Teorijske osnove nadgledanog učenja

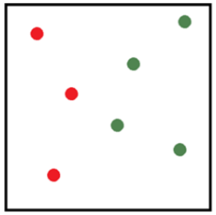
Šta znamo do sada?

- Hoeffdingova nejednakost:

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

M – broj mogućih hipoteza (od koje biramo jednu)

- Dihotomije



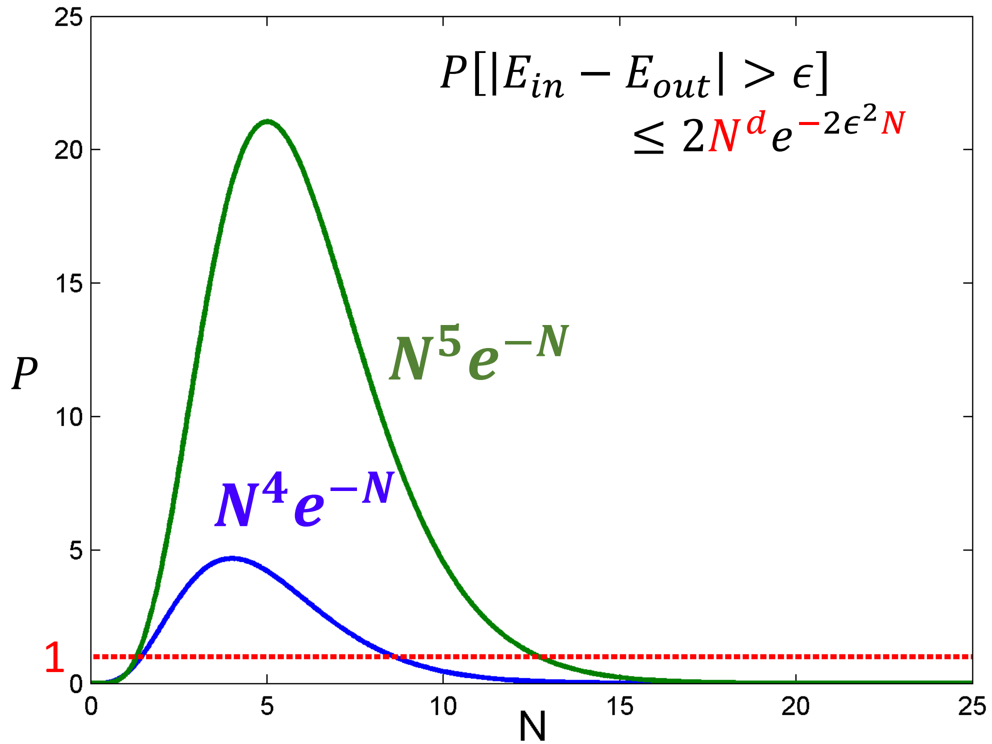
- Growth function

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

- Želimo da u Hoeffdingovoj nejednakosti zamenimo M (potencijalno beskonačno) sa $m_{\mathcal{H}}(N)$ (konačan broj)

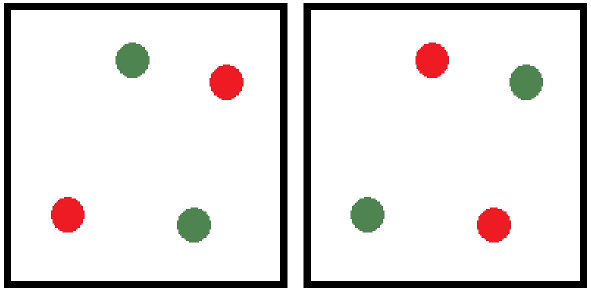
Šta znamo do sada?

- Problem: teško je odrediti $m_{\mathcal{H}}(N)$ - zavisi od skupa hipoteza i ulaznog prostora
- Ali, dovoljno je da znamo da je $m_{\mathcal{H}}(N)$ polinomijalno – za dovoljno veliko *N* granica verovatnoće će postati smislena!



Šta znamo do sada?

- Tačka preloma
  - Ako ne postoji skup podataka veličine k koji može da bude „razbijen“ skupom hipoteza $\mathcal{H}$, onda kažemo da je k tačka preloma za $\mathcal{H}$
  - Npr. 4 je tačka preloma za perceptron jer ne možemo dobiti ove kombinacije:



- Želimo da možemo da tvrdimo: ako postoji tačka preloma $\Rightarrow m_{\mathcal{H}}(N)$ je polinomijalno u N

Teorija generalizacije

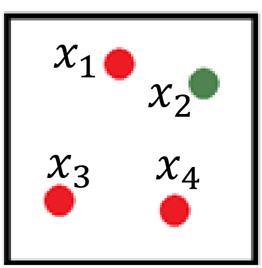
1. Dokazaćemo da je $m_{\mathcal{H}}(N)$ polinomijalno ako postoji tačka preloma
2. $m_{\mathcal{H}}(N)$ može da zameni M u Hoeffdingovoj nejednakosti

1. Gornja granica $m_{\mathcal{H}}(N)$

- Da bismo pokazali da je $m_{\mathcal{H}}(N)$ polinomijalno
- Pokazaćemo $m_{\mathcal{H}}(N) \leq \cdots \leq \cdots \leq \text{polinom}$
- $B(N, k)$ - maksimalan broj dihotomija na *N* tačaka sa tačkom preloma *k*
  - Nije vezano za konkretan ulazni prostor ili konkretan skup hipoteza
- Napravićemo sledeću tabelu:

