Strojno učenje

2. Nadzirano učenje

prof. dr. sc. Bojana Dalbelo Bašić
doc. dr. sc. Jan Šnajder

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Ak. god. 2012/13.

v1.1

Danas. . .

1. Osnovni pojmovi
2. Vapnik-Chervonenkisova dimenzija
3. Induktivna pristranost
4. Primjer: regresija
5. Odabir modela

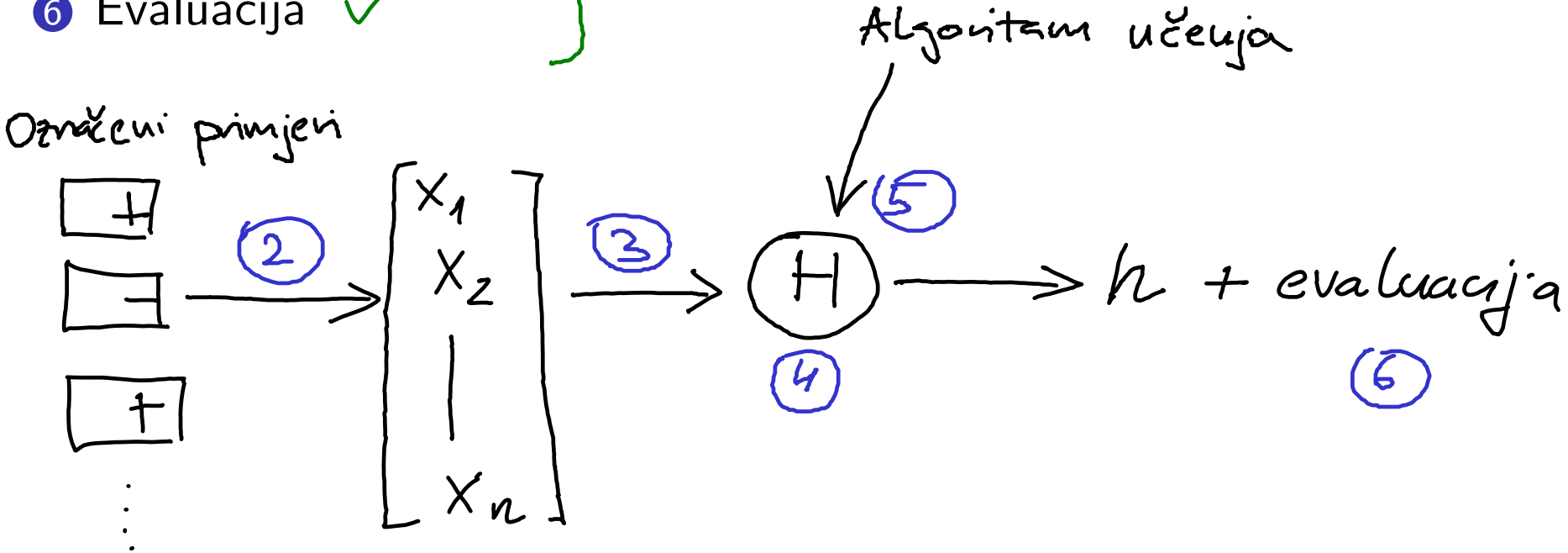
Danas...

