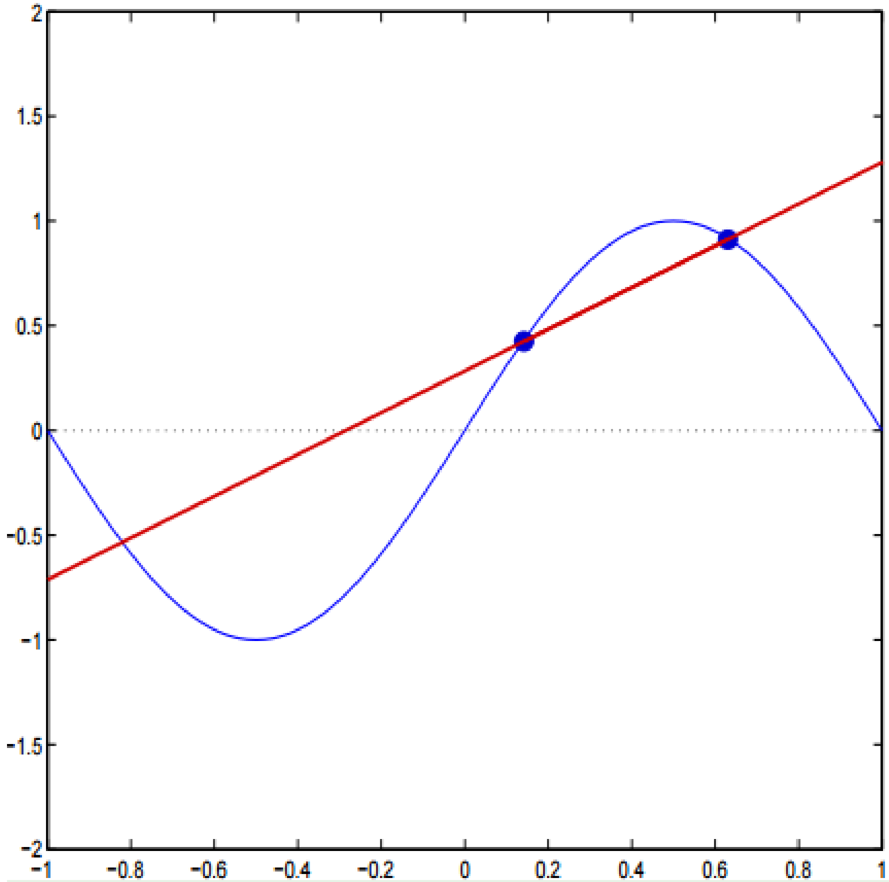


Učenje iz dve konkretne tačke

$\mathcal{H}_0$



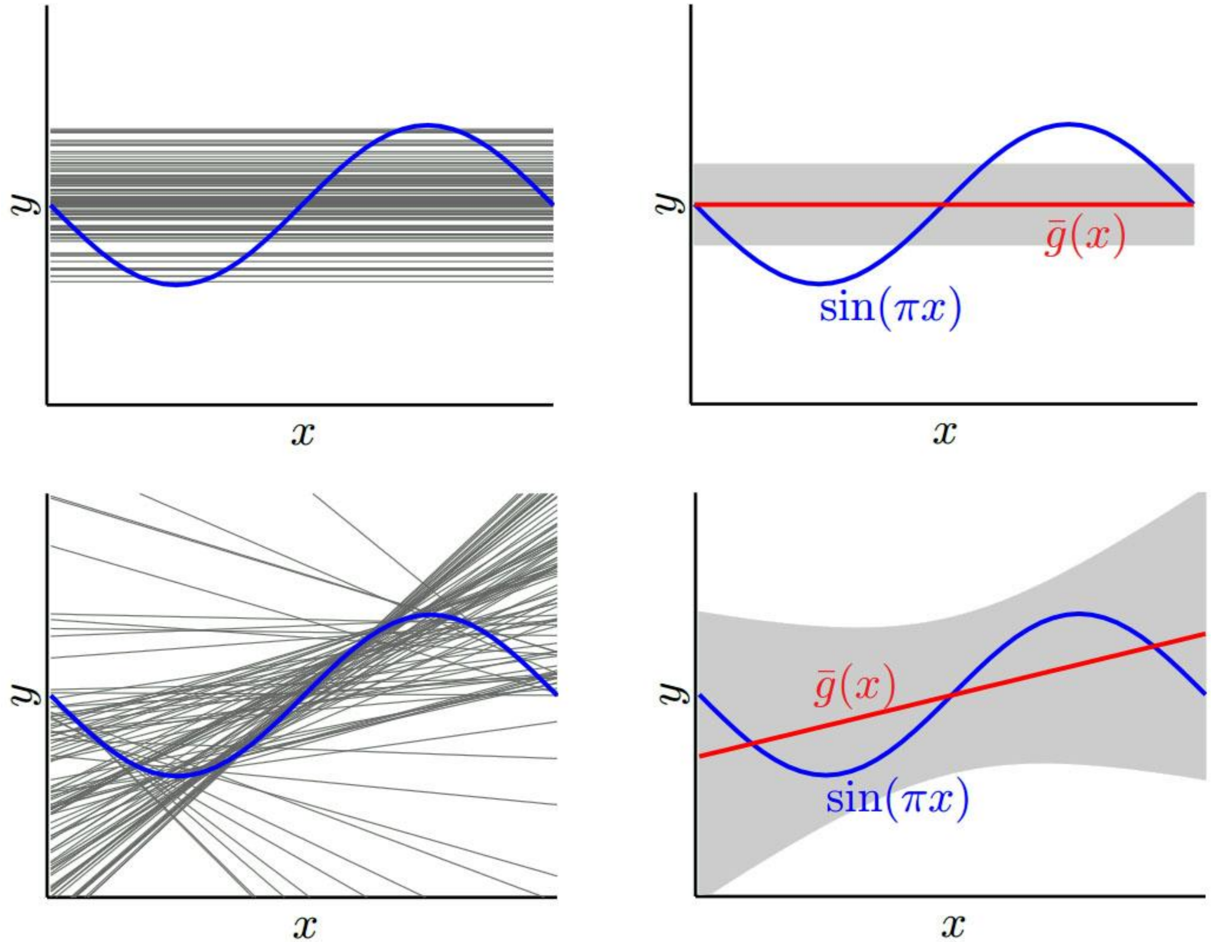
$\mathcal{H}_1$



Sistematsko odstupanje i varijansa

$\mathcal{H}_0$

$\mathcal{H}_1$



bias = 0.50
var = 0.25
E_{out} = 0.75

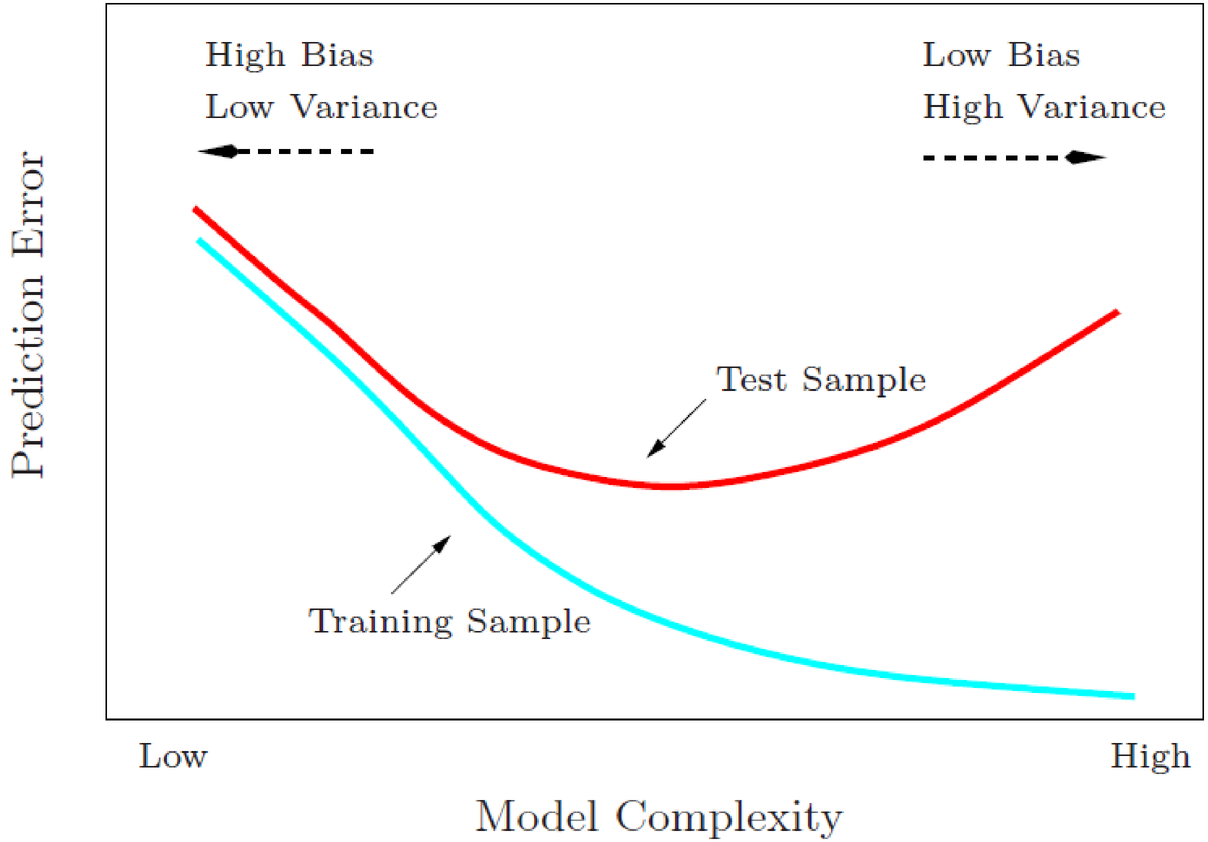
bias = 0.21
var = 1.69
E_{out} = 1.90

Lekcija naučena iz eksperimenta

- Podešavajte kompleksnost modela da odgovara resursima podataka sa kojima raspolažete, a ne prema kompleksnosti ciljne funkcije