$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
$+1$	$+1$	$\ldots$	$+1$	$+1$
-1	$+1$	$\ldots$	$+1$	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Ispisaćemo tabelu sa svim mogućim *različitim* dihotomijama dobijenim pod ograničenjem tačke preloma *k*



x_1	x_2	x_3	x_4
-1	$+1$	-1	-1

- $B(N, k)$ je broj redova ove tabele
  - Broj šablona koje možemo dobiti na N tačaka, tako da ne postoji *k* kolona koje sadrže sve moguće kombinacije
- Odvajamo poslednju kolonu jer želimo rekurzivnu definiciju $B(N, k)$

Određivanje $B(N,k)$

- Napravićemo određeno grupisanje redova ove tabele:

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
			$+1$	$+1$	$\dots$	$+1$	$+1$
			-1	$+1$	$\dots$	$+1$	-1
	S_1	α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	-1	-1
			-1	$+1$	$\dots$	-1	$+1$
			$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	$+1$	$+1$
	S_2^+	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	$+1$
S_2			-1	-1	$\dots$	-1	$+1$
			$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	$+1$	-1
	S_2^-	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	-1	-1

Razlika ovih redova je samo u x_N

- Grupa S_1: kombinacije $x_1, \dots, x_{N-1}$ koje se pojavljuju isključivo sa jednom eksenzijom x_N (ili +1 ili -1, ali, zbog tačke preloma, nemamo obe)
- Grupa S_2^+: kombinacije $x_1, \dots, x_{N-1}$ za koje je $x_N = +1$
- Grupa S_2^-: iste kombinacije $x_1, \dots, x_{N-1}$ kao u S_2^+, samo za koje je $x_N = -1$

$$B(N,k) = \alpha + 2\beta$$

Procena α i β

- Fokusiraćemo se na kolone $x_1, \ldots, x_{N-1}$
- Istaknuti su samo redovi u kojima se kombinacije $x_1, \ldots, x_{N-1}$ razlikuju
  - Za grupu S_1 znamo da svi redovi razlikuju (tu smo stavljali kombinacije za koje imamo samo jedan nastavak x_N)
  - Ako bi se u okviru S_1 ponovila kombinacija $x_1, \ldots, x_{N-1}$, to bi značilo da se ta dva reda moraju razlikovati po nastavku x_N (jedan nastavak +1, a drugi nastavak -1), ali tada bi ovi redovi pripadali u S_2^+ i S_2^-
  - Grupe S_2^+ i S_2^- smo konstruisali tako da imaju identične redove $x_1, \ldots, x_{N-1}$ (a razlikuju se po vrednosti x_N)

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
	S_1	α	$+1$	$+1$	$\ldots$	$+1$	$+1$
			-1	$+1$	$\ldots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	-1	-1
			-1	$+1$	$\ldots$	-1	$+1$
S_2	S_2^+	β	$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	$+1$	$+1$
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	-1	$+1$
	S_2^-	β	$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	-1	-1

Procena α i β

- Da li možemo reći nešto o tački preloma ove manje matrice?
  - U originalnoj matrici nije postojalo *k* kolona sa svim mogućim kombinacijama – u suprotnom, *k* ne bi bila tačka preloma
  - To važi i za ovu podmatricu (ako bi takvih *k* kolona postojalo, istih *k* kolona bi sadržale sve kombinacije i u originalnoj matrici)
  - Broj redova originalne matrice je $B(N, k)$
  - U podmatrici možemo imati maksimalno $B(N-1, k)$ redova
- Na osnovu ovoga možemo zaključiti:

$$\alpha + \beta \leq B(N-1, k)$$

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1		α	$+1$	$+1$	$\dots$	$+1$	$+1$
			-1	$+1$	$\dots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	-1	-1
			-1	$+1$	$\dots$	-1	$+1$
S_2	S_2^+	β	$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	$+1$	$+1$
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	$+1$
			-1	-1	$\dots$	-1	$+1$
	S_2^-	β	$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\dots$	$+1$	-1
			-1	-1	$\dots$	-1	-1

Procena β

- Fokusiraćemo se samo na grupe S_2^+ i S_2^-
  - Redove iz S_2^- smo sakrili jer se radi o identičnim redovima kao u S_2^+
- Ovi redovi imaju tačku preloma $k-1$
  - Recimo da imamo svih $k-1$ kombinacija
  - U tom slučaju bismo dodavanjem kolone x_N imali sve moguće kombinacije *k* kolona – što je u suprotnosti sa zadatom tačkom preloma *k* za *N* kolona
- Na osnovu ovoga možemo zaključiti:

$$\beta \leq B(N-1, k-1)$$

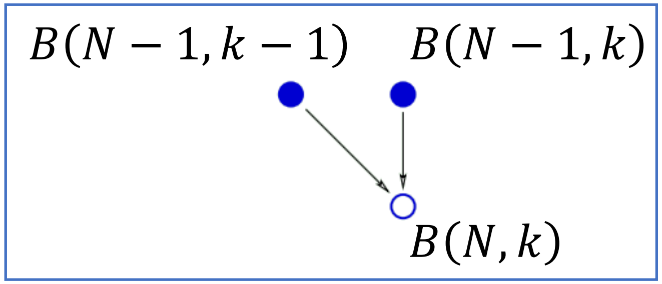
		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
			$+1$	$+1$	$\ldots$	$+1$	$+1$
			-1	$+1$	$\ldots$	$+1$	-1
S_1		α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	-1	-1
			-1	$+1$	$\ldots$	-1	$+1$
			$+1$	-1	$\ldots$	$+1$	$+1$
			-1	-1	$\ldots$	$+1$	$+1$
	S_2^+	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	$+1$
S_2			-1	-1	$\ldots$	-1	$+1$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	$+1$	-1
	S_2^-	β	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	$\ldots$	$+1$	-1
			-1	-1	$\ldots$	-1	-1