1. Osnovni pojmovi
2. Vapnik-Chervonenkisova dimenzija
3. Induktivna pristranost
4. Primjer: regresija
5. Odabir modela

Strojno učenje – tipični koraci

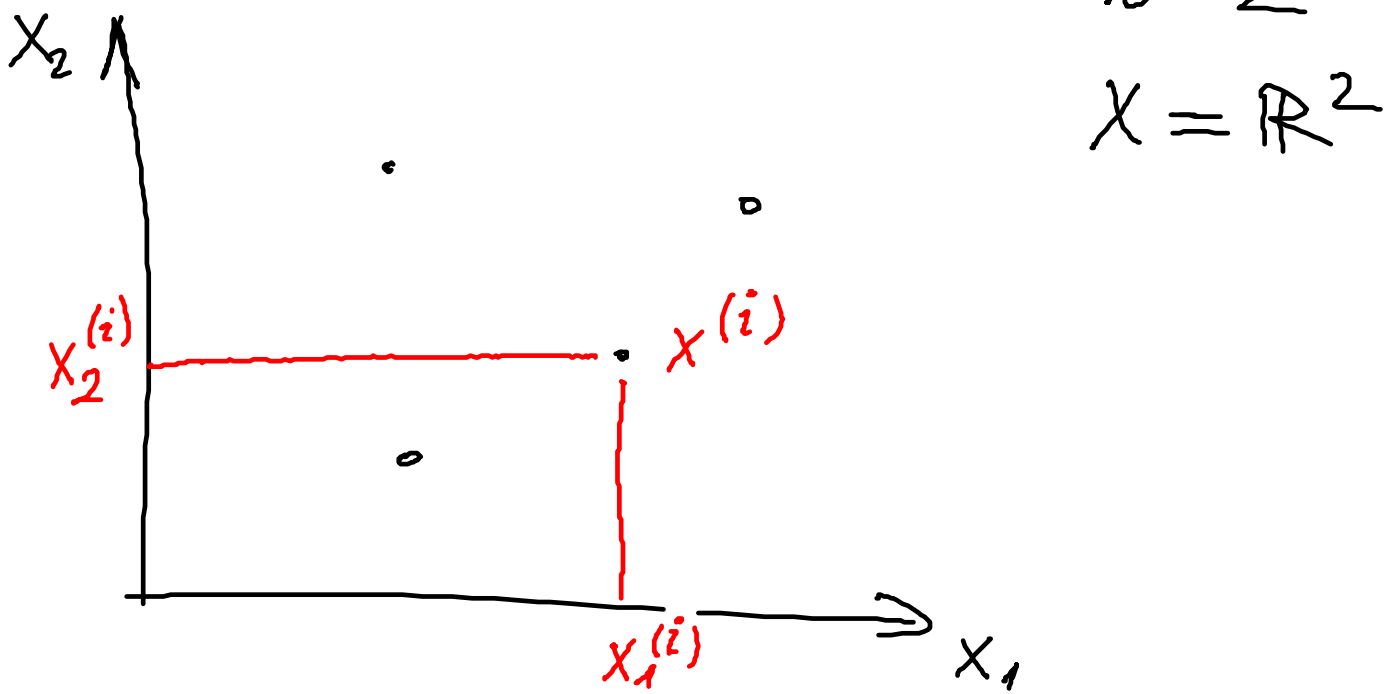
1. (Označavanje skupa podataka) ✗
2. Ekstrakcija značajki ✗
3. (Redukcija dimenzionalnosti)
4. Odabir modela ✓
5. Učenje modela ✓
6. Evaluacija ✓

} su



Primjeri za učenje

- Klasifikacija: odrediti klasu $\mathcal{C}$ kojoj pripada primjer $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathcal{X}$
- $\mathcal{X}$ je n-dimenzijski ulazni prostor (prostor primjera)
- n je broj značajki



Primjeri za učenje

- Nadzirano učenje: poznata nam je oznaka klase y kojoj pripada $\mathbf{x}$
- Skup primjera za učenje: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ → N – broj primjera

x_1	x_2	$\ldots$	x_n	y
$x_1^{(1)}$	$x_2^{(1)}$	$\ldots$	$x_n^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	$\ldots$	$x_n^{(2)}$	$y^{(2)}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$x_1^{(N)}$	$x_2^{(N)}$	$\ldots$	$x_n^{(N)}$	$y^{(N)}$

primjeri N

Poznato!

značajke n

oznake primjera

"Učenje koncepata"