  - Ne znamo šta je ciljna funkcija. A čak i da znamo nivo njene kompleksnosti, možda nemamo dovoljno resursa da naučimo tako kompleksnu funkciju
  - Ciljna hipoteza se čak može nalaziti u skupu hipoteza koji razmatramo, ali, pitanje je da li raspolažemo sa dovoljno podataka da je naučimo. Moramo se ograničiti na skup hipoteza koji možemo efektivno da pretražimo pomoću dostupnih resursa
  - Resursi se odnose na količinu dostupnih podataka, ali i njihov kvalitet (šum će učiniti da stvari budu još gore)

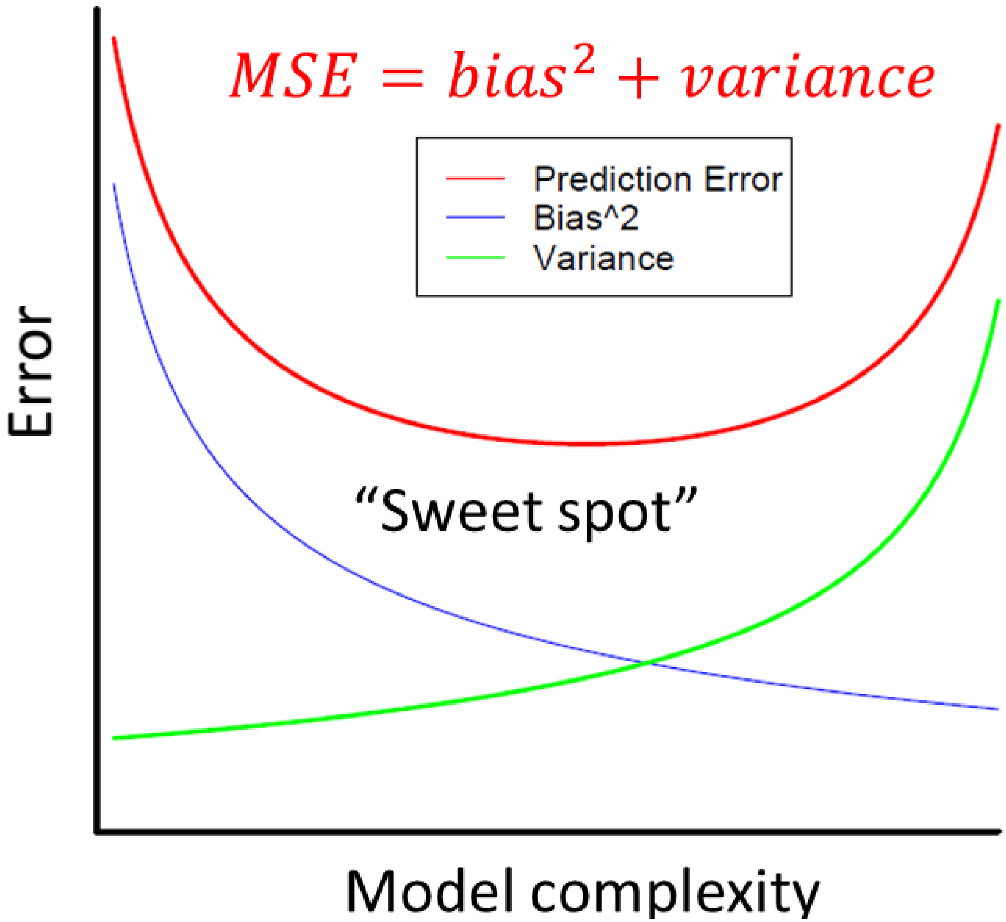
Nagodba sistematskog odstupanja i varijanse