Određivanje $B(N,k)$

$B(N,k) = \alpha + 2\beta \,, \alpha + \beta \leq B(N-1,k), \beta \leq B(N-1,k-1)$

$\Rightarrow B(N,k) \leq B(N-1,k) + B(N-1,k-1)$

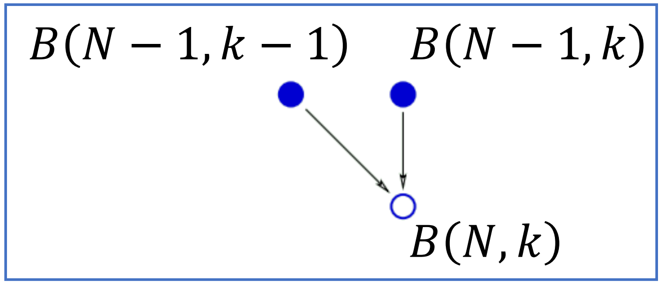
Ovo će biti gornja granica za growth function **bilo kog skupa hipoteza** sa prelomnom tačkom k **na bilo kom ulaznom prostoru**



		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1	3	4	4	4	4	..
	3	1	4	7	8	8	8	..
N	4	1	5	11	..	..	..	..
	5	1	6	⋮	.			
	6	1	7	⋮		.		
	⋮	⋮	⋮	⋮			.	

Određivanje $B(N,k)$

$B(N,k) = \alpha + 2\beta\ , \alpha + \beta \le B(N-1,k), \beta \le B(N-1,k-1)$



$\Rightarrow B(N,k) \le B(N-1,k) + B(N-1,k-1)$

Ovo će biti gornja granica za growth function **bilo kog skupa hipoteza** sa prelomnom tačkom *k* **na bilo kom ulaznom prostoru**

- Na jednoj tački imamo 2 moguće vrednosti +1 i -1
- Ako je tačka preloma 1 ne možemo imati više od jednog reda u tabeli $x_1, \ldots, x_N$ (u tabelu stavljamo samo različite šablone koji se međusobno razlikuju u barem jednoj koloni)

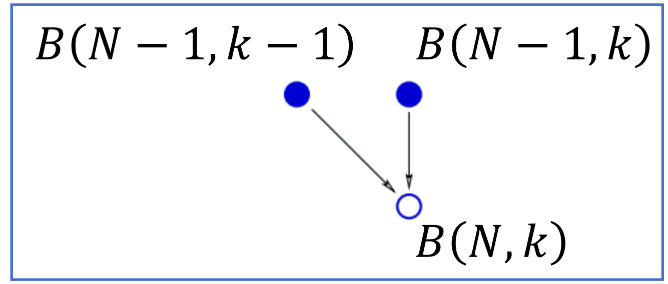
N \ k	1	2	3	4	5	6	..
1	1	2	2	2	2	2	..
2	1	3	4	4	4	4	..
3	1	4	7	8	8	8	..
4	1	5	11	..	..	..	..
5	1	6	:	.			
6	1	7	:		.		
:	:	:	:			.	

Određivanje $B(N,k)$

$B(N,k) = \alpha + 2\beta\,, \alpha + \beta \leq B(N-1,k), \beta \leq B(N-1,k-1)$

$\Rightarrow B(N,k) \leq B(N-1,k) + B(N-1,k-1)$

Ovo će biti gornja granica za growth function **bilo kog skupa hipoteza** sa prelomnom tačkom *k* **na bilo kom ulaznom prostoru**



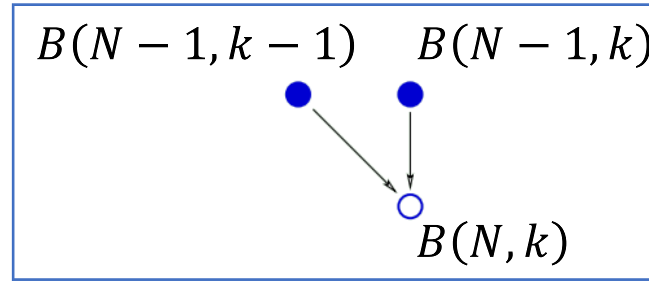
N \ k	1	2	3	4	5	6	..
1	1	2	2	2	2	2	..
2	1	3	4	4	4	4	..
3	1	4	7	8	8	8	..
4	1	5	11	..	..	..	..
5	1	6	:	.			
6	1	7	:		.		
:	:	:	:			.	