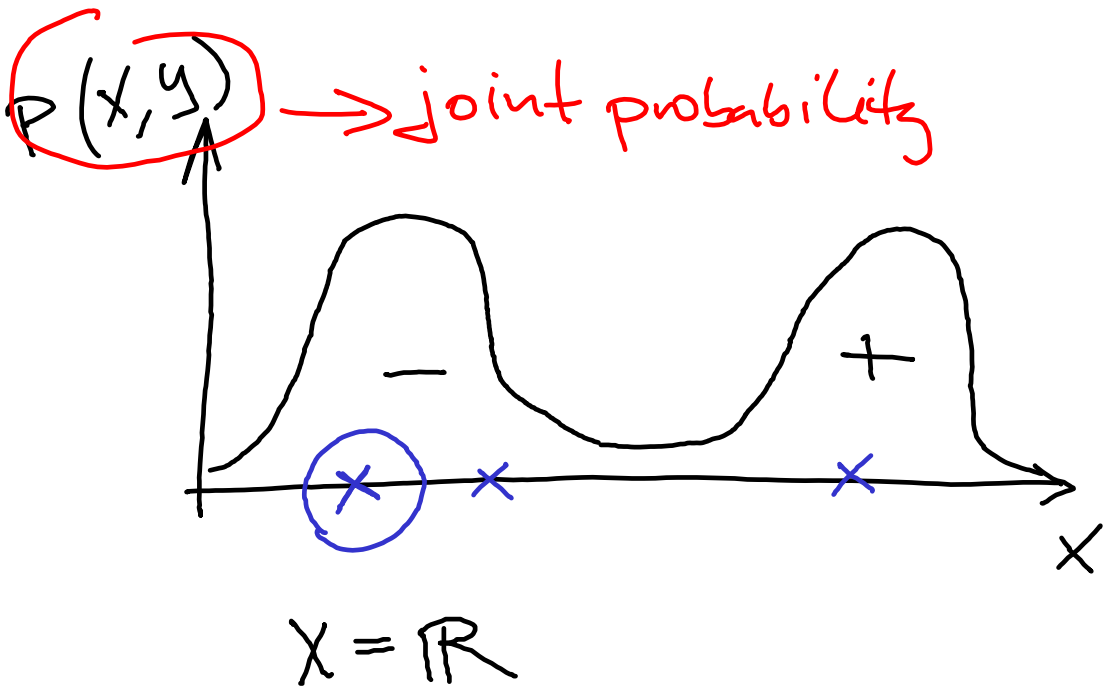
- Binarna klasifikacija:
  $(y = 1) \Rightarrow$ "pozitivan primjer", $(y = 0) \Rightarrow$ "negativan primjer"
  $(y = -1)$

Primjeri za učenje

Pretpostavka iid

Primjeri $\mathbf{x} \in \mathcal{X}$ uzorkovani međusobno nezavisno i iz iste zajedničke distribucije $P(\mathbf{x}, y)$

independently & identically distributed



Od sada nadalje pretpostavljamo uvijek da vrijedi i. i. d.

Hipoteza

- $h : \mathcal{X} \to \{0, 1\}$

funkcija koja klasificira primjere

$$h(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \text{ pripada klasi } \mathcal{C} \\ 0 & \mathbf{x} \text{ ne pripada klasi } \mathcal{C} \end{cases}$$

Zadovoljivost

Primjer $\mathbf{x} \in \mathcal{X}$ zadovoljava hipotezu $h \in \mathcal{H}$ akko $h(\mathbf{x}) = 1$

Konzistentnost

Hipoteza h je konzistentna s primjerom za učenje $(\mathbf{x}, y)$ akko $h(\mathbf{x}) = y$

Hipoteza ispravno klasificira primjer x

Višeklasna klasifikacija (engl. *multiclass*)

- Klasifikacija u klase $\mathcal{C}_j$, gdje $j = 1, \ldots, K$ → broj klasa
- Oznaka klase primjera $\mathbf{x}^{(i)}$ je K-dimenzijski vektor $\mathbf{y}^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_K^{(i)})^{\mathrm{T}}$

vektor oznaka

$$x^{(i)} \in C_j \qquad y_k^{(i)} = \begin{cases} 1 & j = k \\ 0 & \text{inače} \end{cases}$$

$$x^{(1)} \in C_3, \; K = 5 \qquad y^{(1)} = (0, 0, 1, 0, 0) \quad (C_3)$$

- Različito od klasifikacije s višestrukim oznakama (engl. *multilabel*)