- Napomena: u prethodnom eksperimentu smo *znali* šta je ciljna funkcija *f* i to nam je omogućilo da izračunamo tačnu vrednost sistematskog odstupanja i varijanse
- U stvarnoj primeni ne poznajemo *f*, pa ne možemo tačno izračunati sistematsko odstupanje i varijansu, ali, postoji način da *procenimo*



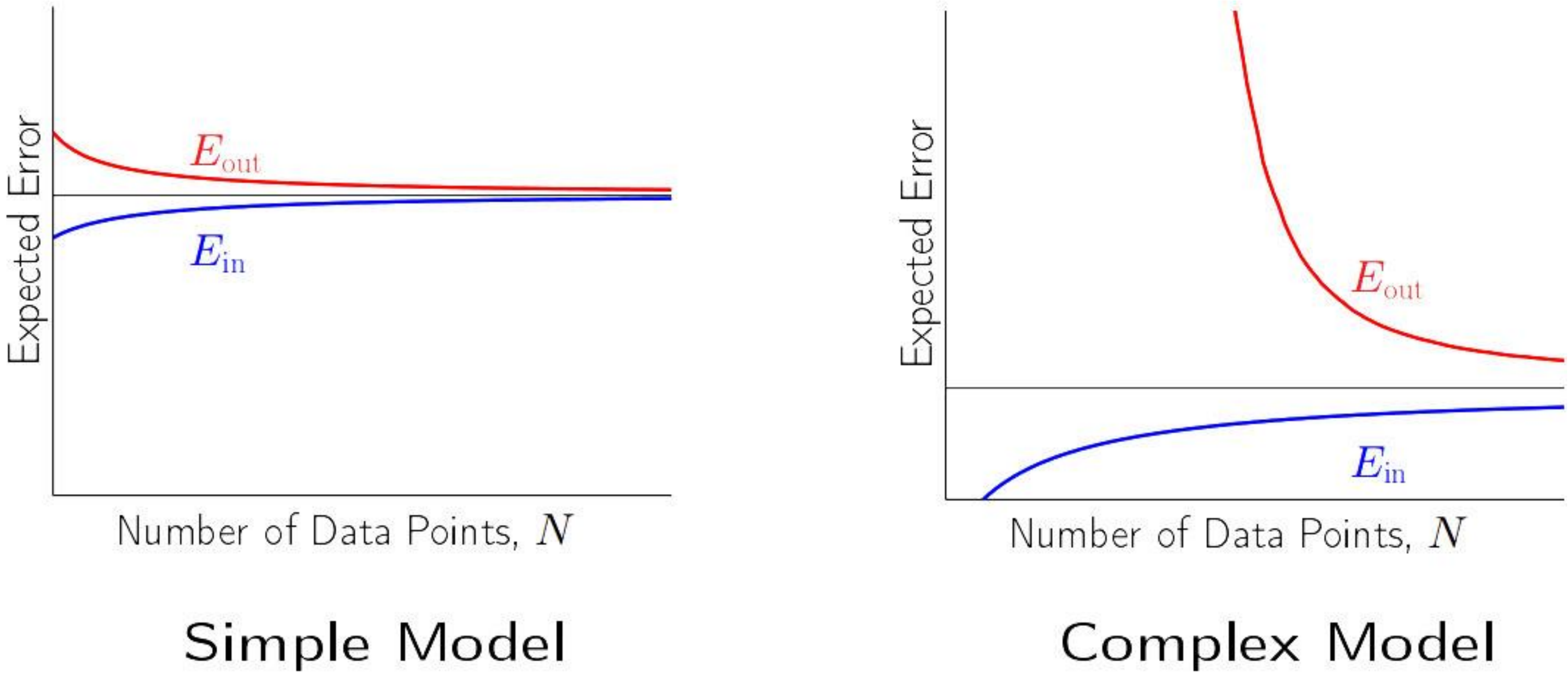
Nagodba sistematskog odstupanja i varijanse

- Često, sa povećanjem kompleksnosti:
  - Sistematsko odstupanje modela opada
  - Varijansa raste
- Postoji *sweet spot* – kompleksnost modela pri kojoj imamo najmanji doprinos sistematskog odstupanja i varijanse