Određivanje $B(N, k)$

$B(N,k) = \alpha + 2\beta \,, \alpha + \beta \leq B(N-1,k), \beta \leq B(N-1,k-1)$

$\Rightarrow B(N,k) \leq B(N-1,k) + B(N-1,k-1)$

Ovo će biti gornja granica za growth function **bilo kog skupa hipoteza** sa prelomnom tačkom *k* **na bilo kom ulaznom prostoru**



- Koliko različitih vrednosti možemo imati na jednoj tački pod ograničenjem da nijednih $k > 1$ tačaka ne mogu da imaju sve moguće kombinacije
  - Ovo ograničenje zapravo nije nikakvo ograničenje
  - Dakle, možemo imati 2 različita mapiranja za 1 tačku (+1 i -1)

	$N \backslash k$	1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1	3	4	4	4	4	..
	3	1	4	7	8	8	8	..
N	4	1	5	11	..	..	..	..
	5	1	6	:	.			
	6	1	7	:		.		
	:	:	:	:			.	

Analitičko rešenje za $B(N,k)$

$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

Teorema:

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Dokaz indukcijom:

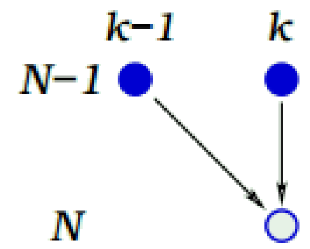
1. Proverimo ispunjenost graničnih uslova (prva kolona i prvi red)
2. Pretpostavićemo da formula važi za $B(N-1,k-1)$ i $B(N-1,k)$ i onda ćemo pokazati da mora da važi i za $B(N,k)$

		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1						
	3	1						
N	4	1						
	5	1						
	6	1						
	⋮	⋮						

Indukcija

$B(N, k)$ $\qquad$ $B(N-1, k)$ $\qquad$ $B(N-1, k-1)$

$$\sum_{i=0}^{k-1}\binom{N}{i} = \sum_{i=0}^{k-1}\binom{N-1}{i} + \sum_{i=0}^{k-2}\binom{N-1}{i} \; ?$$



$$= 1 + \sum_{i=1}^{k-1}\binom{N-1}{i} + \sum_{i=1}^{k-1}\binom{N-1}{i-1}$$

$$= 1 + \sum_{i=1}^{k-1}\left[\binom{N-1}{i} + \binom{N-1}{i-1}\right]$$

$$= 1 + \sum_{i=1}^{k-1}\binom{N}{i} = \sum_{i=0}^{k-1}\binom{N}{i} \; ✓$$

$B(N,k)$ je polinomijalno!

Dokazali smo:

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

- Za konkretan skup hipoteza $\mathcal{H}$, tačka preloma k je fiksirana (ne raste sa *N*)
  - Npr. za perceptron ne možemo moguće dihotomije za 4 tačke, pa je 4 tačka preloma. Ako se zapitamo šta perceptron radi na 100 tačaka – tačka preloma je i dalje 4 (konstanta koja ne zavisi od broja tačaka)
- $B(N,k)$ - maksimum dihotomija koje možemo dobiti pod ograničenjem da je prelomna tačka k
  - pomoću bilo kog skupa hipoteza na bilo kom ulaznom prostoru
- Dakle,

$$m_{\mathcal{H}}(N) \leq B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

$B(N,k)$ je polinomijalno!

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

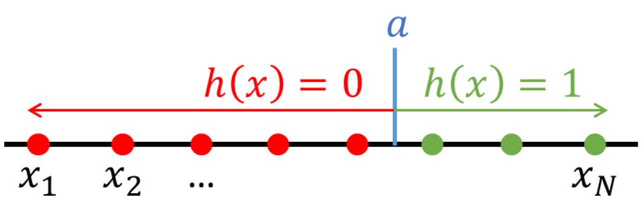
- Da li je $m_{\mathcal{H}}(N)$ polinomijalno po N?
  - Da, u $\sum_{i=0}^{k-1} \binom{N}{i}$ maksimalni stepen je N^{k-1}, a k je konstanta (ne menja se sa N)
  - Ne samo da je $m_{\mathcal{H}}(N)$ polinomijalno po N, već stepen tog polinoma zavisi od tačke preloma

- *Hoeffdingova* nejednakost:

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

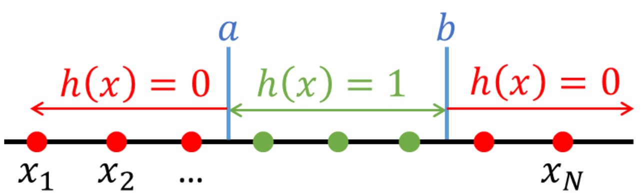
Tri primera (sa prošlog predavanja)

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$



Dobili smo: $k = 2, m_{\mathcal{H}}(x) = N + 1$

Granica: $m_{\mathcal{H}}(x) \leq \sum_{i=0}^{1} \binom{N}{i} = N + 1$



Dobili smo: $k = 3, m_{\mathcal{H}}(x) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

Granica: $m_{\mathcal{H}}(x) \leq \sum_{i=0}^{2} \binom{N}{i} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

Tri primera (sa prošlog predavanja)

- 2D perceptron:
  - Našli smo da je $k = 4$
  - Ali nismo se trudili da pronađemo $m_{\mathcal{H}}(x)$ za $N = 1,2,3, \ldots$ ovo bi bilo veoma naporno uraditi za svako *N*!
  - Sada, *bez poznavanja geometrije prostora i skupa hipoteza*, znamo smo da odredimo granicu:

$$m_{\mathcal{H}}(x) \leq \sum_{i=0}^{3} \binom{N}{i} = \frac{1}{6} N^3 + \frac{5}{6} N + 1$$