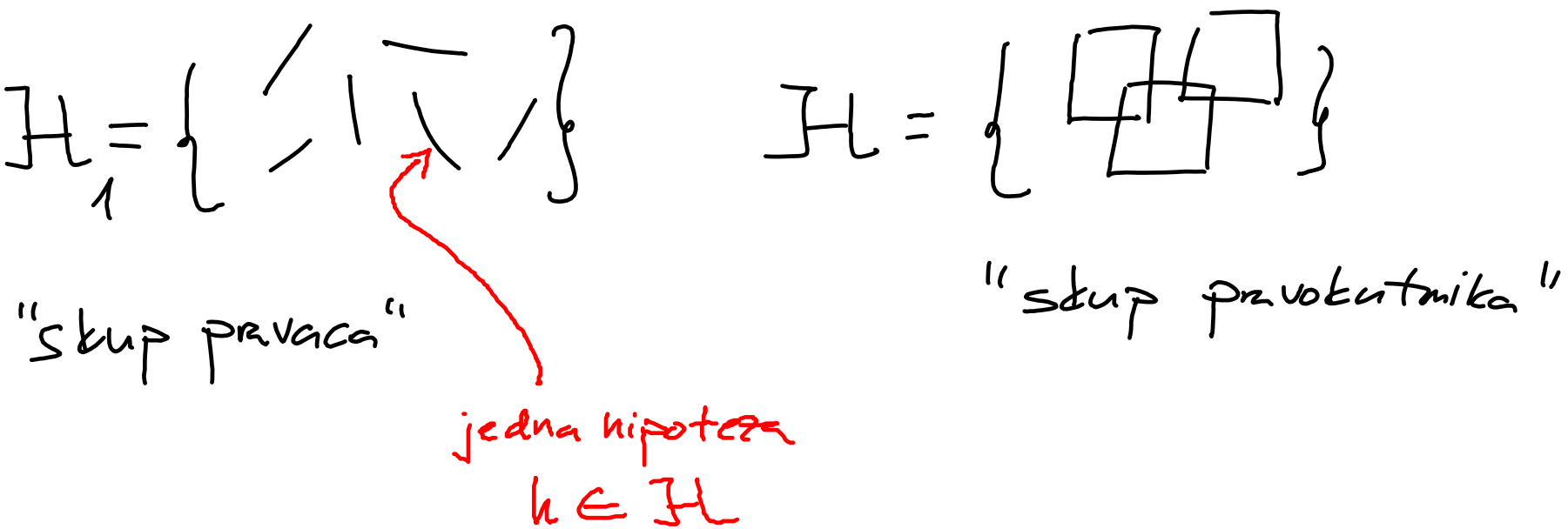
$$x^{(1)} \in \{C_1, C_3\} \qquad y^{(1)} = (1, 0, 1, 0, 0)$$

(Nećemo raditi.)

Model

- Model ili prostor hipoteza: skup mogućih hipoteza $\mathcal{H}$
- Učenje se svodi na pretraživanje prostora hipoteza i nalaženje najbolje hipoteze $h \in H$
- Najbolja hipoteza je ona koja najtočnije klasificira primjere
- $\mathcal{H}$ je vrlo velik – treba nam heurističko pretraživanje

optimiz. problem



Empirijska pogreška

- Empirijska pogreška iskazuje koliko dobro hipoteza klasificira primjere za učenje:

broj pogrešaka

$$E(h|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{h(\mathbf{x}^{(i)}) \neq y^{(i)}\} = \frac{1}{N}\sum_{i=1}^{N} |h(\mathbf{x}^{(i)}) - y^{(i)}|$$

