

Teorijske osnove nadgledanog učenja

Da li je učenje izvodljivo?

- Tri suštinske komponente prisutne u svakom problemu mašinskog učenja:
 1. Postoji šablon
 2. Šablon ne možemo jednostavno izraziti matematičkom zakonitošću
 3. Postoje podaci

Šta ako ne postoji šablon?

- Možemo pokušati da učimo ali nećemo uspeti...
- Teorija učenja pokazaće nam da možemo primeniti određenu tehniku mašinskog učenja i time utvrditi da li šablon postoji ili ne
 - Postoji mera koja nam govori da li smo nešto naučili ili nismo
- Dakle, čak i ako ne postoji šablon, nema štete od toga da probamo da primenimo mašinsko učenje
 - Nećemo misliti da smo „naučili“ neki šablon koji ne postoji

Šta ako nam je poznata matematička zakonitost?

- Mašinsko učenje će raditi, ali ovo nije optimalan način rešavanja problema
- Možemo direktno da isprogramiramo rešenje i odredimo rezultat perfektno
 - Zbog čega generisati primere, definisati skup hipoteza, pretpostavljati neki model, itd.?
 - Sistem dobijen mašinskim učenjem će gotovo sigurno grešiti na nekim primerima

Šta ako ne postoje podaci?

- U tom slučaju ne možemo baš ništa
- Mašinsko učenje se oslanja na podatke iz kojih može da uči

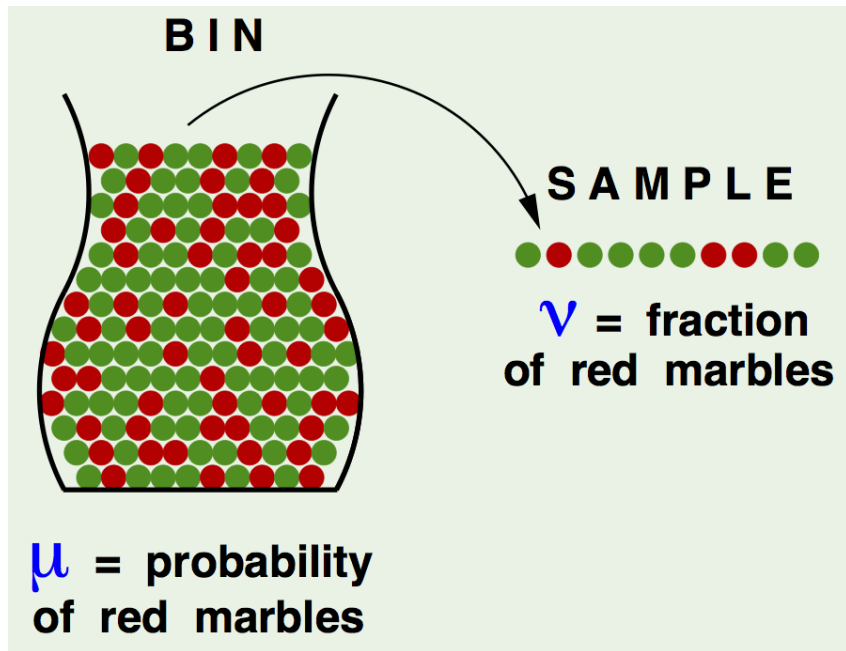
Nadgledano obučavanje

- Fokusiraćemo se na nadgledano obučavanje
 - Imamo nepoznatu ciljnu funkciju $y = f(x)$
 - Imamo skup podataka $T = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$
 - Obučavajući algoritam bira hipotezu $g \approx f$ iz skupa hipoteza \mathcal{H}
- Da li je moguće naučiti nepoznatu ciljnu funkciju?
 - Nije – imamo konačan skup podataka i vrednosti ciljne funkcije samo za njega. Kako možemo reći koja je *stvarna* vrednost ciljne funkcije van tog skupa?
 - Da li je učenje izvodljivo?

Izvodljivost učenja

- Probabilistički model
- Povezanost sa učenjem
- Povezanost sa *stvarnim* učenjem
- Dilema i rešenje

Eksperiment sa klikerima



- Vrednost μ nam je nepoznata konstanta
 - Izabraćemo uzorak od N klikera – deo klikera koji je crven ćemo označiti sa v
 - Različiti uzorci će imati različito v
 - Da li nam v govori nešto o μ ?
-
- Kratak odgovor: ne
 - U uzorku mogu biti svi klikeri crveni, iako u kutiji imamo uglavnom zelene – ne znamo ništa o klikerima koje nismo izabrali
 - Duži odgovor: da
 - Ako je uzorak dovoljno velik, frekvencija v iz uzorka bliska je frekvenciji μ
 - Razlika dva odgovora: moguće naspram verovatnog
 - Iz *probabilističke* perspektive, v nam govori nešto o μ

Eksperiment sa klikerima

$$P[\text{lošeg događaja}] \leq$$

Eksperiment sa klikerima

$$P[|v - \mu| > \varepsilon] \leq$$

Eksperiment sa klikerima

$$P[|v - \mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

Hoeffdingova nejednakost

- Dobra vest: imamo $-N$ u eksponentu
 - Sa porastom uzorka verovatnoća da v loše aproksimira μ (sa tolerancijom ε) brzo opada
- Loša vest: imamo ε^2 u eksponentu
 - Za malu toleranciju ε verovatnoća da v loše aproksimira μ brzo raste
 - Npr. $\varepsilon = 0.1$ (što je izuzetno tolerantno) \rightarrow u eksponentu se N množi sa 0.01 što značajno redukuje efekat velikog uzorka
 - Za $\varepsilon = 10^{-6}$ \rightarrow eksponent pada na nulu, ovo je premala tolerancija za gotovo bilo koji skup podataka!
- Kažemo da je izjava “ $\mu = v$ ” P.A.C. (Probably Approximately Correct)

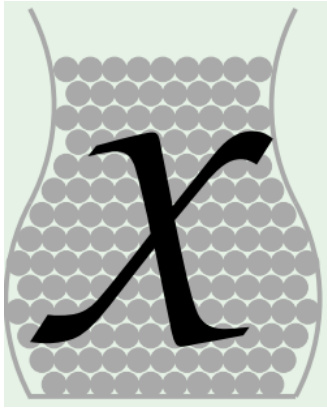
Hoeffdingova nejednakost

$$P[|v - \mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

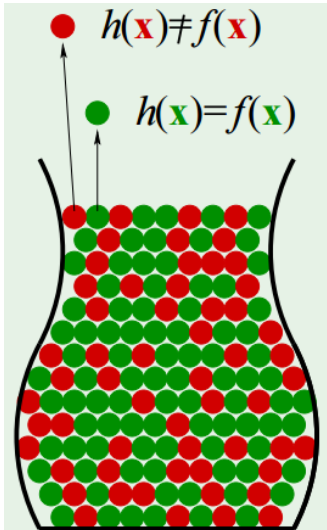
- Validna je za svako N i svako ε veće od 0
- Granica ne zavisi od (konkretne vrednosti) μ (veliĉine koja je nepoznata)
- Tradeoff: N , ε i granica
 - Obiĉno nemamo uticaja na N
 - ε je tolerancija koju mi bираmo – što je veće epsilon, treba nam veći skup podataka
- $v \approx \mu \implies \mu \approx v$
 - Forma verovatnoće je simetriĉna
 - Zapravo μ utiĉe na v
 - Ali mi radimo obrnutu stvar: koristimo v kako bismo procenili μ

Povezanost sa učenjem

- **Eksperiment sa klikerima:** Nepozata je vrednost (broja) μ
- **Učenje:** Nepoznata je funkcija $f: X \rightarrow Y$



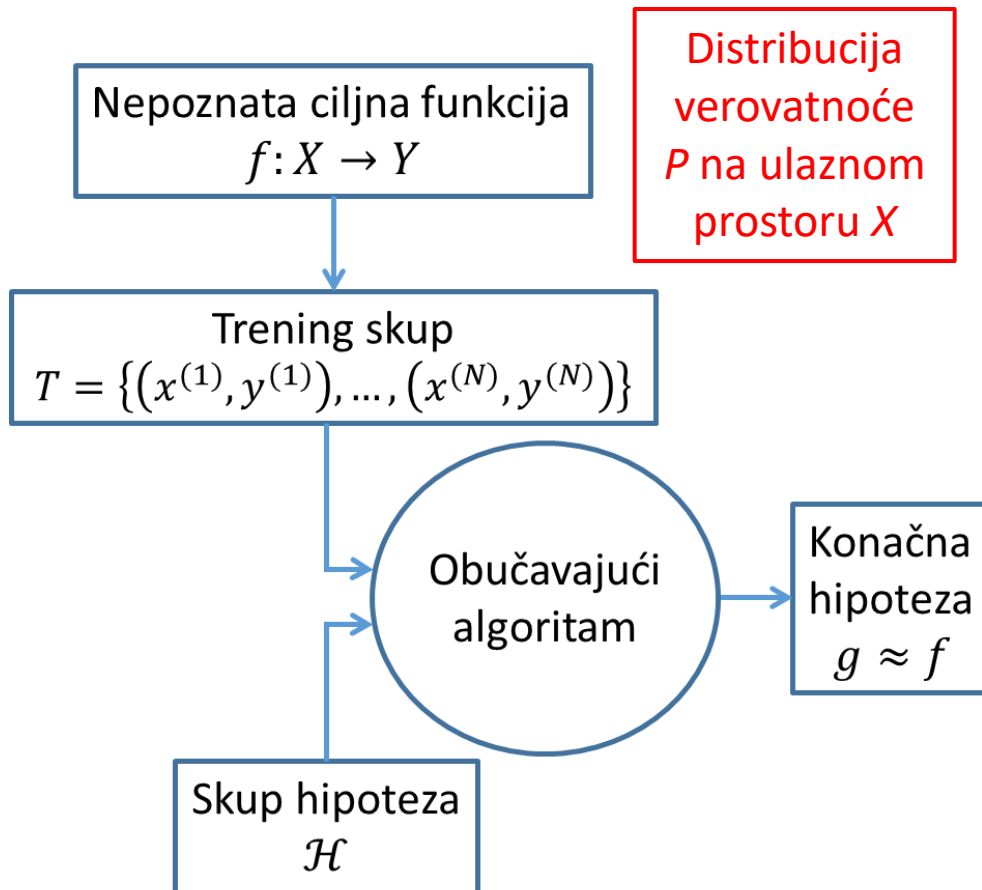
- Kutija sa klikerima \rightarrow ulazni prostor X (*input space*)
- Svaki kliker \rightarrow jedan primer $x \in X$



- : Naša hipoteza je **tačna** za dato x $h(x) = f(x)$
- : Naša hipoteza je **netačna** za dato x $h(x) \neq f(x)$

Ne znamo tačno da kažemo koji je kliker crven a koji zelen – ciljna funkcija je nepoznata! Ovo je samo mapiranje *nepoznato* $\mu \leftrightarrow$ *nepoznata ciljna funkcija* f

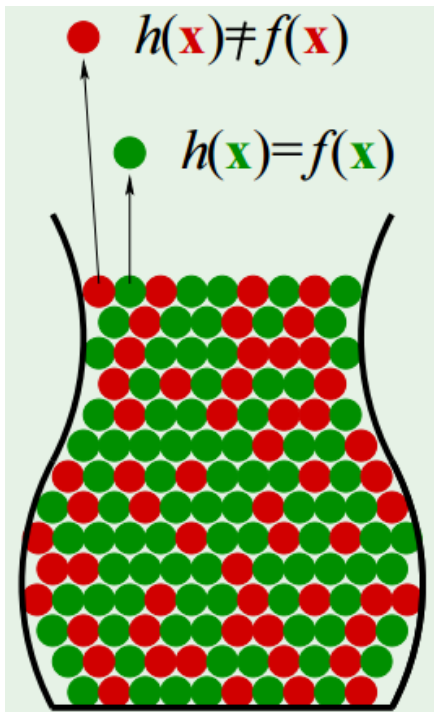
Povezanost sa učenjem



- Distribucija verovatnoće P generiše tačke $x^{(1)}, \dots, x^{(N)}$ **nezavisno** jedna od druge
- Nezavisnost generisanih tačaka je jedina uvedena pretpostavka
 - Nismo uveli pretpostavke o ciljnoj funkciji – može da bude bilo šta
 - Nismo uveli pretpostavke o P – može da bude bilo koja distribucija
- Zahvaljujući Hoeffdingovoj nejednakosti možemo znati granicu performansi