  - I, možemo da tvrdimo da je $m_{\mathcal{H}}(x)$ polinomijalno po *N*

$m_{\mathcal{H}}(N)$ može da zameni M

- Umesto

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

- Imaćemo:

Vapnik-Červonenkinsova nejednakost

$$\boxed{P[|E_{in} - E_{out}| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}$$

- Ovo je tačno za bilo koju hipotezu koja ima tačku preloma
  - $m_{\mathcal{H}}(2N)$ je polinomijalno i mi ćemo (sa dovoljno velikim N) moći da učimo – da greška na uzorku korektno prati stvarnu grešku
- VC nejednakost je najvažniji teorijski rezultat u mašinskom učenju

*Dokaz da $m_{\mathcal{H}}$ može da zameni M možete pronaći u knjizi Abu-Mostafa, Y.S., Magdon-Ismail, M. and Lin, H.T., 2012. *Learning from data* (Vol. 4). New York, NY, USA:: AMLBook. (Appendix)

Spajanje svega otkrivenog

- Dakle, sve ovo je vredelo truda jer imamo jednostavnu karakterizaciju skupa hipoteza:

  Sve što treba da uradimo da bismo tvrdili da (uz dovoljno veliko N) možemo da učimo koristeći taj skup hipoteza jeste da dokažemo da tačka preloma postoji za taj skup hipoteza

  - a ne moramo čak ni da znamo koja tačka preloma je u pitanju

VC dimenzija

Definicija, interpretacija i granice generalizacije

Definicija VC dimenzije

- VC dimenzija je broj koji se definiše za skup hipoteza $\mathcal{H}$. Označićemo je sa $d_{VC}(\mathcal{H})$
- Predstavlja „najviše tačaka koje $\mathcal{H}$ može da razbije“
  - Najveća vrednost N za koju važi $m_h(N) = 2^N$
  - Ne kaže sa *bilo kojih* N tačaka može biti razbijeno (dovoljno je da možemo pronaći pogodnih N tačaka)

$$N \leq d_{VC}(\mathcal{H}) \Rightarrow \mathcal{H} \text{ može da razbije N tačaka}$$
$$N > d_{VC}(\mathcal{H}) \Rightarrow N \text{ je tačka preloma za } \mathcal{H}$$

The growth function

- U pogledu tačke preloma k:

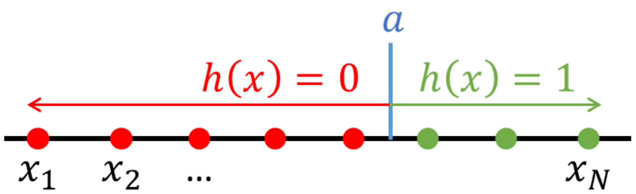
$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

- U pogledu VC dimenzije d_{VC}:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

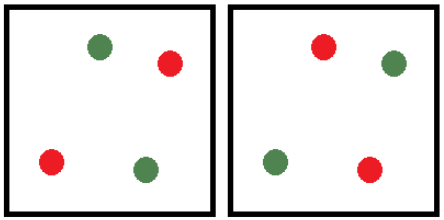
  - Najveći stepen u ovom polinomu je $N^{d_{VC}}$

Primeri



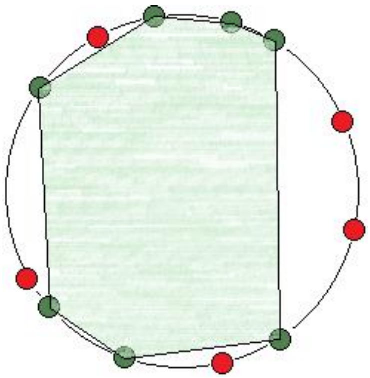
Pozitivni zraci:

- Možemo „razbiti“ najviše jednu tačku (tačka preloma je 2). Dakle, $d_{VC} = 1$



2D perceptroni:

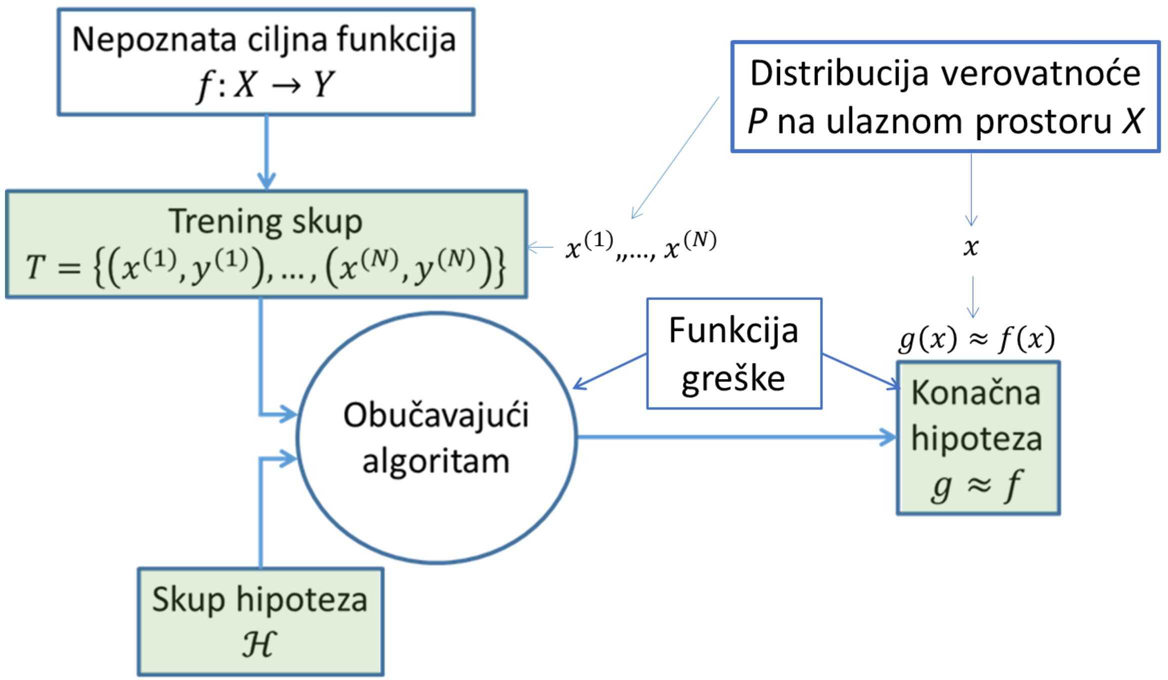
- Tačka preloma je 4, dakle $d_{VC} = 3$