broj primjera

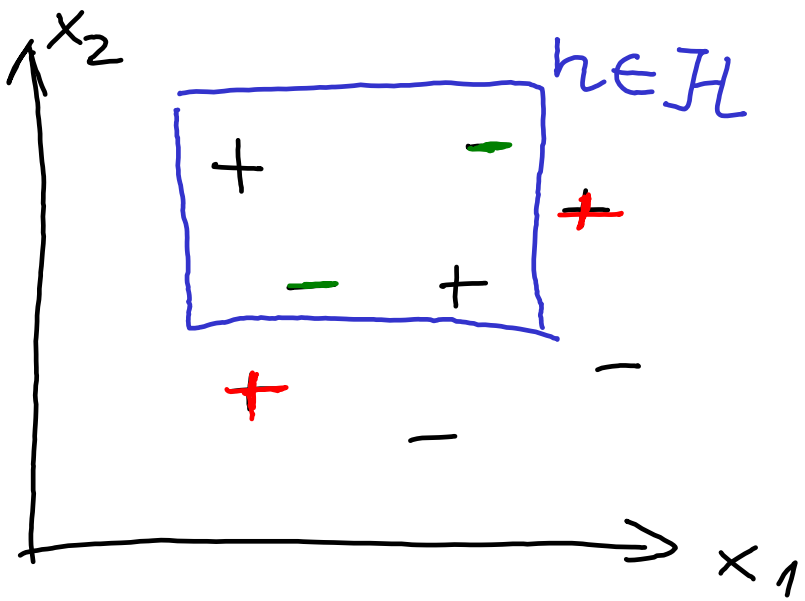
hipoteza

skup primjera za učenje

$1\{P\} = \begin{cases} 1 & P \equiv T \\ 0 & \text{inače} \end{cases}$

$\mathcal{X} = \mathbb{R}^2$

$n = 2$



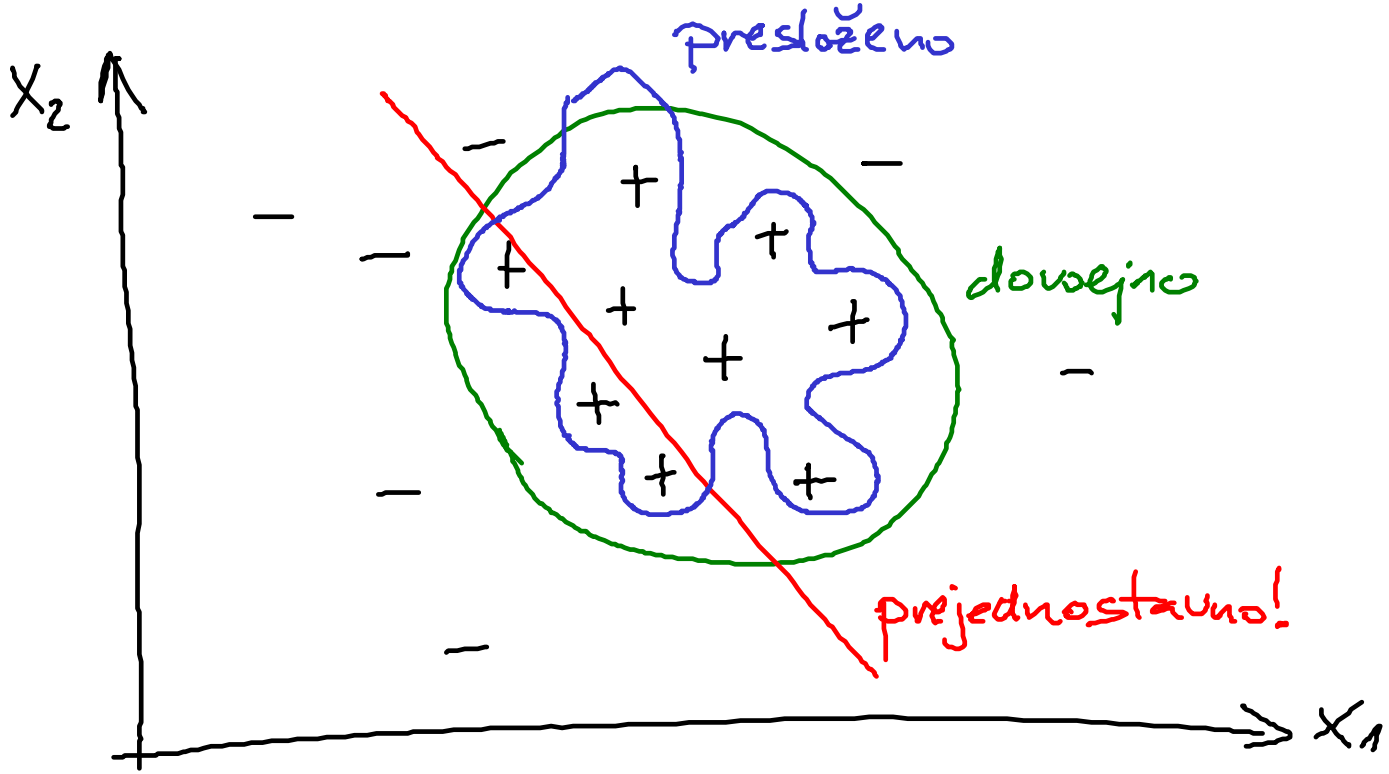
2 pogrešno pozitivna (FP)