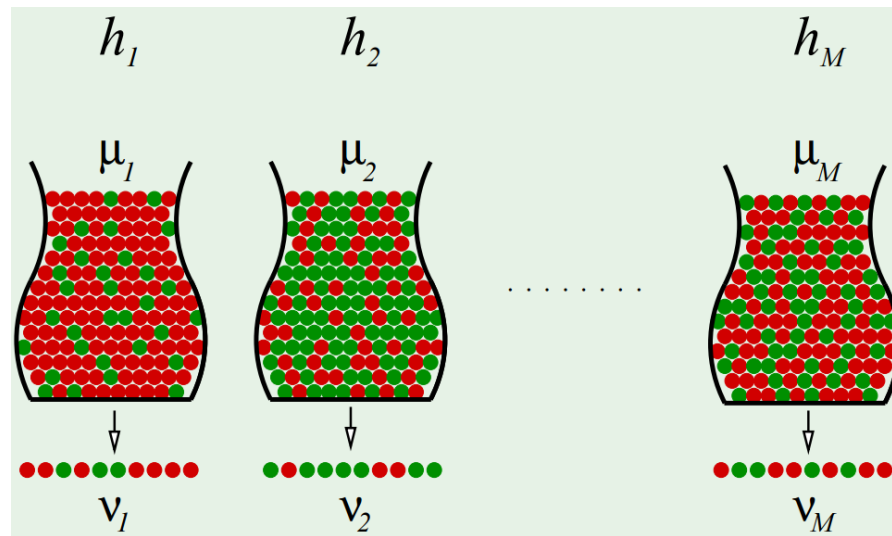
Jedna kutija – verifikacija učenja

$$P[|v - \mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$



- U eksperimentom sa klikerima, hipoteza h je fiksirana
 - Boja klikera zavisi od slaganja između hipoteze h i ciljne funkcije f
 - Malo v u izvučenom uzorku (malo crvenih klikera) govori da konkretno h dobro aproksimira f
 - Hoeffdingova nejednakost nam garantuje da je frekvencija v crvenih klikera jednaka stvarnoj frekvenciji μ crvenih klikera
 - Dakle, nakon što smo odabrali h , možemo da pogledamo podatke i kažemo da li je h dobro ili loše.
- Ono što smo razmatrali je *verifikacija* učenja, a ne samo učenje
- Nema garancije da će v biti malo

Više kutija – učenje



- **Učenje** je pretraga prostora mogućih hipoteza sa ciljem pronađemo onu koja će raditi dobro na podacima
 - Imamo više mogućih **hipoteza** h
 - Obratiti pažnju da je h hipoteza (konkretna funkcija) a ne model
 - Model je prostor hipoteza koji pretražujemo, npr. model konstante može biti $y=3$ ili $y=1$
 - Koristićemo broj mogućih hipoteza kao meru kompleksnosti modela
 - Pogledaćemo uzorak generisan za svaku hipotezu
 - Odabraćemo hipotezu g za koju je v najmanje

Notacija za učenje

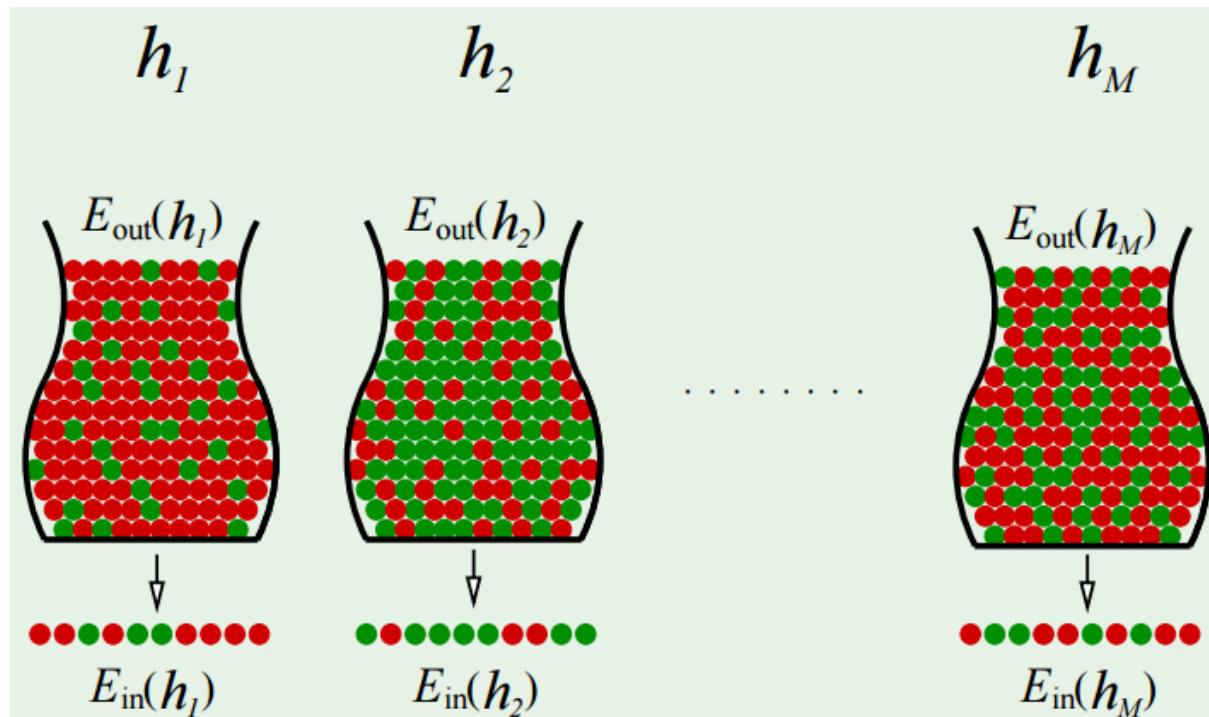
- ν i μ zavise od hipoteze h
 - ν : *na uzorku* (*in sample*) ćemo označiti sa $E_{in}(h)$ (greška na uzorku)
 - μ : *van uzorka* (*out of sample*) ćemo označiti sa $E_{out}(h)$ (greška van uzorka)
- Ako je $E_{in}(h)$ malo, to znači da je greška na *datom uzorku* mala
- Ako je $E_{out}(h)$ malo, to znači da je greška na *neuočenim podacima* mala
- Ono što je naš cilj jeste malo $E_{out}(h)$
 - Ako imamo dobre performanse na nečemu što nismo videli dok smo učili – znači da smo zaista nešto naučili
- *Hoeffdingova nejednakost* ($P[|\nu - \mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$):
$$P[|E_{in}(h) - E_{out}(h)| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

Notacija za učenje

- Hoeffdingova nejednakost

$$P[|E_{in}(h) - E_{out}(h)| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}$$

Verovatnoća da performanse hipoteze dobijene na uzorku odstupaju od performansi van uzorka više od propisane tolerancije epsilon je manja ili jednaka broju koji je (nadamo se) mali

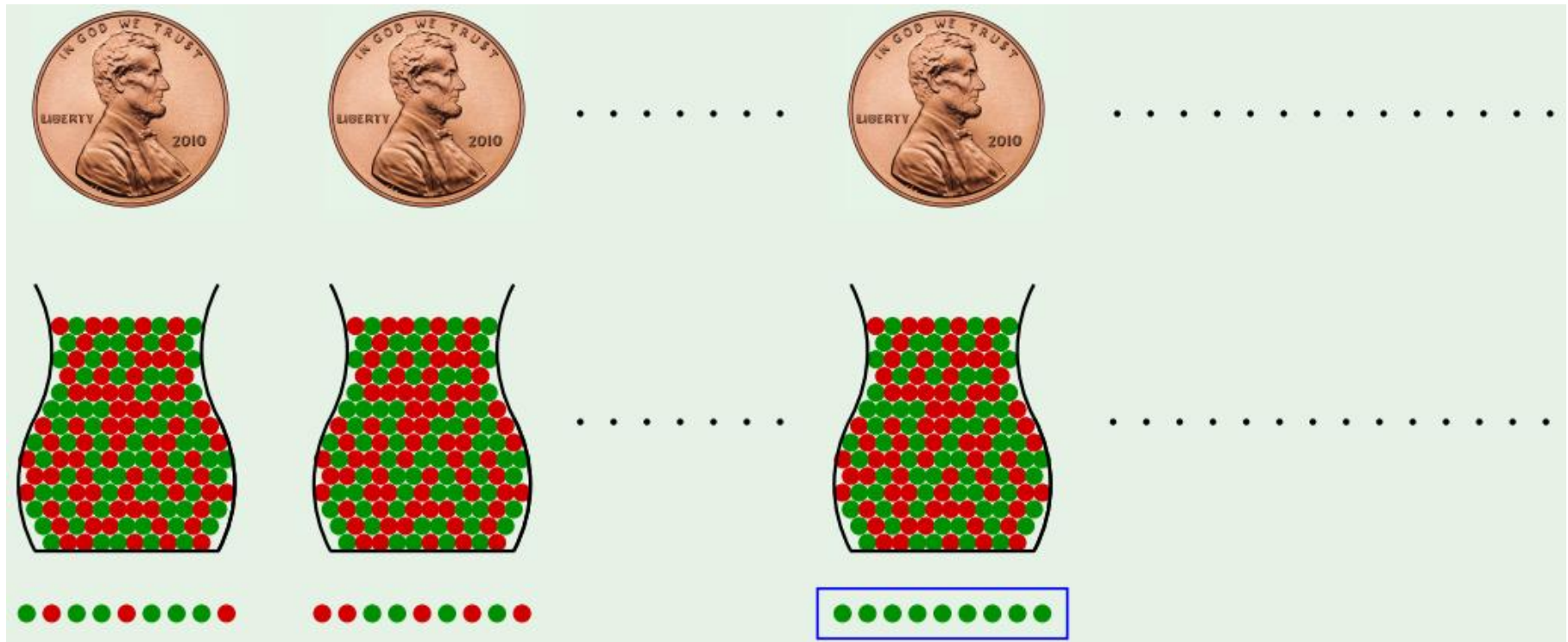


Dilema

- Hoeffdingova nejednakost nije primenljiva u slučaju sa više kutija
- Analogija sa novčićem:
 - Ako bacimo *jedan novčić* 10 puta, koja je verovatnoća da 10 puta padne glava?
 $0.5^{10} \approx 0.1\%$ (jednom u 1000 puta)
 - Ako bacimo *1000 novčića* 10 puta svaki, koja je verovatnoća da će *jedan* od novčića 10 puta pasti na glavu?
 $1 - (1 - 0.5^{10})^{1000} \approx 63\%$ - verovatnije da će se desiti nego da neće. Dakle, 10 glava u ovom slučaju nije indikacija stvarne verovatnoće

Dilema

- Hoeffdingova nejednakost se odnosi na svaki od pojedinačnih novčića. Ali postoji mala verovatnoća da je $\nu \neq \mu$ – ako ponovimo eksperiment dovoljno puta – postoji značajna verovatnoća da će se ovaj slučaj desiti



Bulova nejednakost (*union bound*)

- Neka su A_1, A_2, \dots, A_M M različitih događaja. Tada važi:

$$P(A_1 \cup \dots \cup A_M) \leq P(A_1) + \dots + P(A_M)$$

- Verovatnoća da se desi barem jedan od M različitih događaja je najviše suma verovatnoća pojedinačnih događaja

Jednostavno rešenje

- g – naša konačna hipoteza (odabrana među M mogućih h_1, \dots, h_M)
- Koja je verovatnoća da je hipoteza g loša? Iskoristićemo Bulovu nejednakost:

$$P[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq P \left[\begin{array}{c} |E_{in}(h_1) - E_{out}(h_1)| > \varepsilon \\ \vee |E_{in}(h_2) - E_{out}(h_2)| > \varepsilon \\ \dots \\ \vee |E_{in}(h_M) - E_{out}(h_M)| > \varepsilon \end{array} \right]$$
$$\leq \sum_{i=1}^M P[|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon] \leq \sum_{i=1}^M 2e^{-2\varepsilon^2 N}$$

Jednostavno rešenje

$$P[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq \sum_{i=1}^M 2e^{-2\varepsilon^2 N}$$

- Sada imamo gornju granicu za verovatnoću da je *greška na uzorku* hipoteze g (odabrane među M razmatranih hipoteza) bliska *grešci van uzorka* hipoteze g