Konveksne regije:

- Možemo da razbijemo bilo kojih *N* tačaka ako su raspoređene po kružnicu
- Dakle, $d_{VC} = \infty$ (ne postoji maksimum tačaka koji možemo razbiti)

Povezanost VC dimenzije sa učenjem



- Ako je d_{VC} konačno $\Rightarrow$ g $\in \mathcal{H}$ će da generalizuje
  - Nezavisno od obučavajućeg algoritma
  - Nezavisno od distribucije verovatnoće na ulaznom prostoru X
  - Nezavisno od ciljne funkcije

VC dimenzija perceptrona

- Perceptron ($x \in \mathbb{R}^D$):

$$h(x) = sign\left(\sum_{i=0}^{D} \theta_i x_i\right) = sign(\theta^T x)$$

- Za $D = 2$ (dvodimenzioni prostor):

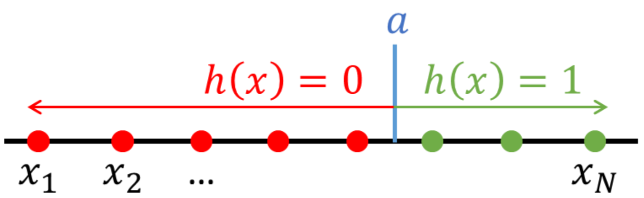
$$d_{VC} = 3$$

  - Ali, ako pređemo u 3D prostor, možemo razbiti slučaj od 4 tačke koje nismo mogli razbiti u 2D prostoru – samo treba da odaberemo tačke koje nisu u ravni
- Generalnije,

$$d_{VC} = D + 1$$

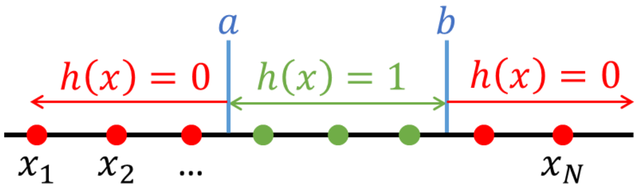
- $D + 1$ je broj parametara perceptrona $\theta = [\theta_0, \theta_1, \dots, \theta_D]$!
  - VC dimenzija nam daje maksimalan broj tačaka koji možemo da razbijemo
  - Sa više parametara imamo višu VC dimenziju

Primeri



$d_{VC} = 1$

Broj parametara: 1 (a)



$d_{VC} = 2$

Broj parametara: 2 (a, b)

Interpretacija VC dimenzije – stepeni slobode

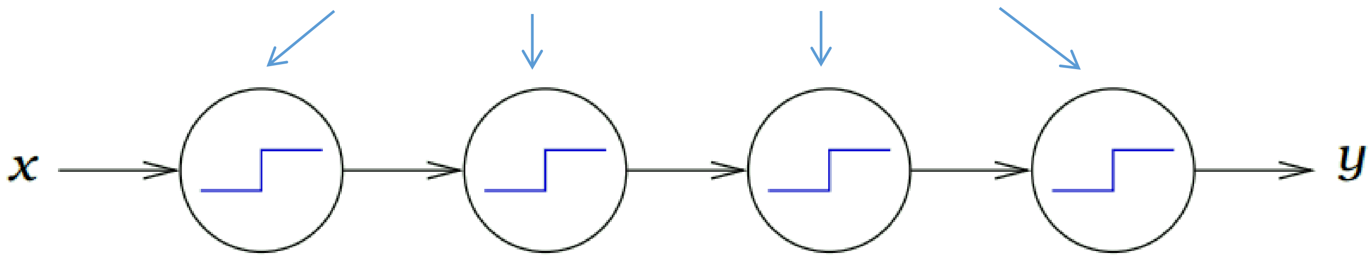
- Broj parametara: analogni stepeni slobode
  - Svaki parametar je realan broj

- d_{VC} je ekvivalentno „binarnim“ stepenima slobode
  - Različite dihotomije

Ali...