2 pogrešno negativna (FN)

$E(h|\mathcal{D}) = \frac{2+2}{8} = 50\,\%$

Složenost/kapacitet modela

- Idealno, $\mathcal{H}$ uključuje klasu $\mathcal{C}$, tj. postoji $h \in \mathcal{H}$ takva da je h konzistentna s $\mathcal{D}$
- No moguće je da takva hipoteza ne postoji, tj. da za sve $h \in \mathcal{H}$ vrijedi $E(h|\mathcal{D}) > 0$
- Tada kažemo da model $\mathcal{H}$ nije dovoljnog kapaciteta ili složenosti



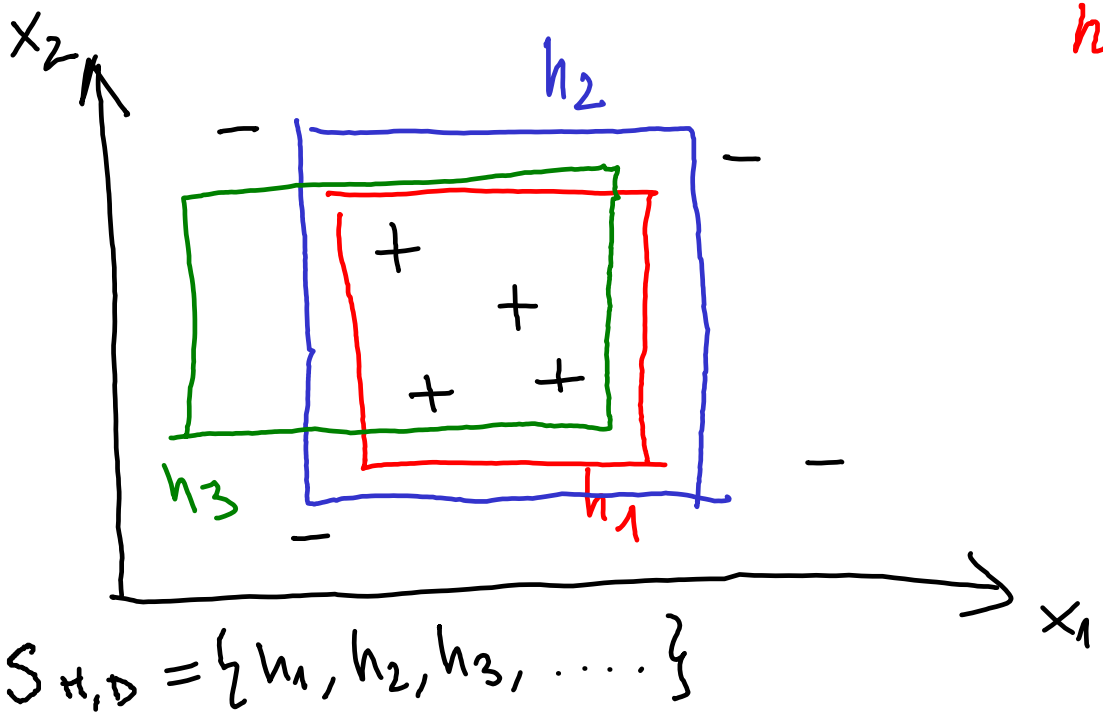
Prostor inačica

Prostor inačica (engl. *version space*)

Prostor inačica $VS_{\mathcal{H},\mathcal{D}} \subseteq \mathcal{H}$ modela $\mathcal{H}$ je skup hipoteza koje su konzistentne s primjerima za učenje $\mathcal{D}$:

$$VS_{\mathcal{H},\mathcal{D}} = \left\{ h \in \mathcal{H} \mid \forall (\mathbf{x}, y) \in \mathcal{D}.\ \left(h(\mathbf{x}) = y\right) \right\}$$

h je konzistentna sa svim primjerima za učenje



$VS_{H,D} = \{h_1, h_2, h_3, \ldots\}$

H = "skup pravokutnika"

$|D| = 8$

$\mathcal{X} = \mathbb{R}^2$

$n = 2$

Danas. . .