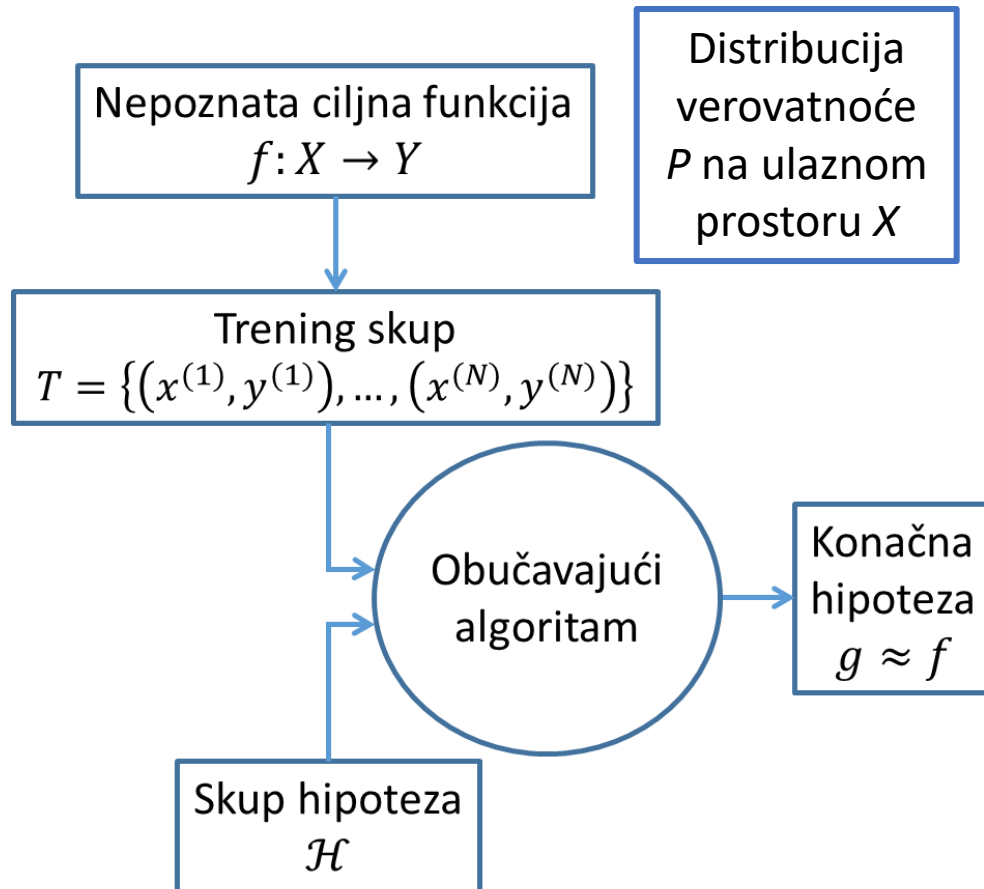
$$P[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2\textcolor{red}{M}e^{-2\varepsilon^2 N}$$

- Ovo nam govori da što je model prilagodljiviji (veće M) – to više gubimo vezu između greške dobijene na uzorku i greške van uzorka
 - Prilagodljiv model može da se izuzetno dobro prilagoditi trening skupu, a da pritom ne generalizuje dobro van uzorka
 - Za sada smo (radi jednostavnosti) razmatrali konačno M , ali kod gotovo svih modela M je beskonačno

Zaključak

- Da li je učenje izvodljivo?
 - Za jako prilagodljive modele gornja granica verovatnoće je beznačajna
- Još nismo utvrdili konačan rezultat učenja
- Utvrdili smo princip da kroz učenje možemo da generalizujemo
- Kasnije ćemo videti granicu generalizacije kroz teoriju

Osnovna teorijska postavka nadgledanog učenja



- Odnos između x i y je određen probabilističkim zakonom $P(x, y)$
- Cilj: odrediti „najbolju“ funkciju g takvu da važi $g \approx f$
- Šta znači „ $g \approx f$ “?
- Uvešćemo funkciju greške (eng. *loss*) $E(g, f)$ koja kvantifikuje koliko hipoteza g odstupa od ciljne funkcije f

Funkcija greške

- Funkcija greške $E(g, f)$ se gotovo uvek definiše preko razlike u konkretnoj tački $e(h(x), f(x))$

- *Squared error loss*

$$e(h(x), f(x)) = (h(x) - f(x))^2$$

- *Binary error loss*

$$e(h(x), f(x)) = \begin{cases} 1, & h(x) \neq f(x) \\ 0, & h(x) = f(x) \end{cases}$$

Funkcija greške

- Globalna greška $E(g, f)$ se računa kao prosek $e(g(x), f(x))$

- Greška na uzorku:

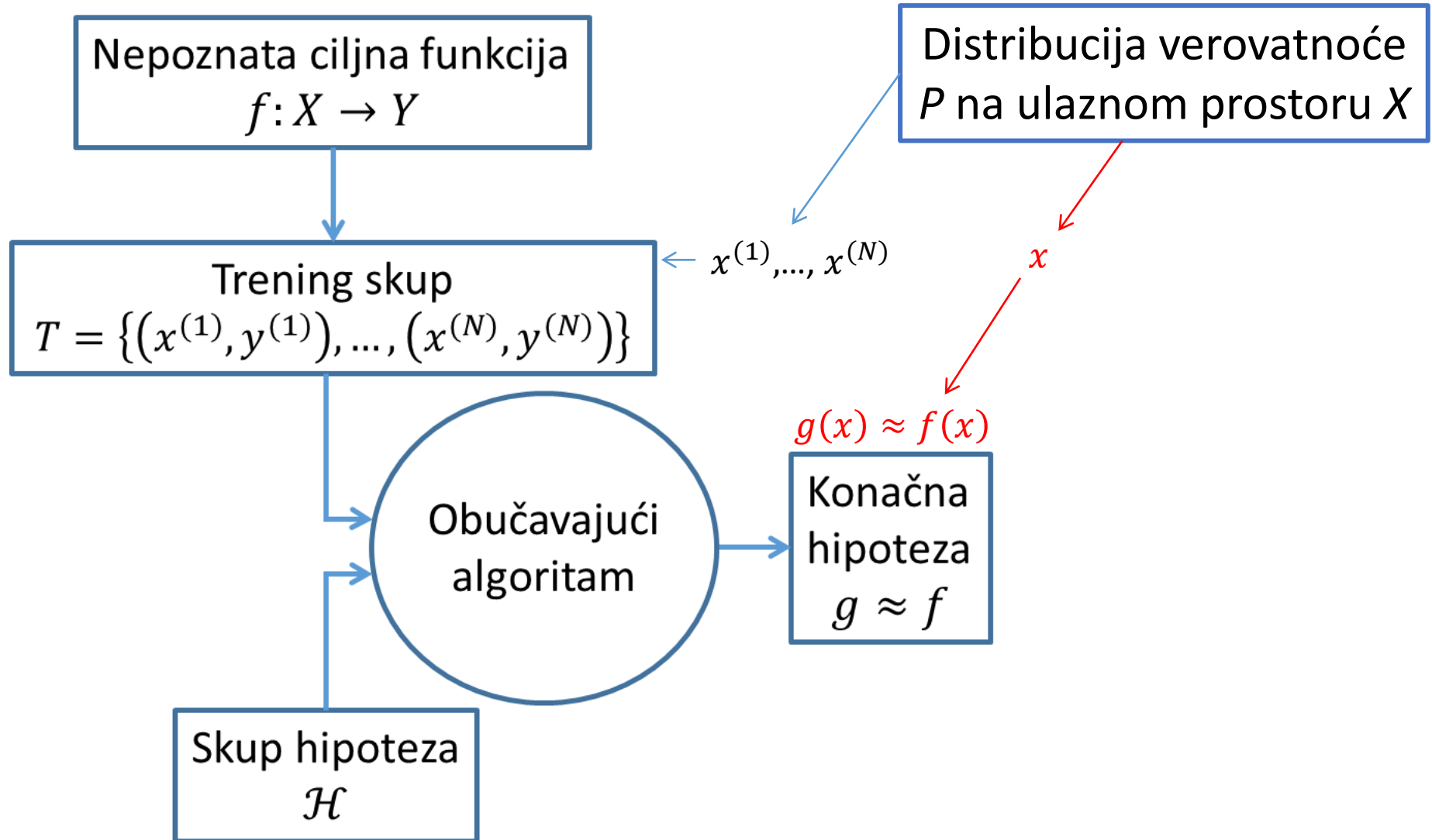
$$E_{in}(g) = \frac{1}{N} \sum_{n=1}^N e(g(x_n), f(x_n))$$

- Greška van uzorka:

$$E_{out}(g) = \mathbb{E}_x[e(g(x_n), f(x_n))]$$

- Potrebno je odrediti funkciju g za koje je $E_{out}(g)$ najmanje

Mera greške



Kako odabrati meru greške?

- Primer: identifikacija pomoću otiska prsta



$$f = \begin{cases} 1, & \text{pristup odobren} \\ 0, & \text{uljez} \end{cases}$$

- Dve vrste greške
 - False accept*
 - False reject*

		f	
		+1	-1
g	+1	No error	False accept
	-1	False reject	No error

- Kako penalizovati ove greške?

Kako odabrati meru greške?

- Ne postoji analitički način. Ovo je pitanje domena primene

- **Supermarket:** potvrđuje otisak prsta radi popusta

- *False accept*: nije veliki problem – ako damo jedan popust viška, to ne utiče previše na poslovanje
- *False reject*: problematičan – možemo izgubiti mušteriju

		f	
		+1	-1
g	+1	0	1
	-1	10	0

- **CIA:** potvrđuje otisak prsta radi sigurnosti

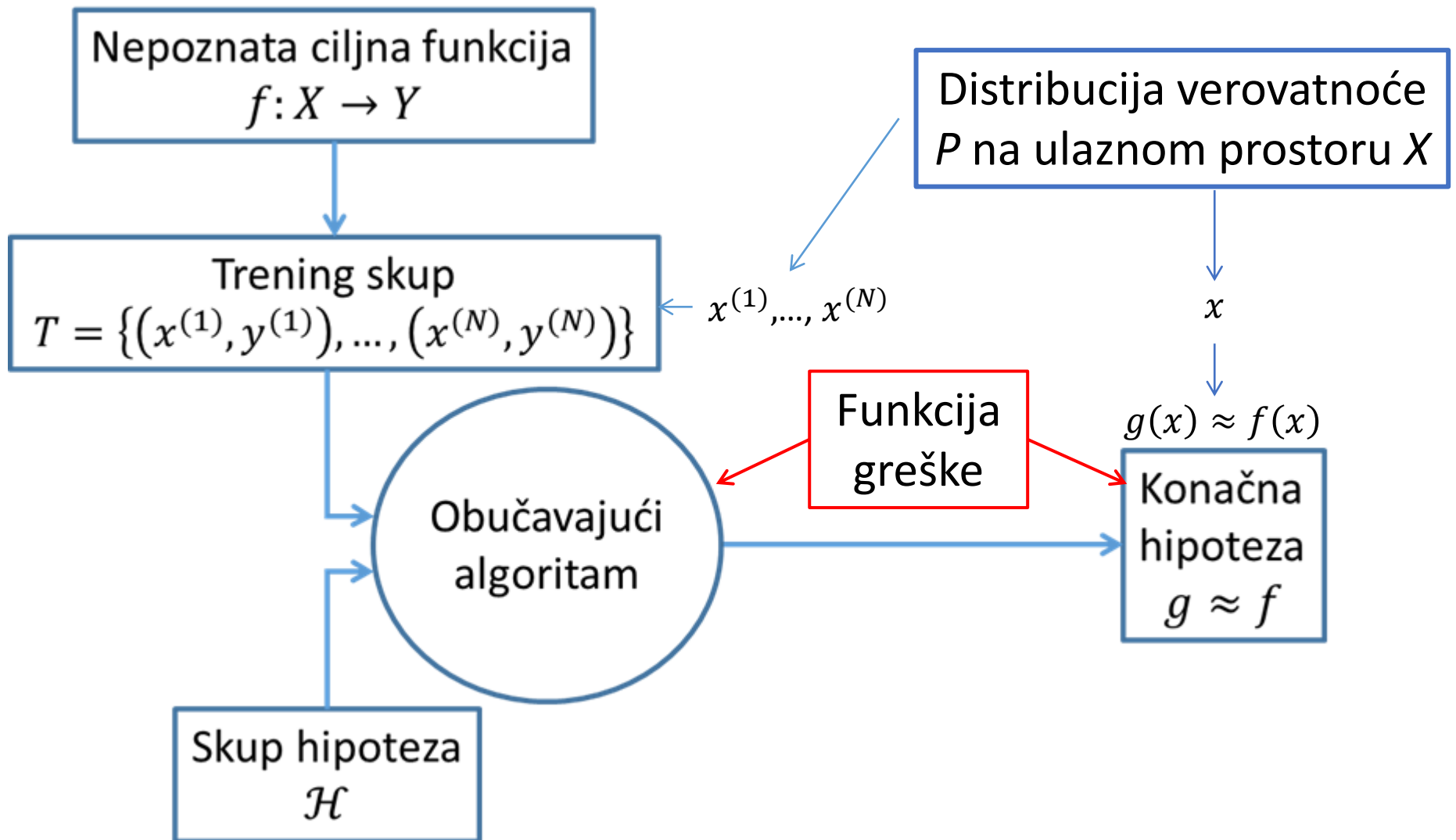
- *False accept*: izuzetno velik problem – neko je dobio pristup osetljivim podacima a nije trebao
- *False reject*: možemo tolerisati – zaposleni mogu pokušavati više puta da se prijave na sistem

		f	
		+1	-1
g	+1	0	1000
	-1	1	0

Kako odabrati meru greške?