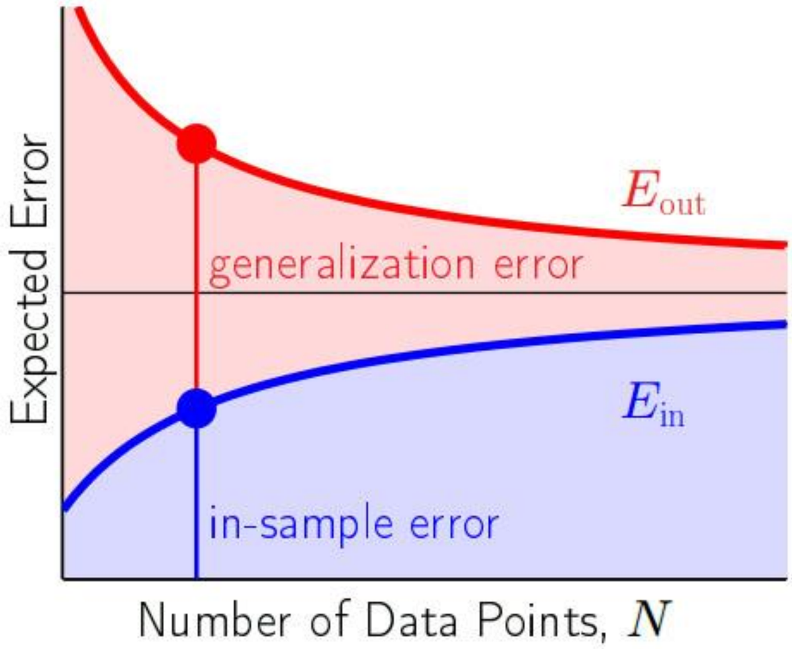


Learning curves

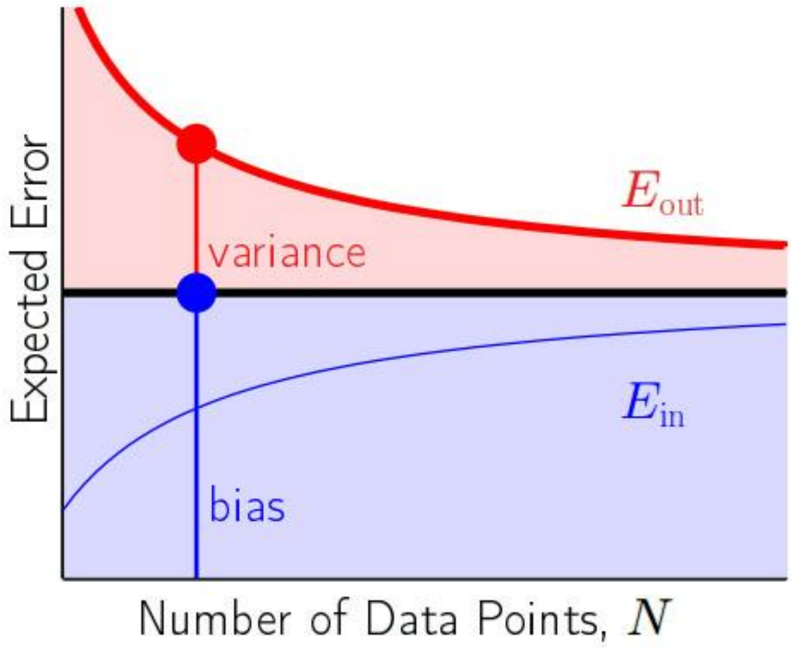
- Pomenuli smo ih kao metod za ML dijagnostiku
  - metod koji nam dopušta da utvrdimo da li naš model pati od velikog sistematskog odstupanja ili od velike varijanse
- U suštini, na njima se iscrtava očekivana vrednost $\mathrm{E}_{\mathrm{T}}\left[E_{out}\left(g^{(T)}\right)\right]$ i $\mathrm{E}_{\mathrm{T}}\left[E_{in}\left(g^{(T)}\right)\right]$ kao funkcija broja primera N



Odnos VC analize sa analizom bias-variance



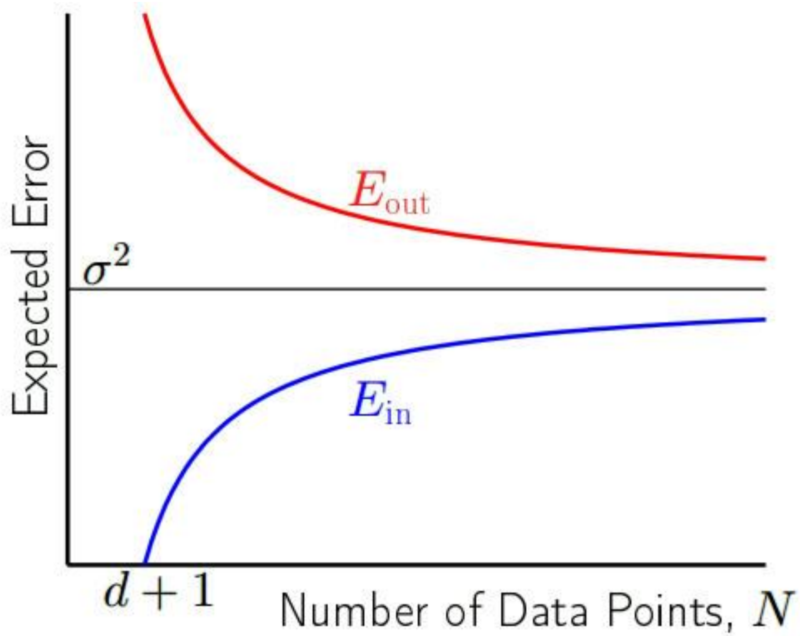
VC analysis



bias-variance

Analiza za slučaj linearne regresije

- Ciljna varijabla (sa šumom):
$$y = \theta^T x + \varepsilon$$
- Skup podataka:
$$T = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$$
- Rešenje:
$$\theta = (X^T X)^{-1} X^T y$$
- Vektor greške na uzorku:
$$X\theta - y$$
  - (ako kvadriramo pojedinačne komponente ovog vektora i sumiramo ih dobijamo E_{in})
- Vektor greške van uzorka:
$$X\theta - y'$$
  - y' smo dobili korišćenjem istih x vrednosti iz skupa T i generisanjem novih y vrednosti. Razlika u odnosu na trening skup će biti u šumu



Broj parametara

Analiza za slučaj linearne regresije

- Najbolja greška aproksimacije:

$$\sigma^2$$

- Očekivana greška na uzorku:

$$\sigma^2\left(1-\frac{d+1}{N}\right)$$

- Očekivana greška van uzorka:

$$\sigma^2\left(1+\frac{d+1}{N}\right)$$

- Očekivana greška generalizacije ($E_{out} - E_{in}$):

$$2\sigma^2\left(\frac{d+1}{N}\right)$$

VC dimenzija podeljena sa brojem primera

Tri principa učenja

- Occam’s Razor
- Sampling Bias
- Data Snooping

Jednostavna hipoteza

- An explanation of the data should be made *as simple as possible, but no simpler* – Albert Ajnštajn

- Okamova oštrica – simbol principa
  - Imamo objašnjenje podataka. Nastavljamo da „orezujemo“ objašnjenje sve dok ne stignemo do minimuma koji je i dalje konzistentan sa podacima
  - Kada dobijemo ovaj minimum, njega smatramo najboljim mogućim objašnjenjem

Okamova oštrica

- Najjednostavniji model koji odgovara podacima je najverodostojniji

1. Šta znači da je model jednostavan?
2. Kako znamo da je jednostavnije bolje?