1. Osnovni pojmovi
2. Vapnik-Chervonenkisova dimenzija
3. Induktivna pristranost
4. Primjer: regresija
5. Odabir modela

Vapnik-Chervonenkisova dimenzija

- Modeli su različitog kapaciteta: neki su složeniji (fleksibilniji), a neki manje složeni
- Statistička/računalna teorija učenja (engl. *COLT*)
- VC-dimenzija iskazuje kapacitet modela $\mathcal{H}$ u smislu broja primjera za klasifikaciju s kojim se $\mathcal{H}$ može uspješno nositi

$X = \mathbb{R}^2$

$n = 2$

Vapnik-Chervonenkisova dimenzija

skup svih primjera!! ($D \subseteq X$)

Razdjeljivanje primjera

Neka je funkcija $y : \mathcal{X} \to \{0, 1\}$ funkcija koja primjerima iz $\mathcal{X}$ dodjeljuje oznake klase. Model $\mathcal{H}$ razdjeljuje N primjera akko

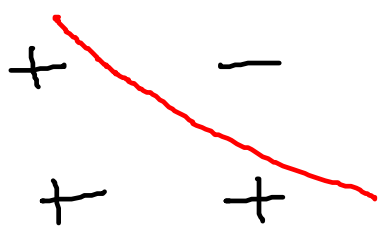
$$\exists\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\} \subseteq \mathcal{X}, \forall y, \exists h \in \mathcal{H}, \forall i \in \{1, \ldots, N\}.$$
$$\left(h(\mathbf{x}^{(i)}) = y(\mathbf{x}^{(i)})\right)$$

primjer je ispravno klasificiran

Npr. $X = \mathbb{R}^2$

$N = 4$

Broj označavanja je $2^N = 2^4 = 16$



← jedno od 16 označavanja

Vapnik-Chervonenkisova dimenzija

VC-dimenzija

VC-dimenzija modela $\mathcal{H}$, označena kao $VC(\mathcal{H})$, jest najveći broj primjera koje model $\mathcal{H}$ može razdijeliti

Napomene:

- odabir N primjera je proizvoljan, ali jednom kad je fiksiran, razdvajanje mora biti moguće za 2^N označavanja
- $VC(\mathcal{H})$ ne ovisi o $\mathcal{D}$ → mjera ne ovisi o konkretnom problemu (konkretnoj distribuciji primjera)
- moguće $VC(\mathcal{H}) = \infty$
- dokaz za $VC(\mathcal{H}) = m$ radimo u dva koraka: (1) $VC(\mathcal{H}) \geq m$ i (2) $VC(\mathcal{H}) \neq m + 1$ → zbog toga
- teorijska i vrlo pesimistična ocjena složenosti modela → u praksi nije tako loše!

Vapnik-Chervonenkisova dimenzija

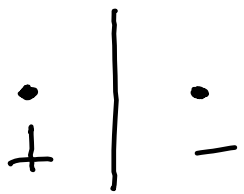
$\mathcal{H}$ = "skup pravaca", $\mathcal{X} = \mathbb{R}^2$, $VC(\mathcal{H}) = ?$

Općenito:

$VC(H) = n+1$ za n-dimenzijske linearne modele

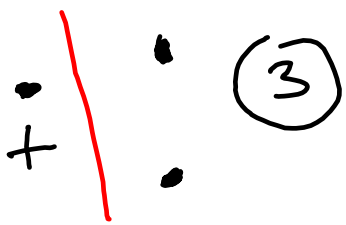
$N = 1$ ✓ (2)

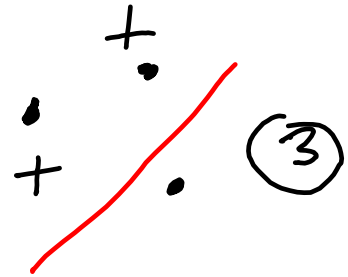
$N = 2$ ✓ (4)