- Kada radimo na praktičnom problemu mera greške treba da bude **specificirana od strane korisnika**
- Ovo nije uvek moguće
 - Nekada je mušteriji teško da formalizuje ovu grešku
 - Čak i ako je formalizuju, može biti prekompleksna za optimizaciju
- Alternative:
 - Konceptualno verodostojne mere – možemo analitički argumentovati da su dobre
 - Npr. *squared error* pod pretpostavkom da šum prati normalnu raspodelu
 - Praktične mere
 - Npr. zahvaljujući *squared error* smo u stanju da pronađemo *closed form solution*
 - *Squared error* je jednostavna za korišćenje i ukoliko koristimo GD jer rezultuje konveksnom optimizacijom (globalni optimum)

Komponente obučavanja

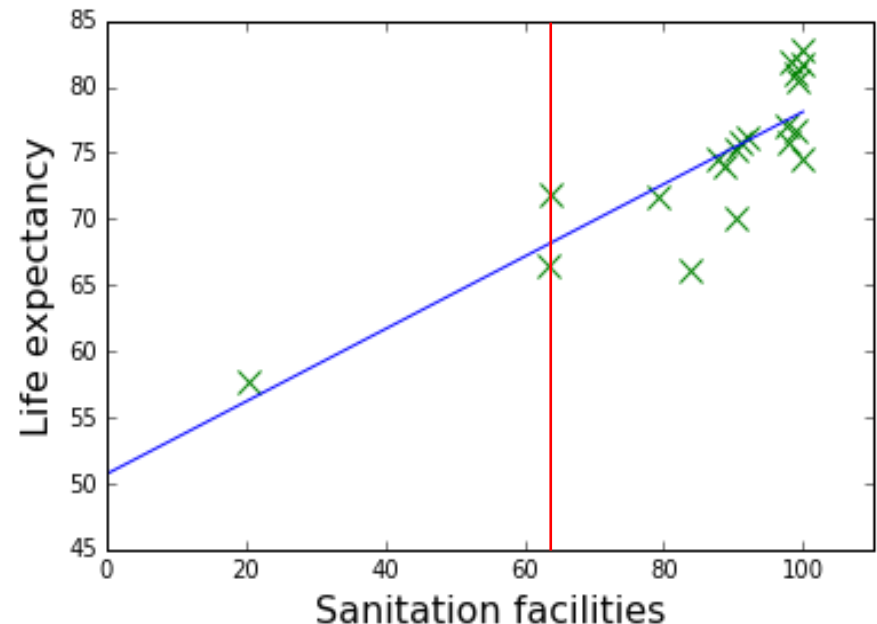


- Ciljna funkcija nije uvek *funkcija*
 - U matematičkom smislu funkcija $f(x)$ bi trebala da ima jedinstvenu vrednost za svaku tačku x

Dodela kredita

starost	23 godine	23 godine
pol	Muški	Muški
Godišnja zarada	\$30 000	\$30 000
Trenutni dug	\$15 000	\$15 000
Posедује nekretninu	Da	Da
Odobriti kredit	Da	Ne

Predikcija životnog veka

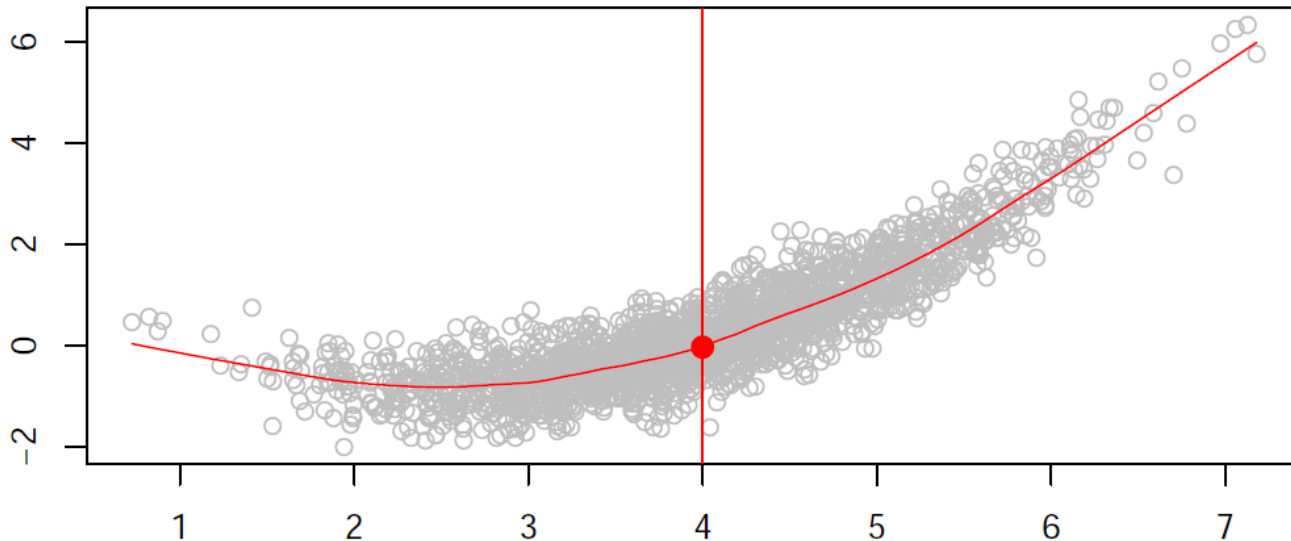


Šum

- Podaci suštinski sadrže šum
 - Slučajne greške prilikom merenja
 - Druge (neuočene) varijable koje utiču na y
- Naša pretpostavka je da šum ima srednju vrednost 0
 - Npr. kod regresije smo pretpostavili $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$
 - Da to nije tako (npr. šum raste linearno sa porastom x), ovo bi bilo uočeno kao šablon i kao takvo uključeno u našu hipotezu

Ciljna distribucija (*target distribution*)

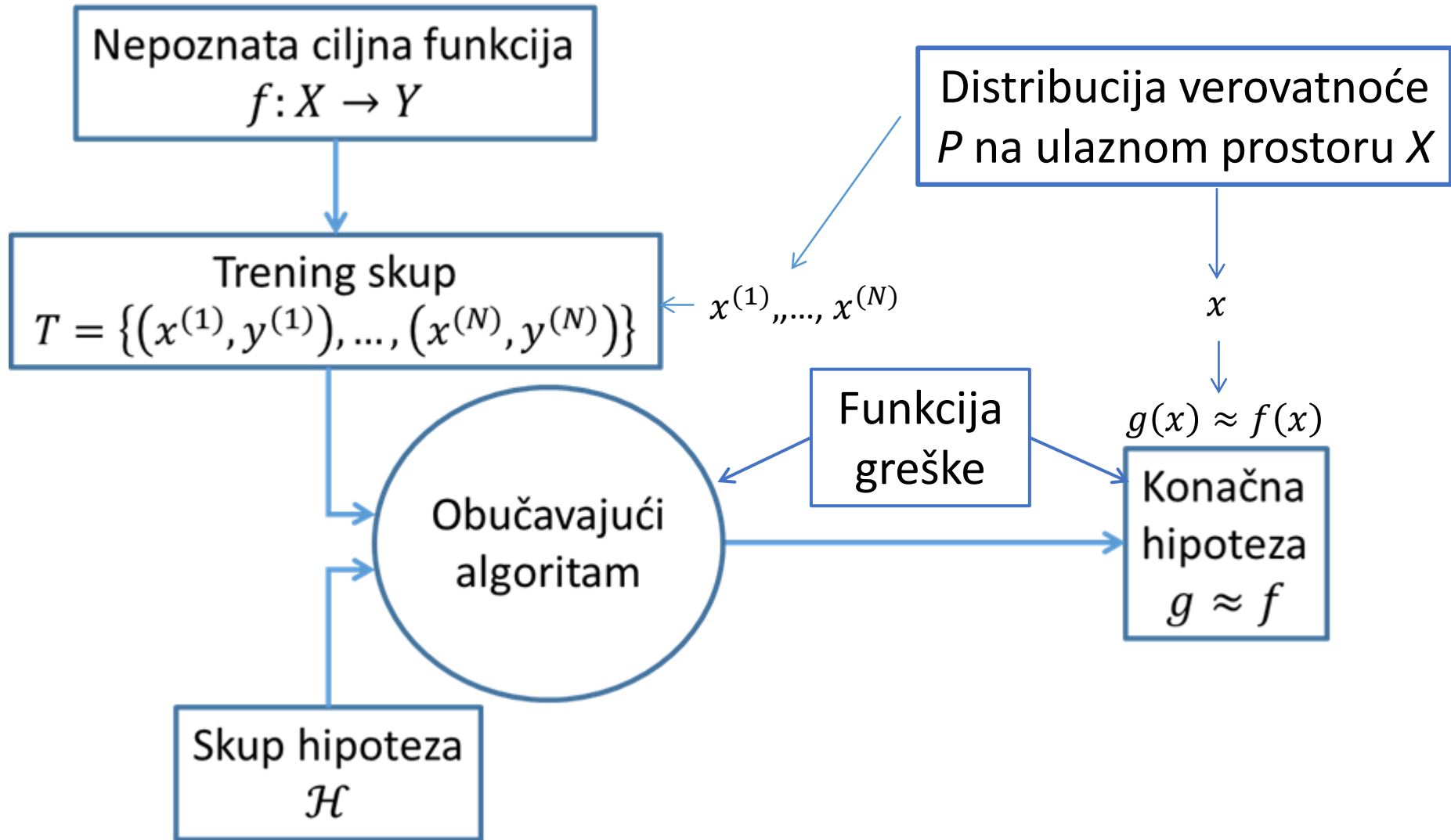
- Umesto $y = f(x)$ koristimo ciljnu *distribuciju*:
- Instance (x, y) su generisane iz združene (*joint*) distribucije:
 $P(x)P(y|x)$
- *Noisy target* =
 Deterministic target $f(x) = \mathbb{E}(y|x)$ +
 Noise $y - f(x)$



Ciljna distribucija (*target distribution*)

- $P(y|x)$ - x ima uticaja na y , ali to nije striktno deterministička veza
 - Ranije smo imali zavisnost $y = f(x)$, tj. za jedno x , moguće je samo jedno y , sve ostalo je nemoguće
 - Sada, imamo probabilističku zavisnost – ako je dato neko x , neki y -oni su verovatniji
- Obučavanje:
 - Mi obučavamo algoritam da pokupi šablon $E(y|x)$
 - $E(y|x)$ je (naša ciljna) funkcija
 - Ostalo proglašavamo šumom – nema ništa što možemo da uradimo po pitanju šuma