- Parametri ne moraju nužno da doprinose stepenima slobode

Perceptron $h(x) = sign(\theta^T x)$ (1D slučaj sa 2 parametra)



- Imamo ukupno 8 parametara u ovom modelu. Da li to znači da imamo 8 stepeni slobode?
  - Ne, jer su ovi modeli veoma redudantni – nakon prvog, samo vršimo mapiranja između klasa
- Sa druge strane, VC dimenzija tretira ovaj model kao *black box* i samo gleda šta on može da postigne (koliko maksimalno tačaka može biti „razbijeno“)
  - Zato je VC dimenzija mnogo realnija mera stepena slobode sa kojima raspolažete
  - Možemo reći da se VC dimenzija meri *efektivan* broj parametara

Broj instanci koje nam trebaju

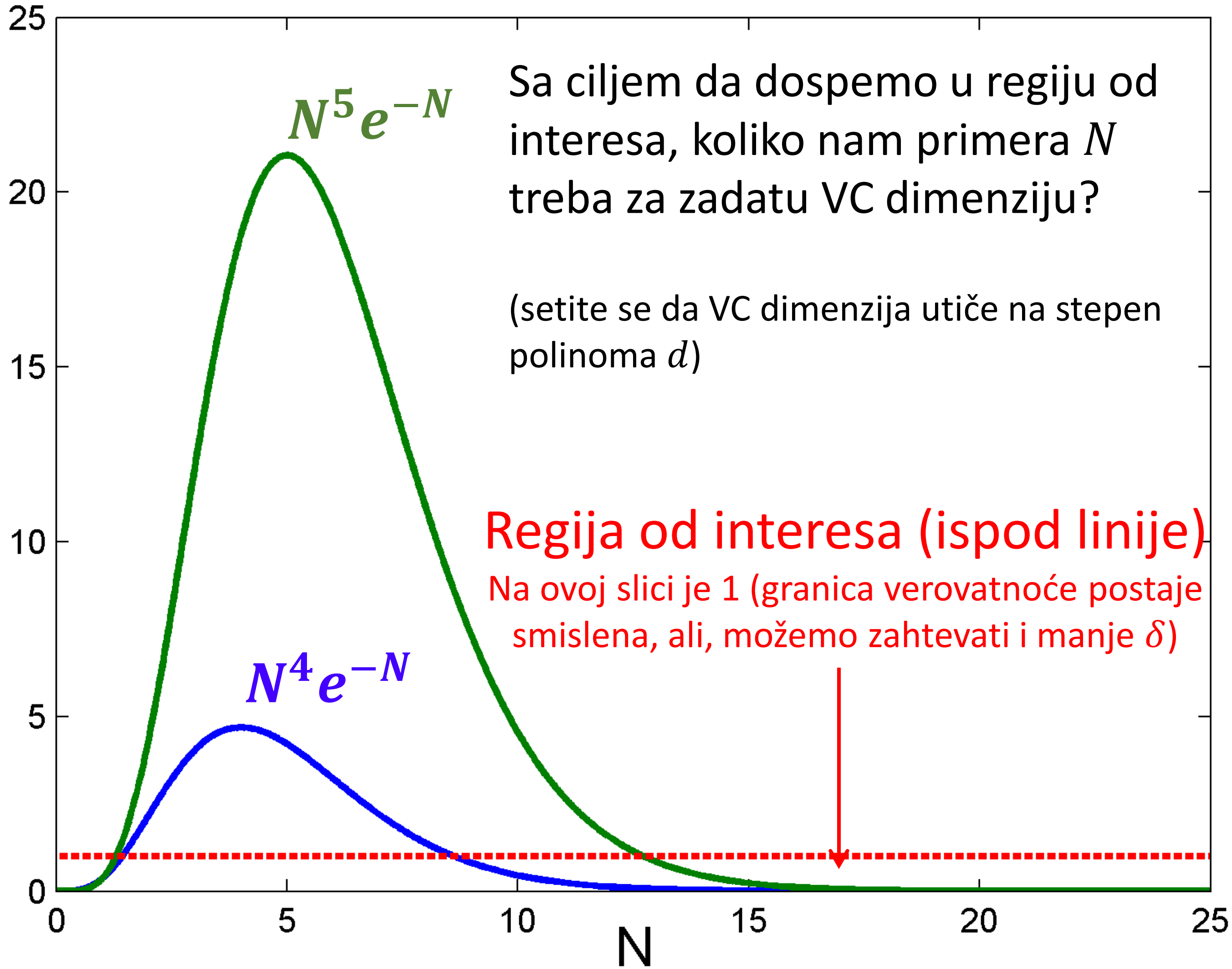
- Dve veličine u VC nejednakosti za koje želimo da budu male:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \underbrace{\boxed{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}}_{\delta}$$

- Ako želimo određene ϵ i δ koliko nam primera treba?
  - Npr., koliko nam primera treba ako želimo da smo za najviše 10% pogrešno procenili E_{out} ($\epsilon = 0.1$) i želimo da je ova izjava tačna u 95% slučajeva ($\delta = 0.05$)
  - Koliko nam primera treba: kako N zavisi od d_{VC}?