Šta znači da je model jednostavan?

- Dva tipa merenja kompleksnosti:
  1. Kompleksnost pojedinačne hipoteze h
  2. Kompleksnost skupa hipoteza $\mathcal{H}$

Kompleksnost pojedinačne hipoteze h

- Ove mere se odnose na pojedinačan objekat – kompleksnost je svojstvo samog objekta

- MDL (*Minimum Description Length*)
  - Dati objekat pokušavamo da specificiramo sa najmanje moguće „bitova“. Što nam manje „bitova“ treba, objekat je jednostavniji
  - Npr. integer od 10^6 cifara. Kompleksnost pojedinačnih integera te dužine varira. Npr. $2^6 - 1$ je jednostavan jer smo ga mogli tako (jednostavno) opisati

- Stepen polinoma

Kompleksnost skupa hipoteza $\mathcal{H}$

- Kompleksnost se izračunava za *skup* objekata

- Entropija
  - Izvršite eksperiment
  - Razmotrite sve moguće ishode i verovatnoće koje idu uz te ishode
  - Formirate jedinstvenu kolektivnu funkciju koja obuhvata ovu verovatnoću

$$\sum p(x_i)\log\left(\frac{1}{p(x_i)}\right)$$

  - Izračunava se za *klasu* objekata – svaki ishod je jedan objekat

- VC dimenzija
  - Svojstvo skupa hipoteza
  - Posmatra skup hipoteza kao celinu i predstavlja jedinstven broj koji označava raznolikost tog skupa hipoteza
  - U ovom slučaju, raznolikost znači kompleksnost

Šta znači da je model jednostavan?

- Kada mislimo „jednostavno“, obično mislimo na jedan objekat h
  - Ne mislimo na alternative koje postoje u opisu podataka

- Dokazi Okamove oštrice su obično u terminima skupa hipoteza $\mathcal{H}$
  - Sa ovim smo se već susreli, npr. VC dimenzija

- Ovo je malo zabrinjavajuće – intuitivan koncept je jedno, a matematički dokaz drugo

Koncept koja ih povezuje: prebrojavanje

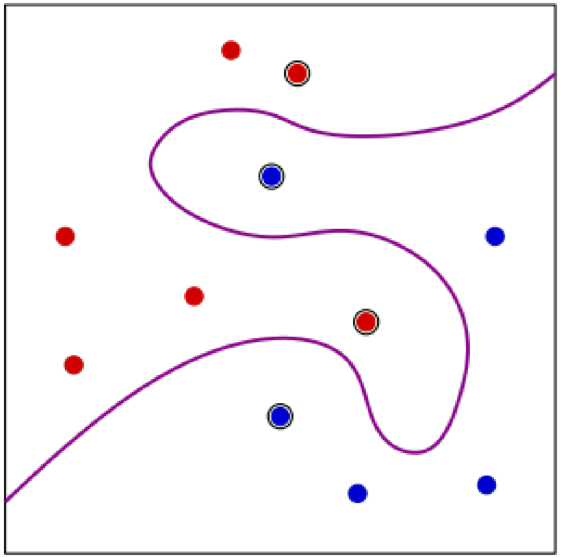
- Dobra vest jeste da su koncepti kompleksnosti objekta i kompleksnosti skupa objekata veoma povezani (skoro identični)
  - l bitova specificiraju h
  - Ovo implicira da postoji 2^l elemenata sličnih h (koji se takođe mogu opisati sa l bitova)
  - Skup svih sličnih objekata možemo označiti sa $\mathcal{H}$
  - „Jedan od 2^l“ možemo koristiti kao opis kompleksnosti $\mathcal{H}$

- l bitova specificira $h \Rightarrow h$ je jedan od 2^l elemenata skupa $\mathcal{H}$
  - Koncept: objekat je kompleksan ako je jedan od mnogih. Objekat je jednostavan ako je jedan od nekolicine
- Šta je sa parametrima koji su realni brojevi (npr. polinom 17. stepena)?
  - I dalje odgovaraju našem opisu „jedan od mnogih“

Izuzetak od pravila

- Izuzetak od ovog pravila (koji izgleda kompleksan ali je samo „jedan od nekolicine“)
  - Namerni izuzetak – želeli smo kompleksan model koji može dobro da se prilagodi podacima. Ali ipak je jedan od nekolicine – nismo želeli da platimo punu cenu kompleksnosti

SVM



Pitanje: predikcija ishoda fudbalske utakmice

- Dobili ste pismo koje predviđa ishod utakmice koja se to veče održava
- Ispostavilo se da je predikcija tačna! Ali možda je samo srećan pogodak...
- Tokom sledećih 5 nedelja dobili ste još 5 ovakvih pisama. Sve predikcije su ispale tačne!
- U šestoj nedelji dobijate pismo: „Želite još? 50$"
- Da li da platite?

0000 0000 0000 0000 1111 1111 1111 1111	0
0000 0000 1111 1111	1
0000 1111	0
0011	1
01	1

Kako znamo da je jednostavnije bolje?