Komponente obučavanja



Komponente obučavanja

Nepoznata ciljna distribucija

$$P(y|X)$$

Ciljna funkcija: $f: X \rightarrow Y + \text{šum}$

Distribucija verovatnoće
 P na ulaznom prostoru X

Trening skup
 $T = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$

$x^{(1)}, \dots, x^{(N)}$

x

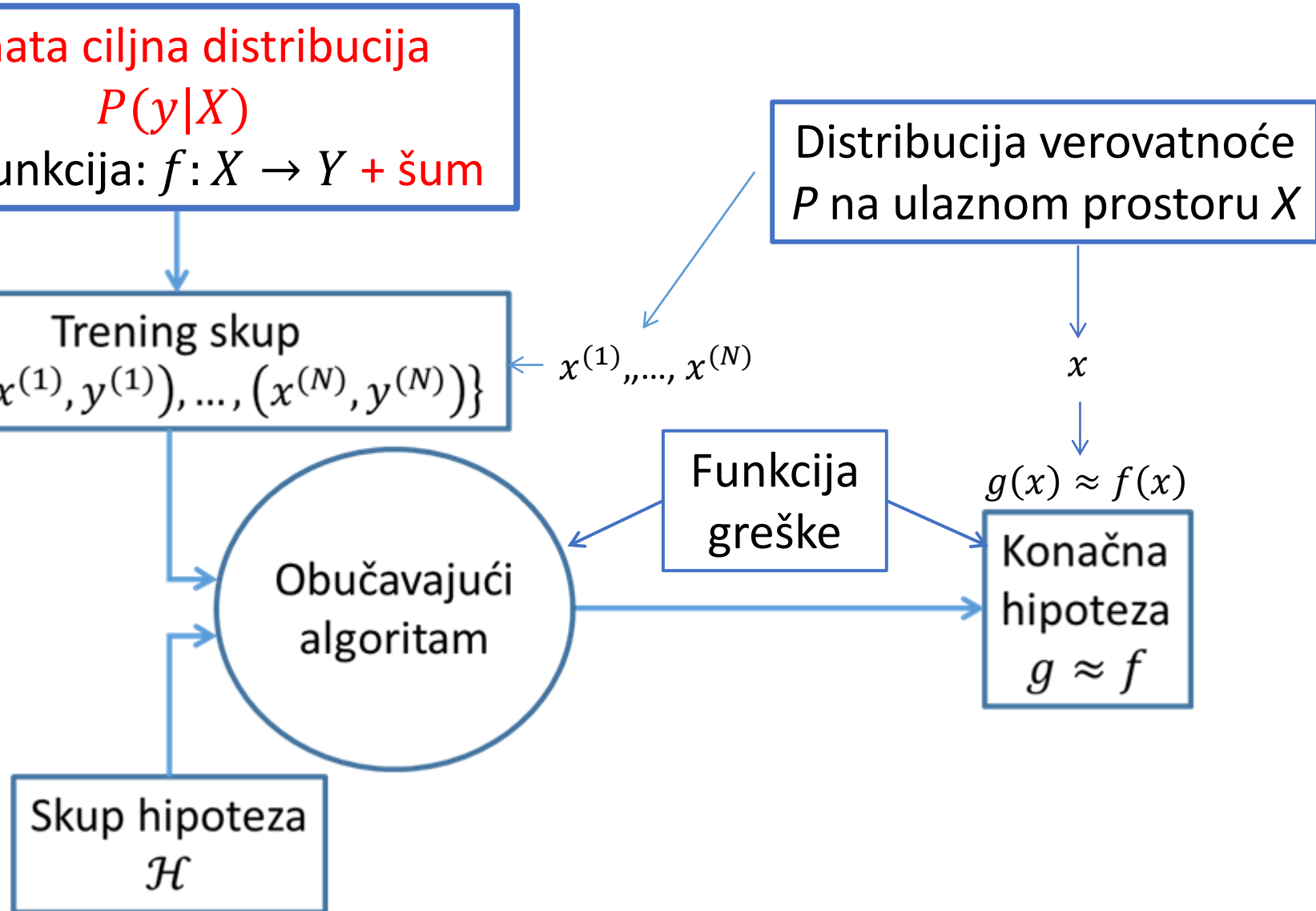
$$g(x) \approx f(x)$$

Funkcija
greške

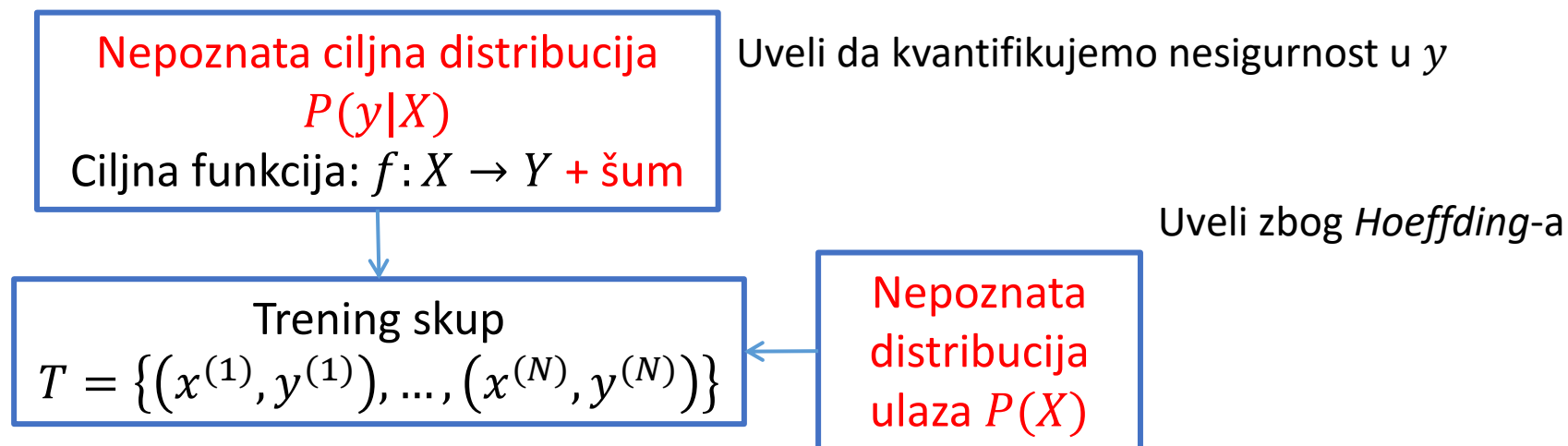
Obučavajući
algoritam

Konačna
hipoteza
 $g \approx f$

Skup hipoteza
 \mathcal{H}



Razlika $P(y|x)$ i $P(x)$



- Distribucija ulaza $P(x)$ kvantifikuje relativnu važnost ulaza x
 - Npr. odlučujemo da li da dodelimo kredit ili ne (y) na osnovu plate (x)
 - Ulazna distribucija $P(x)$ nam govori koja je raspodela plata u opštoj populaciji – koliko ljudi zarađuje platu od \$30 000, koliko sa \$100 000, itd.
 - A ono što želimo da naučimo jeste $P(y|x)$: da li odobriti kredit ili ne za datu platu x
- Združena distribucija $P(x, y) = P(y|x)P(x)$ nije cilj za nadgledano obučavanje
- Ciljna distribucija $P(y|x)$ je ono što želimo da naučimo

Šta znamo do sada?

- Učenje je izvodljivo. Verovatno je da važi:

$$E_{out}(g) \approx E_{in}(g)$$

- Da li je ovo uslov koji znači učenje?
 - Treba nam $g \approx f$, tj. $E_{out}(g) \approx 0$
- $E_{out}(g) \approx 0$ je ono što želimo, a $E_{out}(g) \approx E_{in}(g)$ ono što imamo
 - Ono što je $E_{out}(g) \approx E_{in}(g)$ ustvari jeste dobra generalizacija
 - Ako bi nam cilj bio $E_{out}(g) \approx 0$ (izjednači nepoznatu veličinu sa nulom) ne možemo ništa da uradimo
 - Sa $E_{out}(g) \approx E_{in}(g)$ možemo reći da pomoću $E_{in}(g)$ možemo dobiti „prozor“ za $E_{out}(g)$

Dva pitanja učenja

- $E_{out}(g) \approx 0$ se postiže kroz dva uslova:

$$E_{out}(g) \approx E_{in}(g) \text{ i } E_{in}(g) \approx 0$$

- **Učenje možemo podeliti na dva pitanja:**
 1. **Da li možemo da se osiguramo da je $E_{out}(g)$ dovoljno blizu $E_{in}(g)$?**
 2. **Da li možemo dovoljno da smanjimo $E_{in}(g)$?**
- Pitanje 1 je teorijsko
- Pitanje 2 je praktično – odrediti model tako da greška bude „što je moguće manja“

Šta znači da je E_{in} dovoljno malo?

- Do sada smo pominjali da treba da je blizu 0 ali...
- Postoje mnoge aplikacije gde ne možemo ni sanjati da ćemo E_{in} i E_{out} približiti nuli
- Npr. predviđanje cena akcija
 - Podaci sadrže veoma veliki šum
 - Ako bismo svega 53% vremena bili u pravu *konzistentno* bili bismo veoma srećni – pod tim uslovima bismo mogli vremenom zaraditi puno para>
 - Dakle, u ovom slučaju E_{out} je blizu 50% - u nekim aplikacijama “dovoljno malo” ne mora da znači 0
- Sve dok znamo (ili barem imamo teoretske garancije pomoću Hoeffding-a) da je greška manja od 50% *konzistentno* dobro je

Teorija učenja

- Pomaže nam da steknemo intuiciju kako da najbolje primenimo obučavajuće algoritme u različitim problemima
- Okarakterisaćemo izvodljivost učenja za slučaj kada je M (broj mogućih hipoteza) beskonačan
- Formalizovaćemo nagodbu:
 - Kompleksnost modela \uparrow : $E_{in} \downarrow$
 - Kompleksnost modela \uparrow : $E_{out} - E_{in} \uparrow$

Training vs. testing

- Testiranje:

$$P[|E_{in} - E_{out}| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- Obučavanje:

$$P[|E_{in} - E_{out}| > \epsilon] \leq 2M e^{-2\epsilon^2 N}$$

- M – broj hipoteza
- Čak i za jednostavne modele poput linearne regresije M je beskonačno, što čini da ova nejednakost ne daje nikakvu garanciju
- Da bismo govorili o izvodljivosti učenja, moramo zameniti M u nejednakosti sa veličinom koja neće biti beskonačna kada je broj mogućih hipoteza beskonačan

Odakle nam M ?

- „Loši“ događaji \mathcal{B}_m su:

$$„|E_{in}(h_m) - E_{out}(h_m)| > \epsilon“$$

- Ono što želimo da možemo tvrditi jeste da je verovatnoća ovakvih događaja mala:

$$P[\mathcal{B}_1 \vee \mathcal{B}_2 \vee \dots \vee \mathcal{B}_M] \leq ?$$

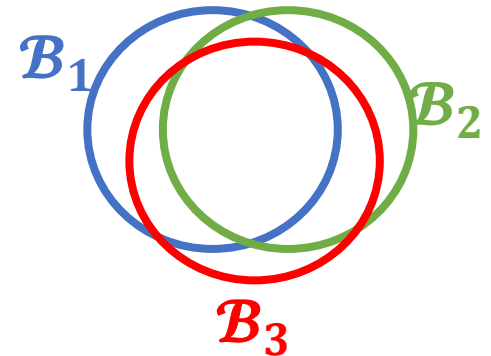
- \mathcal{B}_i - desio se „loš“ događaj za i -tu hipotezu
- Želimo da $P[\mathcal{B}_1 \vee \mathcal{B}_2 \vee \dots \vee \mathcal{B}_M]$ bude malo jer naš algoritam može da izabere bilo koju od ovih M hipoteza

Odakle nam M ?

- Događaji $\mathcal{B}_1, \dots, \mathcal{B}_M$ mogu biti u različitoj međusobnoj korelaciji – nezavisni, isključivi, podudarajući,...
- Želimo da damo gornju granicu za $P[\mathcal{B}_1 \vee \mathcal{B}_2 \vee \dots \vee \mathcal{B}_M]$ nezavisno od korelacije ovih događaja – koristimo Bulovu nejednakost:

$$P[\mathcal{B}_1 \vee \mathcal{B}_2 \vee \dots \vee \mathcal{B}_M] \leq P(\mathcal{B}_1) + P(\mathcal{B}_2) + \dots + P(\mathcal{B}_M)$$

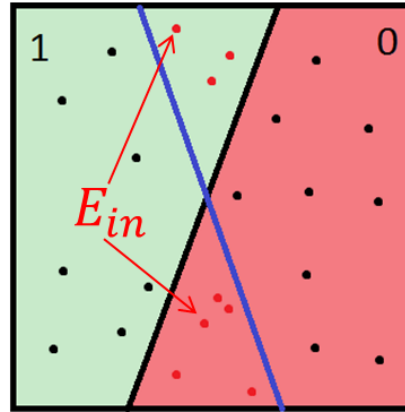
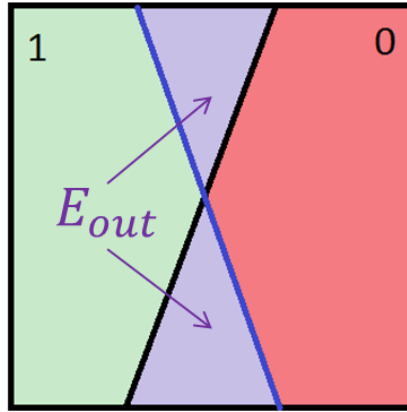
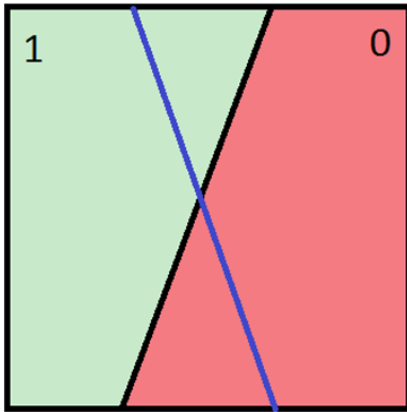
- $P(\mathcal{B}_1) + P(\mathcal{B}_2) + \dots + P(\mathcal{B}_M)$ je najgori slučaj kada su ovi događaji disjunktni (nema preklapanja)
- Ali ovo je prilično nezgrapna granica – npr. u slučaju prikazanom na slici daje gotovo 3 puta veću oblast



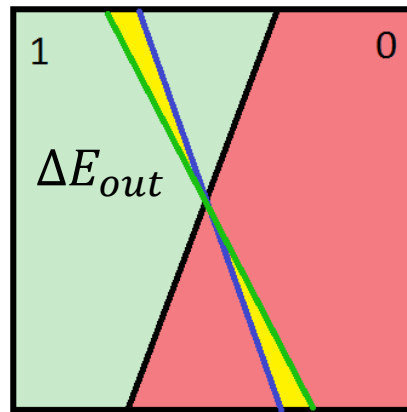
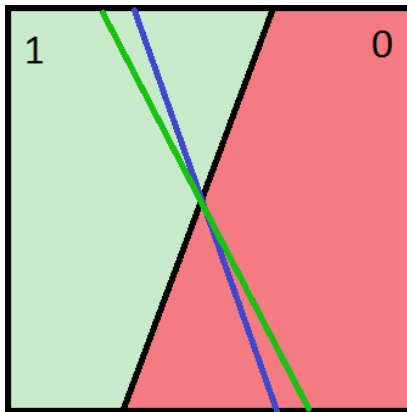
Možemo li bolje od M ?