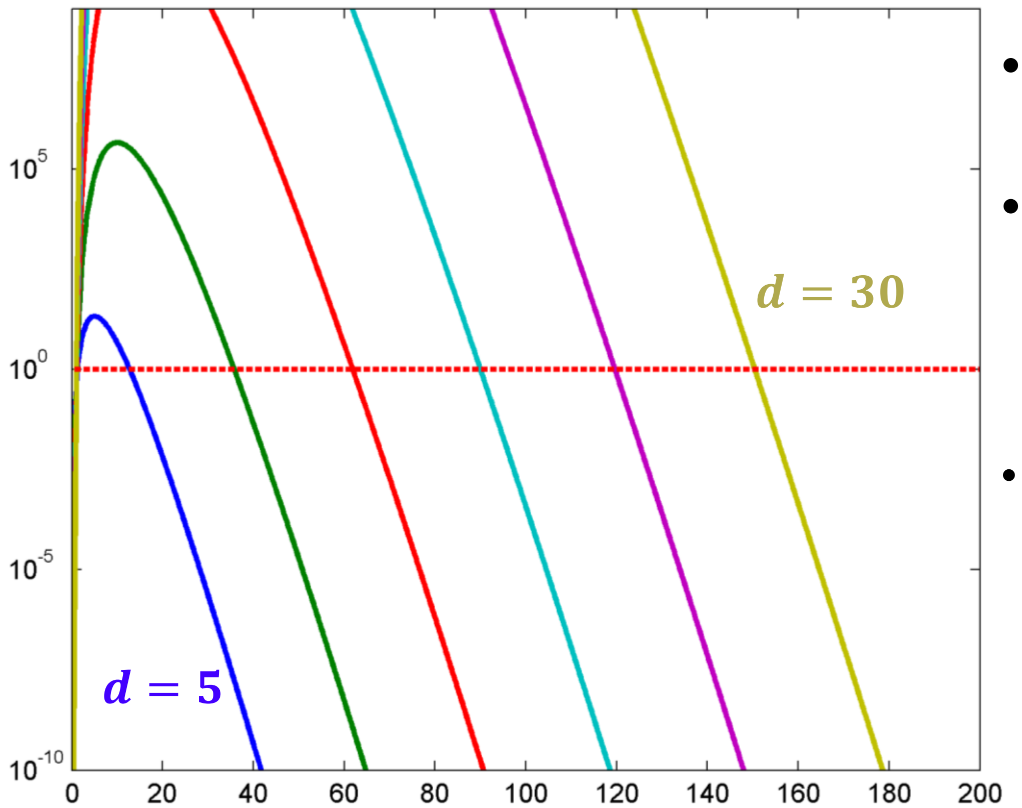
- Razmotrićemo ovu funkciju: $N^d e^{-N}$ (δ sa izbačenim konstantama – zanima nas nagodba između d i N)

$N^d e^{-N}$ (verovatnoća, želimo da bude mala)



$N^d e^{-N}$(verovatnoća, želimo da bude mala)

- Da bismo mogli da prikažemo grafik za različite stepene d (a da ostane vidljivo), prikazaćemo isti grafik na logaritamskoj skali:



- Za veću VC dimenziju, treba nam više primera
- Broj primera koji nam treba za željene garancije je gotovo proporcionalan VC dimenziji

- Napomena: ono što je prikazano na grafiku su *gornje granice* veličina koje merimo i nema garancije da će se prave veličine ponašati na isti način. Međutim, praksa nam pokazuje da to jeste slučaj

Koliko nam N treba za datu VC dimenziju?

- Rezultat iz prakse: da dobijemo *razumnu* generalizaciju:

$$N \geq 10 d_{VC}$$

- Ovo može da zavisi u odnosu na konkretno ϵ i δ i konkretnu primenu, ali se u praksi pokazalo da često važi za širok spektar razumno odabranih ϵ i δ

Granice generalizacije

- Polazeći od VC nejednakosti:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \overbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}^{\delta}$$

- Ako mi specificirate toleranciju ϵ, ja mogu da izračunam verovatnoću δ u odnosu na dati broj primera
- Možemo i obrnuto – da računamo ϵ na osnovu datog δ
  - Želimo da damo izjavu sa pouzdanošću od 95%
  - Možemo da kažemo koju toleranciju garantujemo pod 95%
  - $\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \Rightarrow \epsilon = \boxed{\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}}$ Ω

Granice generalizacije (*generalization bound*)

- Sa verovatnoćom $\geq 1 - \delta$,
$$|E_{out} - E_{in}| \leq \Omega(N, \mathcal{H}, \delta)$$
  - Ω je funkcija:
    - Broja primera (opada sa porastom N)
    - Skupa hipoteza (raste sa VC dimenzijom)
    - Verovatnoćom naše izjave (raste sa smanjenjem δ)

- Ovo ćemo još malo pojednostaviti. Uklonićemo apsolutnu vrednost
  - Ovo je smer koji nam je važan
  - U praksi, E_{in} će uglavnom biti manje od E_{out} jer je to veličina koju ciljano minimizujemo

Generalization bound

- Generalization bound:

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

- E_{out} ne znamo. Ali, E_{in} i Ω su nam poznati i imamo donekle kontrolu nad njima:
  - E_{in} je ono što minimizujemo
  - Ω zavisi od odabranog skupa hipoteza

- Da li je veći skup hipoteza dobar ili loš?
  - Dobar je za E_{in}
  - Loš je za generalizaciju
  - Zato nije dobra ideja prosto odabrati veći skup hipoteza – postoji ravnoteža koja će napraviti da suma $\mathrm{E}_{in} + \Omega$ bude najmanja