- Bolje ne znači elegantnije! Bolje znači bolje performanse van uzorka E_{out}
- Argument*:
  1. Postoji manje jednostavnih hipoteza (u poređenju sa brojem kompleksnih hipoteza) $m_{\mathcal{H}}(N)$
  2. Pošto ih ima manje, manje je verovatno da će savršeno odgovarati datom skupu podataka:

$$\frac{m_{\mathcal{H}}(N)}{2^N}$$

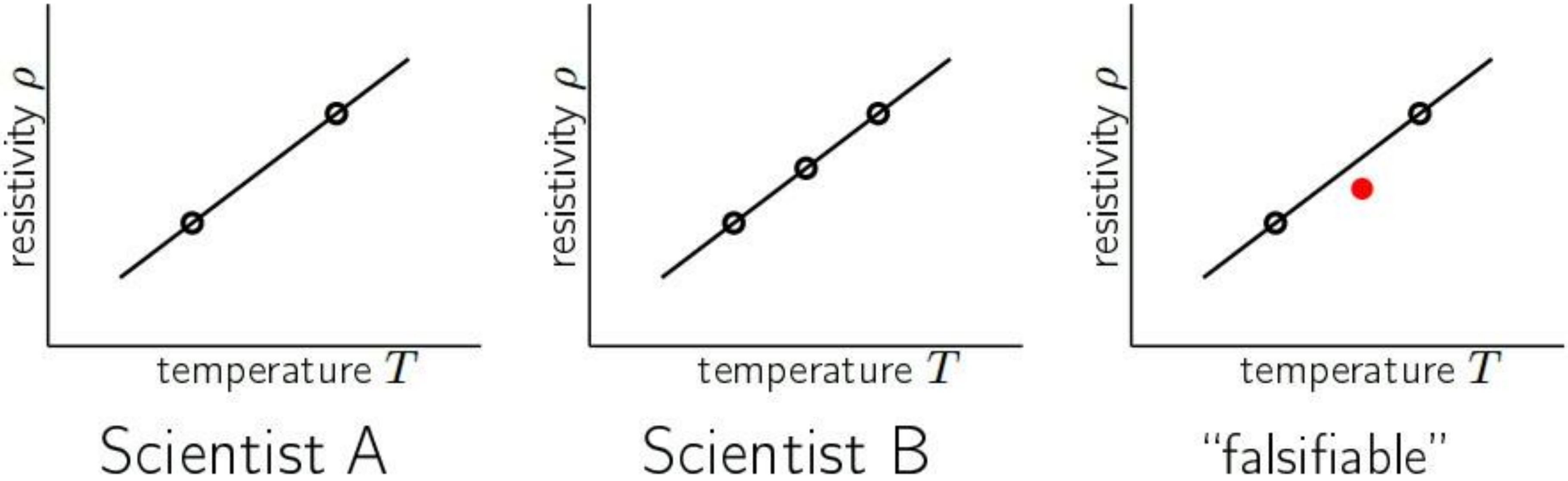
  3. Ako je nešto manje verovatno, značajnije je kada se to desi

- Npr. u prevari sa pismima
  - Onaj ko je prevaren vidi samo jednu hipotezu i ona je perfektna – zato joj pridodaje veliki značaj jer je prilično neverovatno da će se to desiti
  - Onaj ko vidi širu sliku zna da je $m_{\mathcal{H}}(N) = 2^N$ - sigurno je da će se desiti, zbog toga je beznačajno

*Formalan dokaz postoji pod različitim idealizovanim uslovima

Primer beznačajnog eksperimenta

- Želimo da utvrdimo da li su provodljivost određenog metala i temperatura u linearnoj zavisnosti
- Kakav dokaz nam sledeći rezultati na slici pružaju?
- Aksiom neopovrgljivosti (*the Axiom of Non-Falsifiability*) – ako sa datim podacima nemamo mogućnost da opovrgnemo tvrdnju, oni zbog toga ni ne mogu pružiti nikakav dokaz u korist te tvrdnje



Sampling bias – izbor predsednika

- 1948, prvi predsednički izbori nakon Drugog svetskog rata – Truman i Dewey
- Prema anketama, kandidati su veoma blizu i nije jasno ko će pobediti
- Nakon što su izbori završeni, ali glasovi još nisu prebrojani, jedna redakcija je sprovela telefonsku anketu i pitala ljude kako su glasali
- Dobili su rezultat da je Dewey neosporno pobedio
- Rezultat je izgledao toliko očigledan da su odlučili da budu prvi koji će izvestiti o tome i ištampali novine sa naslovom da je Dewey pobedio
- Pobednik (koji na slici drži novine) je Truman

HOME

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Tops Coghlan in Hot Race for Attorney

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburban Ballot Near 375,000

BULLETINS ON ELECTIONS

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Sizable Electoral Margin Seen

Sampling bias – izbor predsednika

- Šta je pošlo po zlu? Da li je ovo posledica probabilističkih garancija?

$$P[|E_{in} - E_{out}| > \epsilon] \leq \delta$$

- U ovom slučaju, nije δ krivo. Dobijeni rezultat je bio posledica sistematske greške (*bias*) u uzorkovanju:
  - 1948 telefoni su bili skupi – kuće sa telefonom su obično bile bogatije
  - Bogati ljudi su preferirali Dewey-a

Sampling bias principle

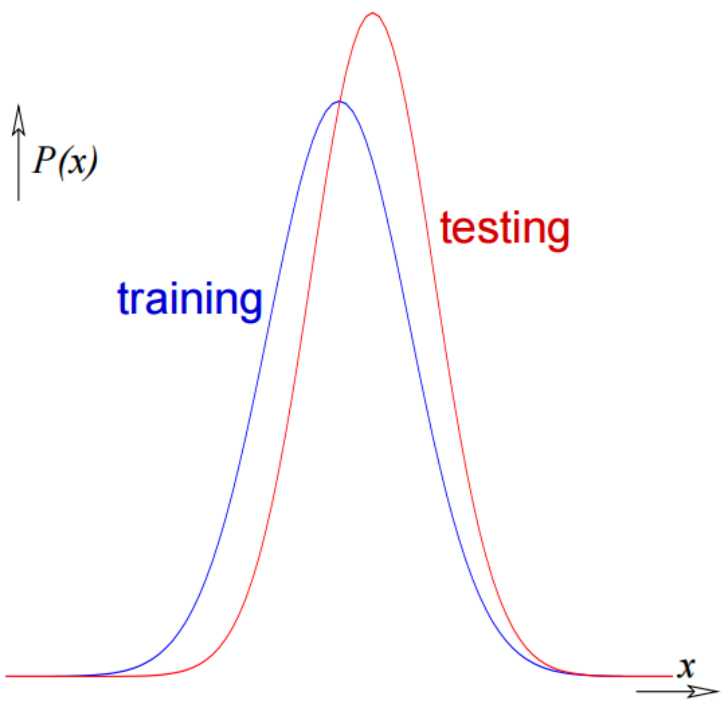
- If the data is sampled in a biased way, learning will produce a similarly biased outcome