- Da, loši događaji se u praksi prilično preklapaju!

Hipoteza 1

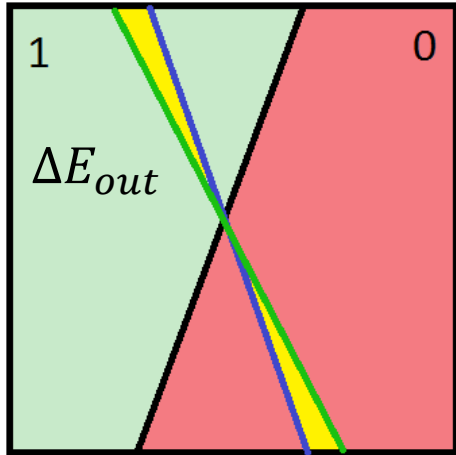


Hipoteza 2



ΔE_{in} : promena labele jedne od tačaka uzorka (tačke koje upadnu u žutu regiju)

Možemo li bolje od M ?

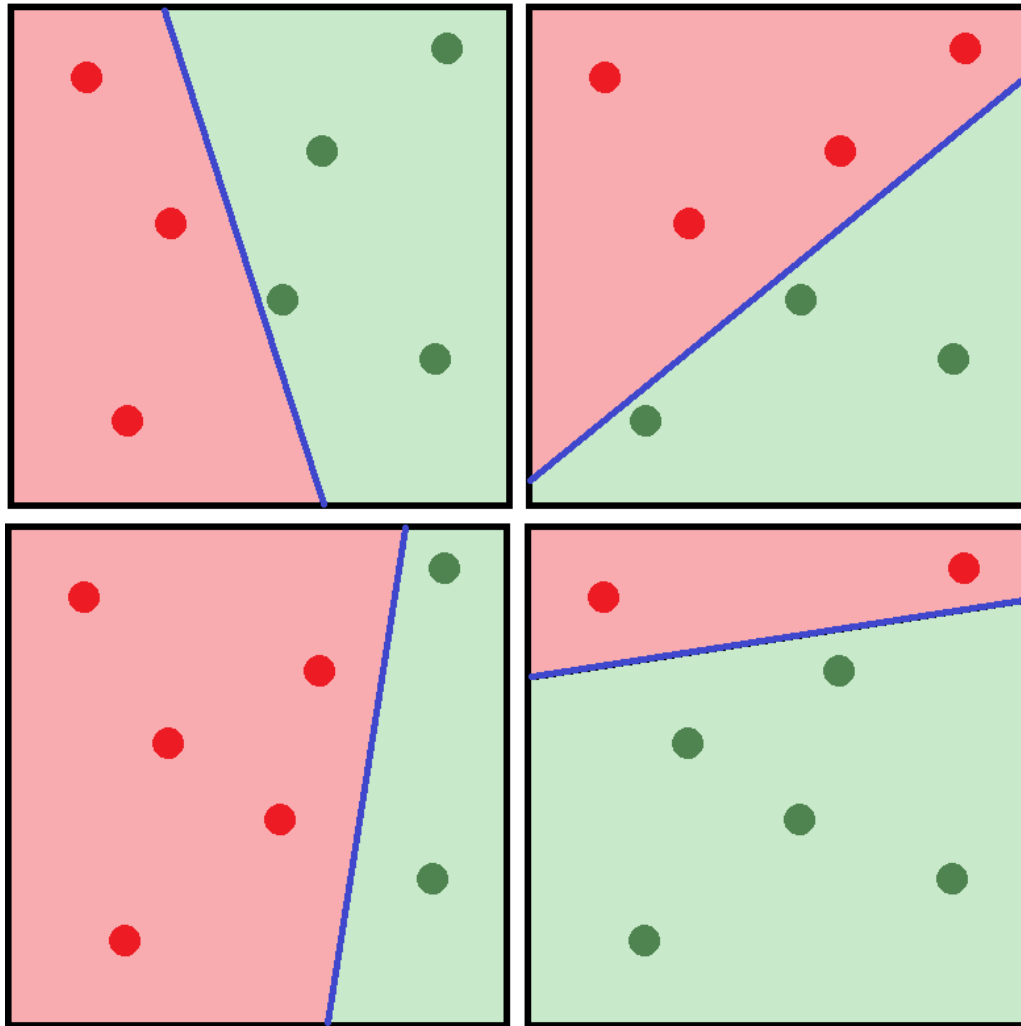


- Prelaskom sa hipoteze 1 na hipotezu 2:
 - ΔE_{out} (žuta regija) je malo
 - ΔE_{in} je takođe malo jer je verovatnoća da neka od tačaka uzorka promeni labelu (upadne u žutu regiju) mala
 - Štaviše, sa porastom regije raste i verovatnoća da će neka se neka od tačaka uzorka naći u regiji pa su ΔE_{out} i ΔE_{in} u korelaciji

$$|E_{in}(h_1) - E_{out}(h_1)| \approx |E_{in}(h_2) - E_{out}(h_2)|$$

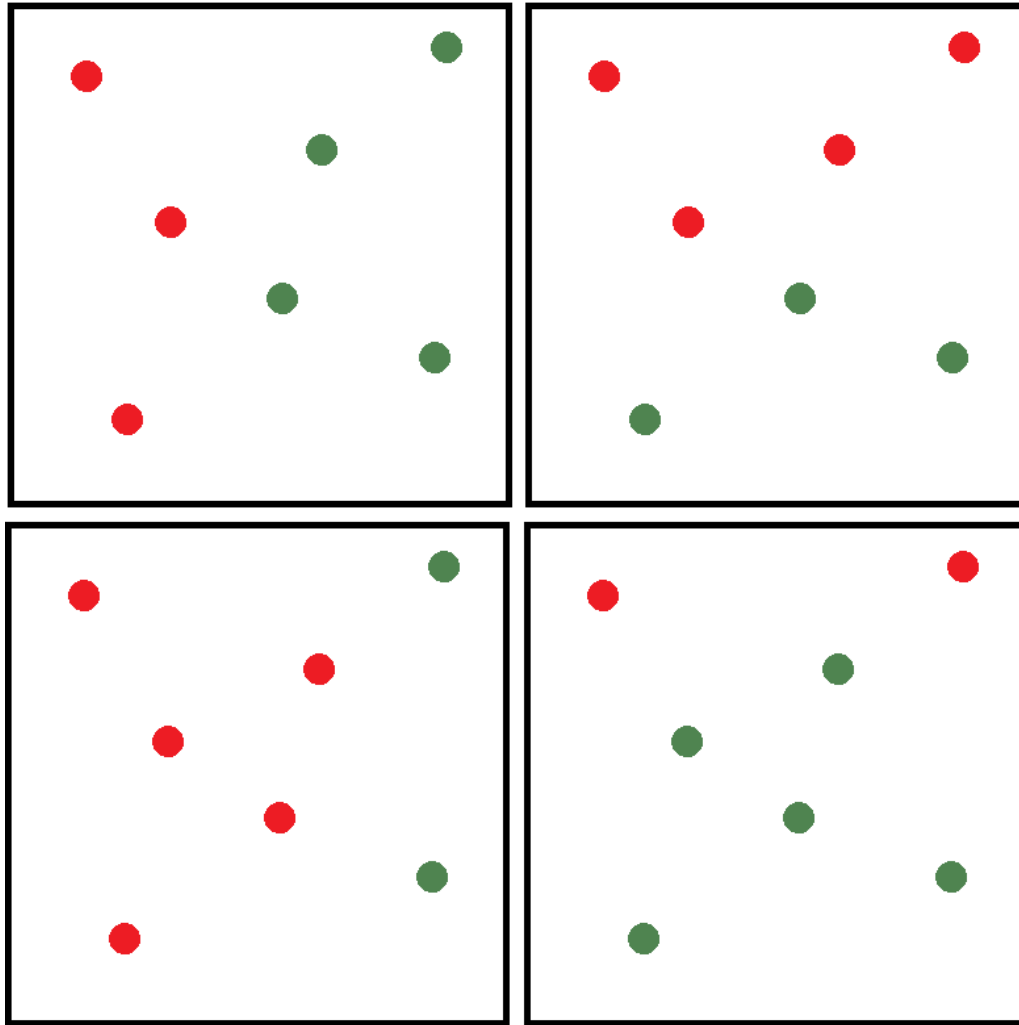
- Ako bi se desio $\mathcal{B}_1(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon)$, velika je verovatnoća da će se desiti i \mathcal{B}_2 - loši događaji se često preklapaju
- Dakle, sigurno možemo bolje od Bulove nejednakosti gde računamo da su ovi događaji disjunktni (i broj hipoteza M figuriše)

Sa čime zameniti M ?



- Sve četiri hipoteze su različite jer dodeljuju različite labele na barem jednoj tački ulaznog prostora
- Pošto je ulazni prostor kontinualan (beskonačan), imamo beskonačno mnogo hipoteza
- Umesto celog ulaznog prostora, razmatraćemo samo uzorak – konačan skup tačaka
- Hipoteze ćemo zvati dihotomije jer posmatramo samo konačan broj tačaka

Sa čime zameniti M ?



- Za ovaj konkretan uzorak, koliko različitih kombinacija crvenih i zelenih tačaka možemo dobiti?
 - Ako imamo model koji može da rezultuje svim mogućim kombinacijama crvenih/zelenih tačaka – to je veoma sofisticiran model
 - Ako model može da proizvede svega nekoliko kombinacija – ovo nije toliko sofisticiran model
- Ne brojimo hipoteze kako bi trebalo jer smo ograničeni samo na ovaj konačan skup tačaka ali ipak možemo da poredimo modele

Dihotomije: mini-hipoteze

- Hipoteza $h: X \rightarrow \{0, 1\}$
- Dihotomija $h: \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\} \rightarrow \{0, 1\}$
- Broj hipoteza $|\mathcal{H}|$ može biti beskonačan
- Broj dihotomija $|\mathcal{H}(x^{(1)}, x^{(2)}, \dots, x^{(N)})|$ je maksimalno 2^N
- Growth function:
 - Najveći broj dihotomija koje se mogu dobiti na bilo kojih N tačaka koristeći dati skup hipoteza
 - kandidat da zameni M u Hoeffdingovoj nejednakosti

Growth function

- Prebrojava najveći broj dihotomija na bilo kojih N tačaka

$$m_{\mathcal{H}}(N) = \max_{x^{(1)}, x^{(2)}, \dots, x^{(N)} \in X} |\mathcal{H}(x^{(1)}, x^{(2)}, \dots, x^{(N)})|$$

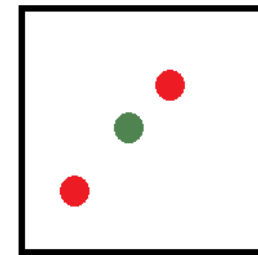
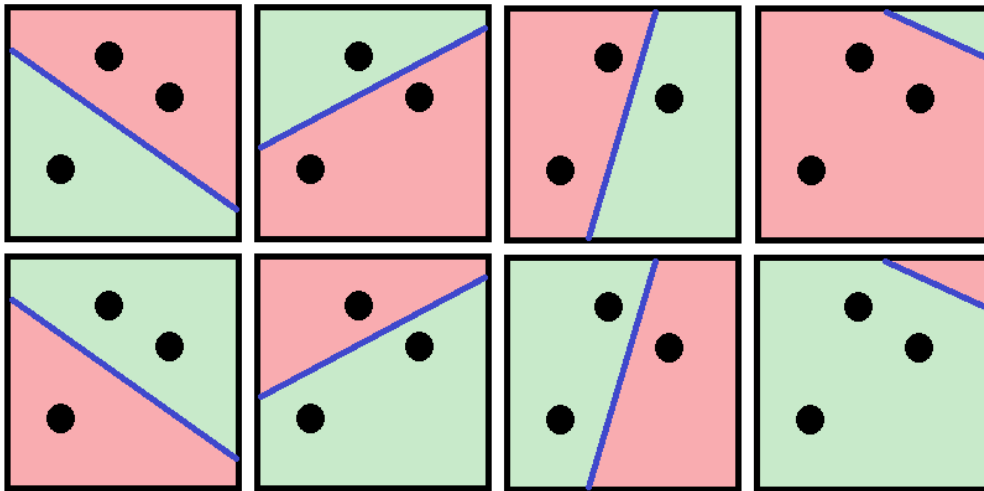
- Za growth funkciju važi:

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

Koliko je model prilagodljiv?

- Razmatraćemo slučaj binarne klasifikacije
- Naš model je perceptron
- Koliko je $m_{\mathcal{H}}(N)$ u ovom slučaju?
 - $m_{\mathcal{H}}(N)$ je funkcija od N – treba da nađemo celu funkciju, dakle vrednost za $N = 1, N = 2, N = 3, \dots$

$$m_{\mathcal{H}}(3) = 8$$



Šta ako su tačke kolinearne?

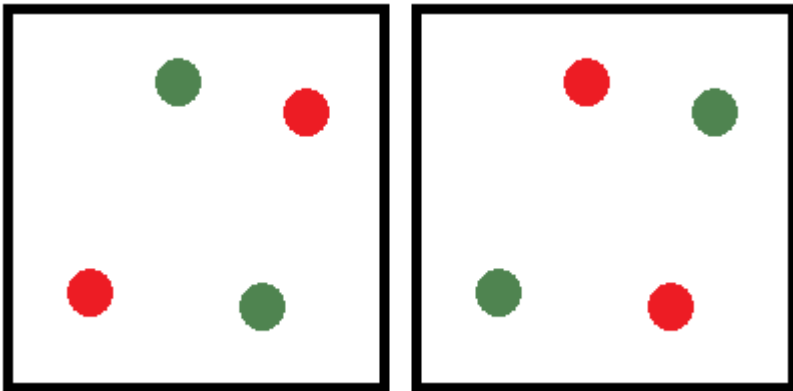
Koliko je model prilagodljiv?

- Razmatraćemo slučaj binarne klasifikacije
- Naš model je perceptron
- Koliko je $m_{\mathcal{H}}(N)$ u ovom slučaju?
 - $m_{\mathcal{H}}(N)$ je funkcija od N – treba da nađemo celu funkciju, dakle vrednost za $N = 1, N = 2, N = 3, \dots$

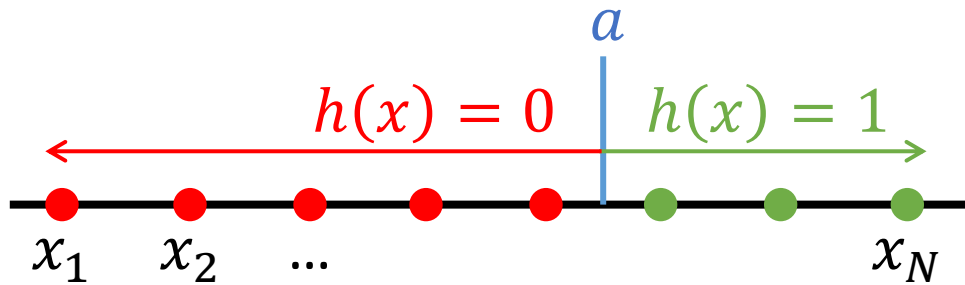
Koliko je model prilagodljiv?

- Razmatraćemo slučaj binarne klasifikacije
- Naš model je perceptron
- Koliko je $m_{\mathcal{H}}(N)$ u ovom slučaju?
 - $m_{\mathcal{H}}(N)$ je funkcija od N – treba da nađemo celu funkciju, dakle vrednost za $N = 1, N = 2, N = 3, \dots$

$$m_{\mathcal{H}}(4) = 14$$



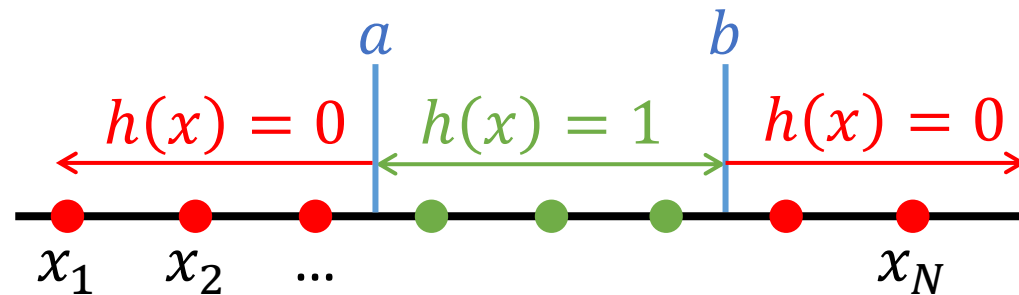
Growth function – ilustrativni primeri



\mathcal{H} je skup hipoteza $h: \mathbb{R} \rightarrow \{0, 1\}$

$$h(x) = \text{sign}(x - a)$$

$$m_{\mathcal{H}}(N) = N + 1$$



\mathcal{H} je skup hipoteza $h: \mathbb{R} \rightarrow \{0, 1\}$

Postaviti krajeve intervala na dva od $N + 1$ mesta

$$m_{\mathcal{H}}(N) = \binom{N + 1}{2} + 1$$

$$= \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

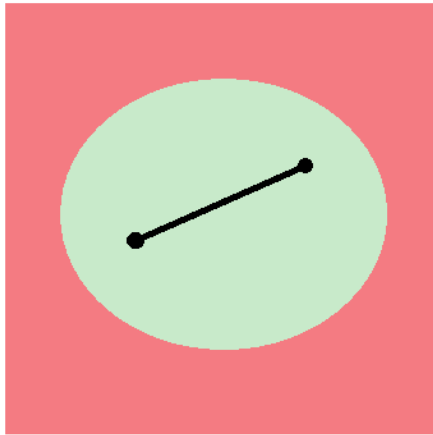
Growth function – ilustrativni primeri

Konveksni skupovi

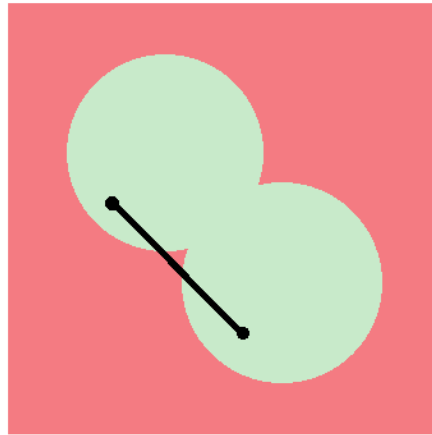
\mathcal{H} je skup hipoteza $h: \mathbb{R}^2 \rightarrow \{0, 1\}$

$h(x) = 1$ je konveksna regija

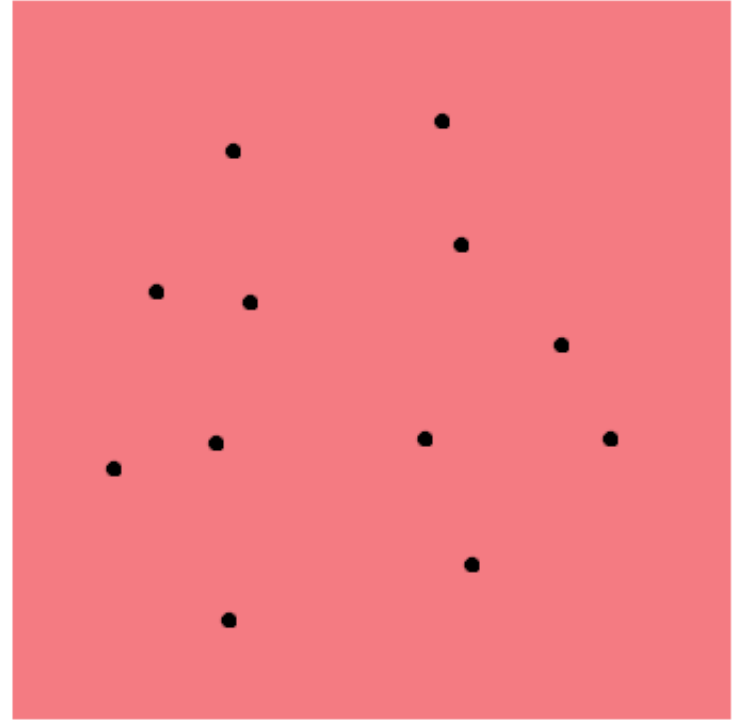
$$m_{\mathcal{H}}(N) = 2^N$$



Konveksna regija
(validna hipoteza)



Nekonveksna regija
(nevalidna hipoteza)



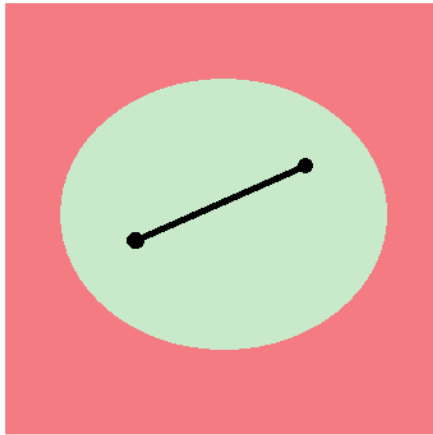
Growth function – ilustrativni primeri

Konveksni skupovi

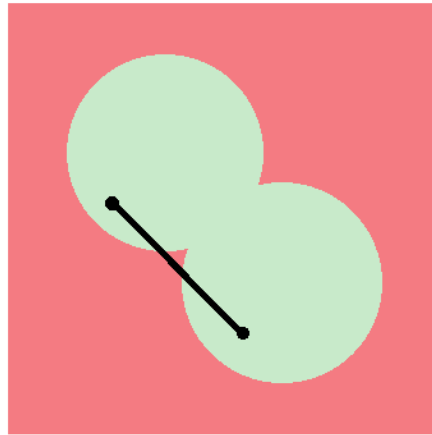
\mathcal{H} je skup hipoteza $h: \mathbb{R}^2 \rightarrow \{0, 1\}$

$h(x) = 1$ je konveksna regija

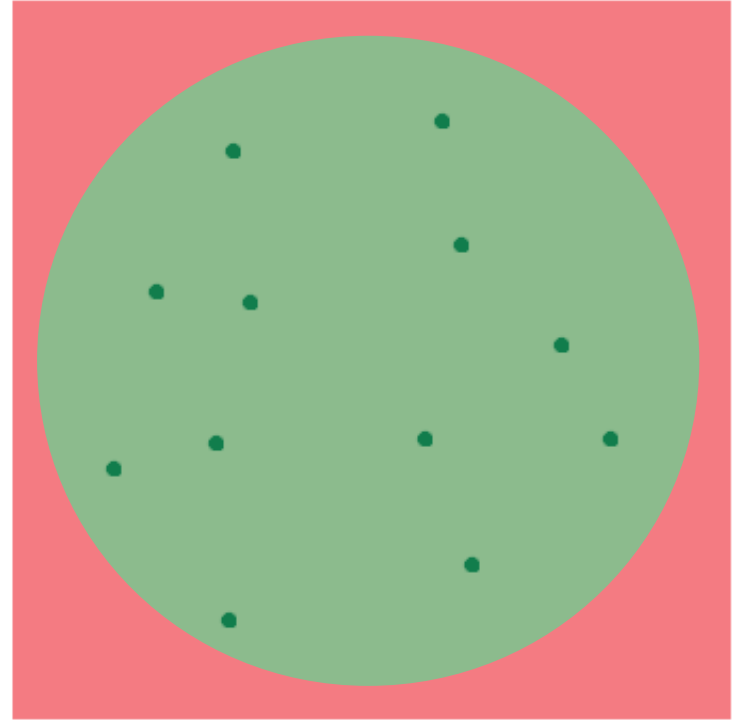
$$m_{\mathcal{H}}(N) = 2^N$$



Konveksna regija
(validna hipoteza)



Nekonveksna regija
(nevalidna hipoteza)