- Moramo se obezbediti da podaci budu reprezentativni u odnosu na ono šta želimo da pokažemo

- Praktičan primer: financial forecasting
  - Situacija je prilično nepredvidiva, pod uticajem neke glasine tržište može drastično da se promeni
  - Ako želimo da pronađemo šablon koji postoji u podacima, razmatramo „normalne“ periode u kojima se taj šablon vidi
  - Kada testiramo model, testiramo ga na pravom tržištu. Može se desiti da uvidimo da postoji sistematsko odstupanje

Poklapanje distribucija

- Jedan način da se borimo sa sistematskim greškama u uzorkovanju jeste poklapanje distribucija
- Pretpostavka uvedena kako bi *Hoeffding*-ova nejednakost važila:
  - Instance korišćene za obučavanje su odabrane iz iste distribucije kao i instance koje ćemo koristiti za testiranje
  - Ako imamo sistematsku grešku u uzorkovanju, ovo je narušeno



- Ako imamo pristup trening i test distribuciji, možemo proveriti da li se poklapaju, npr., u datom primeru su trening i test distribucije donekle različite
- Ako poznajemo ove distribucije, možemo:
  - dodeliti različite težine primerima trening skupa
  - Iz datih podataka ponovo uzorkovati trening skup tako da ispadne kao da je izvučen iz druge distribucije

Poklapanje distribucija

- Ovaj metod radi i u praksi, čak i ako ne znamo konkretne distribucije, možemo da ih procenimo

- Međutim, metod ne radi ako postoji regija u ulaznom prostoru gde je $P = 0$ za trening, ali je $P > 0$ za testiranje
  - Npr. ljudi koji nisu imali telefon u 1948
  - Ne možemo ništa da uradimo pomoću poklapanja distribucija jer nemamo predstavu šta se u tom delu dešava

- Dakle, u određenim situacijama postoji rešenje za sampling bias
- Ali, u nekim situacijama, sve što možemo da uradimo jeste da priznamo da ne možemo garantovati performanse našeg rešenja u delovima koji nisu pokriveni uzorkom

Pitanje – pokušajte da detektujete sampling bias

- Odobravanje kredita mušteriji
- Istorijski podaci mušterija iz prethodne 2-3 godine
  - Dostupne su nam informacije koje svaka mušterija daje prilikom aplikacije za kredit (jer su to jedine informacije koje ćemo imati za nove mušterije)
  - I dostupno nam je ciljno obeležje – u retrospektivi, da li je banka zaradila na ovim mušterijama
- Gde je ovde sampling bias?
- Koristimo podatke o mušterijama koje smo ranije odobrili (zato što su to jedine mušterije za koje imamo podatke o vraćanju kredita)
- Mušterije koje smo odbili nisu deo ovog trening skupa
- Za novu mušteriju ne znamo da li bi ova mušterija bila odbijena ili ne prema starim kriterijumima banke, dakle, ona bi mogla biti deo skupa koji nikada nije pokriven trening skupom
- Međutim, u ovoj primeni, sampling bias nema katastrofalne posledice
  - Banke prilično agresivno dodeljuju kredit pa imamo i dovoljno primera mušterija na kojima je banka pogrešila

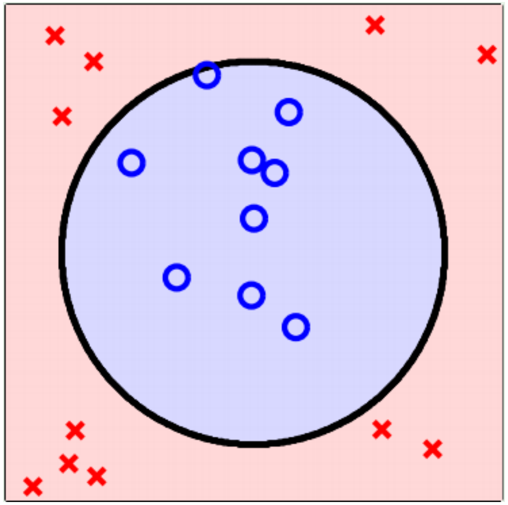
Data snooping

- Princip: ako je skup podataka imao uticaja na bilo koji korak učenja, onda je mogućnost ovog skupa podataka da proceni ishod učenja kompromitovana

- Ovo je zamka u koju upadaju mnogi
  - Ima mnogo načina da se uhvatimo u nju
  - I veoma je privlačno da upadnemo u nju jer dobijamo bolje performanse
  - Manifestuje se na mnogo različitih načina

Data snooping primeri

1. Pogledali smo podatke



- Recimo da imamo primere sa slike
- Bez gledanja smo rešili smo da primenimo transformaciju drugog stepena $z = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$. Greška na uzorku je mala, ali plaćamo cenu generalizacije jer imamo više parametara
- Ali, nakon inspekcije, shvatimo da nam ne trebaju sva obeležja – samo $z = (1, x_1^2, x_2^2)$ ili čak $z = (1, x_1^2 + x_2^2)$