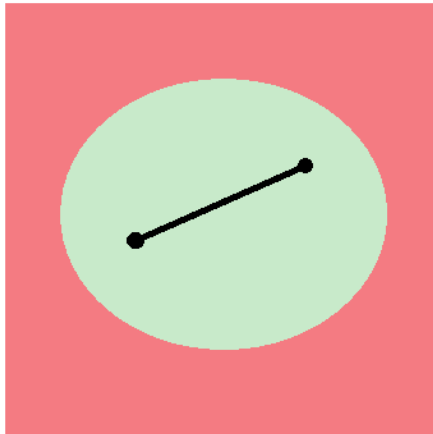
Growth function – ilustrativni primeri

Konveksni skupovi

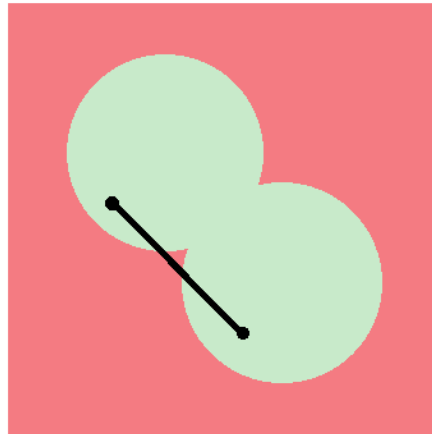
\mathcal{H} je skup hipoteza $h: \mathbb{R}^2 \rightarrow \{0, 1\}$

$h(x) = 1$ je konveksna regija

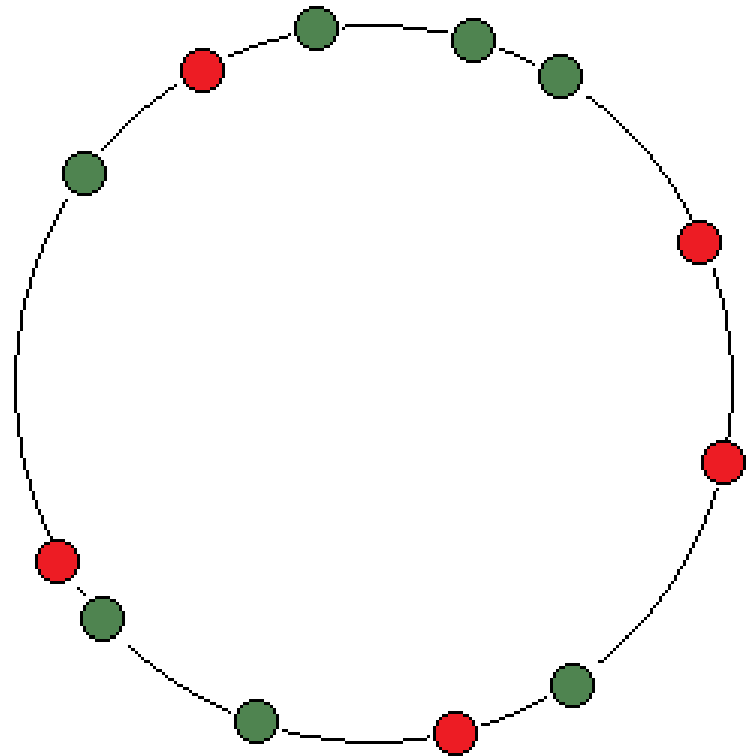
$$m_{\mathcal{H}}(N) = 2^N$$



Konveksna regija
(validna hipoteza)



Nekonveksna regija
(nevalidna hipoteza)



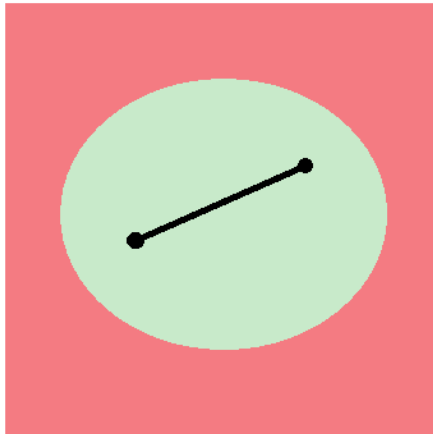
Growth function – ilustrativni primeri

Konveksni skupovi

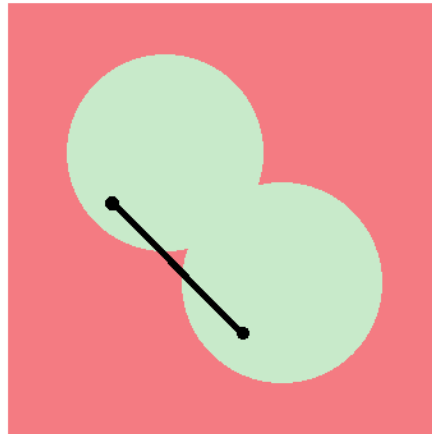
\mathcal{H} je skup hipoteza $h: \mathbb{R}^2 \rightarrow \{0, 1\}$

$h(x) = 1$ je konveksna regija

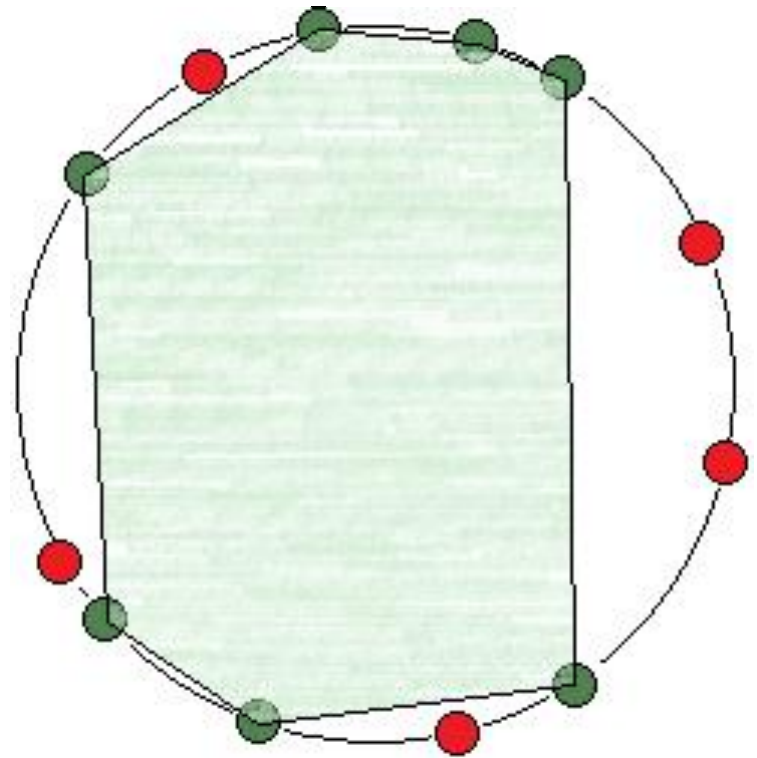
$$m_{\mathcal{H}}(N) = 2^N$$



Konveksna regija
(validna hipoteza)



Nekonveksna regija
(nevalidna hipoteza)



Growth function – ilustrativni primeri

- Tačka:

$$m_{\mathcal{H}}(N) = N + 1$$

- Interval:

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

- Konveksni skupovi:

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

- Što je sofisticiraniji model to je $m_{\mathcal{H}}(N)$ veće

Growth function u Hoeffdingovoj nejednakosti

$$P[|E_{in} - E_{out}| > \epsilon] \leq 2M e^{-2\epsilon^2 N}$$

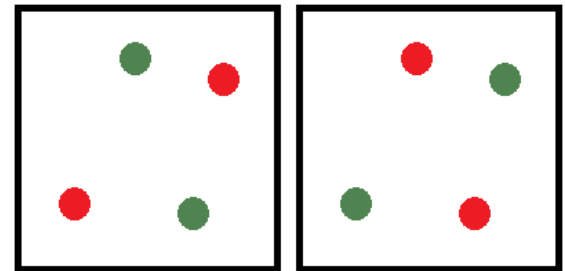
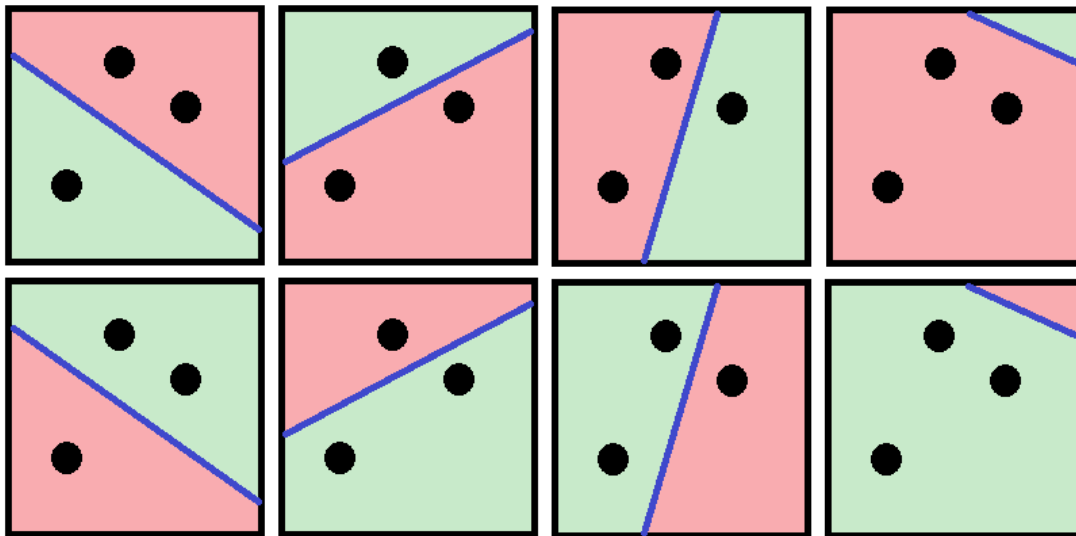
- Šta dobijamo ako M zamenimo sa $m_{\mathcal{H}}(N)$?
 - M može biti beskonačno, $m_{\mathcal{H}}(N)$ je konačan broj
 - Ako je $m_{\mathcal{H}}(N)$ polinomijalno, situacija je dobra – za dovoljno veliko N broj sa desne strane će postati smislen!
 - Treba da pronađemo način da možemo tvrditi da je $m_{\mathcal{H}}(N)$ polinomijalno

Tačka preloma (*break point*)



- Ako na N tačaka sa datim skupom hipoteza možemo dobiti sve dihotomije, kažemo da skup hipoteza „razbija“ N tačaka
 - Npr. konveksni skupovi razbijaju N tačaka jer možemo dobiti svih 2^N hipoteza
- Ako ne postoji skup podataka veličine k koji može da bude „razbijen“ skupom hipoteza \mathcal{H} , onda kažemo da je k **tačka preloma** za \mathcal{H}

$$m_{\mathcal{H}}(k) < 2^k$$

- Npr. za perceptron tačka preloma je 4



Tačka preloma – ilustrativni primeri

- Za koje k više ne možemo dobiti 2^k kombinacija?
- Tačka: $m_{\mathcal{H}}(N) = N + 1$
 - $k = 2$ 
- Interval: $m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$
 - $k = 3$ 
- Konveksni skupovi: $m_{\mathcal{H}}(N) = 2^N$
 - $k = \infty$
- Tačka preloma (ponovo broj) ima željenu osobinu: raste sa porastom sofisticiranosti modela

Zaključak

- Kada kažemo da je skup hipoteza \mathcal{H} složen/prilagodljiv/bogat?
- Ako može da razlikuje različita obeležavanja podataka
- Koliko različitih obeležavanja?
 - Sva
- Na koliko tačaka?
 - Što je veći broj tačaka, veća je prilagodljivost

Nema tačke preloma $\Rightarrow m_{\mathcal{H}}(N) = 2^N$

Postoji tačka preloma $\Rightarrow m_{\mathcal{H}}(N)$ je **polinomijalno** u N

Sumarizacija

- Krenuli smo od broja hipoteza M . To je bilo beznačajno jer M , čak i za proste modele, može biti beskončno i *Hoeffding-ova* nejednakost nam nije ništa značila
- Zatim smo se ograničili na konkretan uzorak, definisali *growth function* $m_{\mathcal{H}}(N)$ i pokušali da ga odredimo za neke modele. Ovo je bilo teško i za jednostavne slučajeve
- Pa smo zaključili da možda ne moramo da znamo konkretno $m_{\mathcal{H}}(N)$, možda je samo dovoljno da možemo tvrditi da je polinomijalno
- Pa smo definisali tačku preloma (možda je jednostavnije da samo izračunamo tačku preloma). Ovo je jednostavnije (treba samo da nađemo jedan pametan primer gde ne možemo razbiti skup tačaka), ali opet nije dovoljno jednostavno
- Na kraju smo došli do toga da ne moramo da znamo ni tačnu tačku preloma – dovoljno je samo da znamo da ona postoji!
- Jer nije važno koji polinom dobijete – važno je samo da možete da naučite ako vam neko da dovoljno podataka (veliko N)
- Ako želimo da znamo tačno koliko nam primera treba za učenje da bismo dobili neke određene performanse – onda moramo da nađemo tačku preloma
- Ali, u principu, ako samo želim da kažem da mogu da učim sa ovim konkretnim skupom hipoteza dovoljno je da kažem da imam tačku preloma