Data snooping primeri

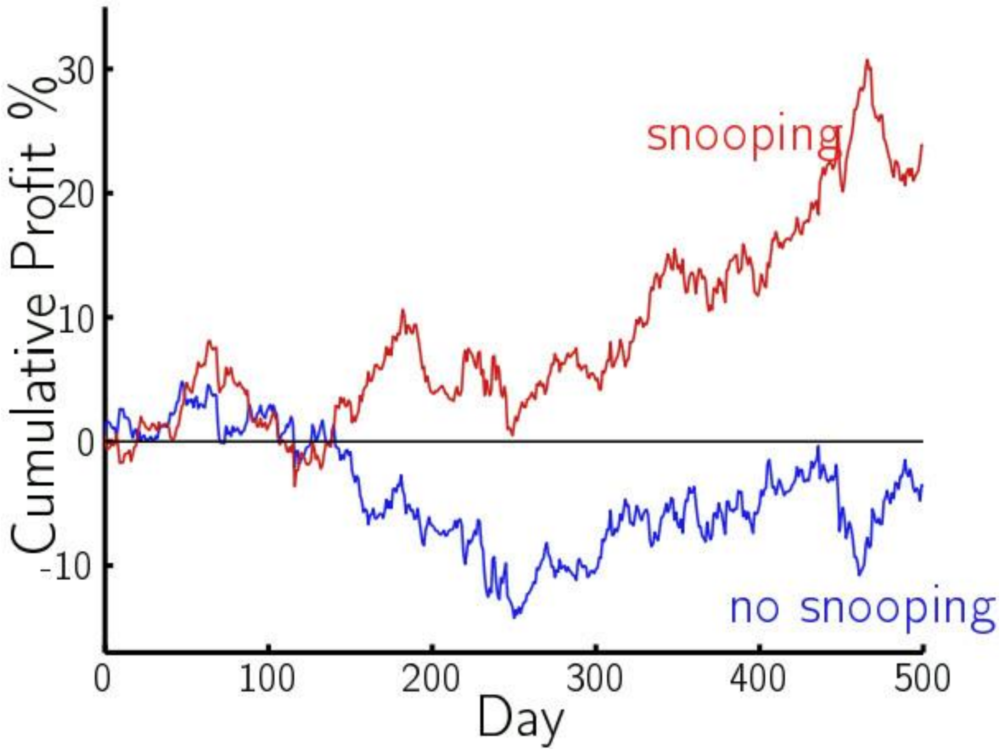
1. Pogledali smo podatke

- Ako pogledamo skup podataka, lako se može desiti da dizajniramo model da bude prilagođem tom konkretnom skupu podataka
- Iako radimo dobro na tom skupu podataka ne znamo da li ćemo jednako dobro raditi na drugom skupu podataka (generisanom iz iste distribucije)
- Data snooping uključuje skup podataka T, ali ne i druge informacije
  - U proces obučavanja možemo (i dobro je) uključiti domensko znanje (npr. koliko imamo ulaza, koliki su im opsezi, kako su mereni, da li su fizički korelirani, ...)
  - Samo ne treba razmatrati konketan skup podataka

Pitanje

- Da li možete identifikovati data snooping u ovom slučaju
- Financial forecasting: predviđanje odnosa između \$ i £
  - Imamo osam godina podataka o odnosu na dnevnom nivou (oko 2000 tačaka)
  - Izlaz je promena odnosa \$ i £ u datom danu u odnosu na prethodni Δr_0, a ulazi predstavljaju promene tog odnosa u prethodnih 20 dana: $\Delta r_{-20}, \Delta r_{-19}, \dots, \Delta r_{-1} \rightarrow \Delta r_0$
  - Prvo, normalizujemo podatke ($\mu = 0, \sigma^2 = 1$)
  - Zatim ih podelimo na trening T_{train} $(N_{train} = 1500)$ i test skup T_{test} $(N_{test} = 500)$
  - Za test skup smo odabrali primere na slučajan način, ne samo poslednje zabeležene dane
  - Ni u jednom trenutku nismo gledali podatke, sve analize smo izvršili automatski

Pitanje



- Data snooping se desio kada smo normalizovali podatke pre podele na trening i test skup
  - Koristili smo srednju vrednost i standardnu devijaciju test skupa!

- Korektno:
  1. Podeliti na T_{train} i T_{test}
  2. Normalizovati T_{train}. Sačuvati μ_{train} i σ^2_{train} kako bismo identičnu transformaciju primenili na test podatke

Data snooping

3. Recikliranje skupa podataka
   - *If you torture the data long enough, it will confess*

- Isprobavamo mnogo obučavajućih algoritama na istom skupu podataka
  - Podelili smo podatke na T_{train} i T_{test}
  - Obučavamo različite modele na T_{train} i evaluiramo ih na T_{test} (E_{test})
  - Kao rezultat vratimo model koji je imao najbolje performanse na T_{test} i kažemo da su njegove performanse E_{test}
- Problem jeste što uvećavamo VC dimenziju, bez da to shvatimo - prava VC dimenzija je unija svih modela koje smo isprobali
- Ovo može da obuhvati i ono šta su drugi probali!
  - Ako koristimo javno dostupan skup podataka i drugi ljudi su već isprobali stvari na njemu
  - Mi pročitamo te radove. Npr. saznamo da se najbolje pokazao SVM sa polinomijalnim kernelom

Data snooping

- Ključni problem u svim primerima je što se prilagođavamo konkretnom skupu podataka – počinjemo da se prilagođavamo šumu koji postoji u njemu

Dva rešenja za data snooping

1. Izbegavanje
   - Disciplina – stavite test podatke u sef i nemojte ga otvarati sve dok nemate finalnu hipotezu
2. Uračunajte i efekat data snooping-a u performanse
   - Kolika je kontaminacija podataka (VC dimenzija, ...)
   - Najteže je uraditi ako ručno pogledamo podatke – teško je modelovati sebe (koliki skup hipoteza je razmatran)

Pitanje: bias via snooping

- Testiranje dugoročnih performansi „buy and hold“ kupovine akcija. Hoćemo da predvidimo kako ćemo proći
  - „buy and hold“ - ne možemo prodavati/menjati u nekom povoljnom trenutku pa kasnije ponovo kupovati, moraju ostati u našem posedstvu od početka do kraja

- Koristićemo podatke od prethodnih 50 godina
  - Hoćemo da test bude što širi – uzmemo sve *Standard & Poor's 500* kompanije (uobičajen reper za U.S. akcije)
  - Pretpostavimo striktno „buy and hold“ model za sve akcije

- Recimo da smo predvideli fantastičan profit. Da li problem postoji?

Pitanje: bias via snooping

- Postoji. Sampling bias – gledamo akcije koje se trenutno preprodaju. Ne razmatramo sve one koje su propale
- Ljudi ovo često ne tretiraju kao sampling bias, već kao data snooping (iako se ne uklapa sasvim u našu raniju definiciju)
    - Jeste snooping – kao da gledamo 50 godina u budućnost i neko nam kaže kojim akcijama se još trguje u toj tački
    - Ali je više sampling bias prouzrokovan pomoću data snooping

Zaključak

- Sva tri koncepta koja smo prešli predstavljaju „zamke“ sa kojima se možemo sresti u primeni mašinskog učenja
  - Npr. Okamova oštrica – vodite računa o kompleksnosti modela koji primenjujete, prilagodite je resursima