$N = 3$ ✓

$VC(H) \geq 3$ (3)



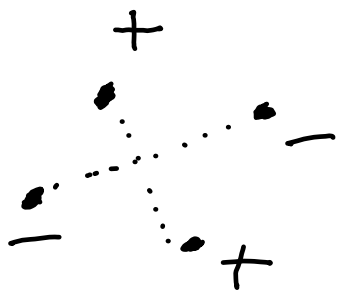


(3) (2)

$VC(H) \neq 4$

$N = 4$

$2^4 = 16$ označavanja



XOR-problem

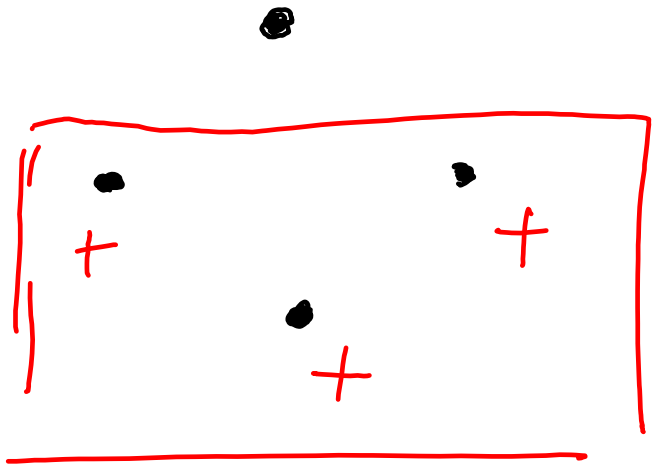
$\Rightarrow VC(H) = 3$

Vapnik-Chervonenkisova dimenzija

$\mathcal{H} =$ "skup pravokutnika čije su stranice poravnate s osima", $\mathcal{X} = \mathbb{R}^2$, $VC(\mathcal{H}) = ?$

$N = 4$

$VC(H) \geq 4$ ✓



$2^4 = 16$

$4 + 2 + 4 + 2 + 4 = 16$ ✓

$VC(H) \neq 5$ ⟵ DZ ⟹ $VC(+1) = 4$

$N = 5$

Danas. . .

1. Osnovni pojmovi
2. Vapnik-Chervonenkisova dimenzija
3. Induktivna pristranost
4. Primjer: regresija
5. Odabir modela

Induktivna pristranost (engl. *inductive bias*)

- Učenje hipoteze je loše definiran problem: h ne slijedi deduktivno iz $\mathcal{D}$
- Npr. učenje Booleove funkcije:

$n = 3$

$\mathcal{X} = \{0, 1\}^3$

8 primjera $\mathcal{X}$

x_1	x_2	x_3	y	
0	0	0	?	
0	0	1	?	
0	1	0	1	✓
0	1	1	0	✓
1	0	0	1	✓
1	0	1	0	✓
1	1	0	?	
1	1	1	1	✓

$|\mathcal{D}| = 5$

- $n = 3,\ N = 5,\ |VS| = 2^{2^n - N} = 8$ → 8 mogućih hipoteza
- Generalizacija: mogućnost klasifikacije još neviđenih primjera
- Učenje i generalizacija nisu mogući bez dodatnih pretpostavki!

Induktivna pristranost (engl. *inductive bias*)

Induktivna pristranost

$\mathcal{L}$ – algoritam za učenje
$h_{\mathcal{L}}$ – hipoteza inducirana pomoću $\mathcal{L}$ na $\mathcal{D}$
$h_{\mathcal{L}}(\mathbf{x})$ – klasifikacija primjera $\mathbf{x} \in \mathcal{X}$.

Induktivna pristranost od $\mathcal{L}$ je bilo koji skup minimalnih pretpostavki $\mathcal{B}$ takvih da

$$\forall \mathcal{D}. \forall \mathbf{x} \in \mathcal{X}. \left((\mathcal{B} \wedge \mathcal{D} \wedge \mathbf{x}) \vdash h_{\mathcal{L}}(\mathbf{x}) \right)$$

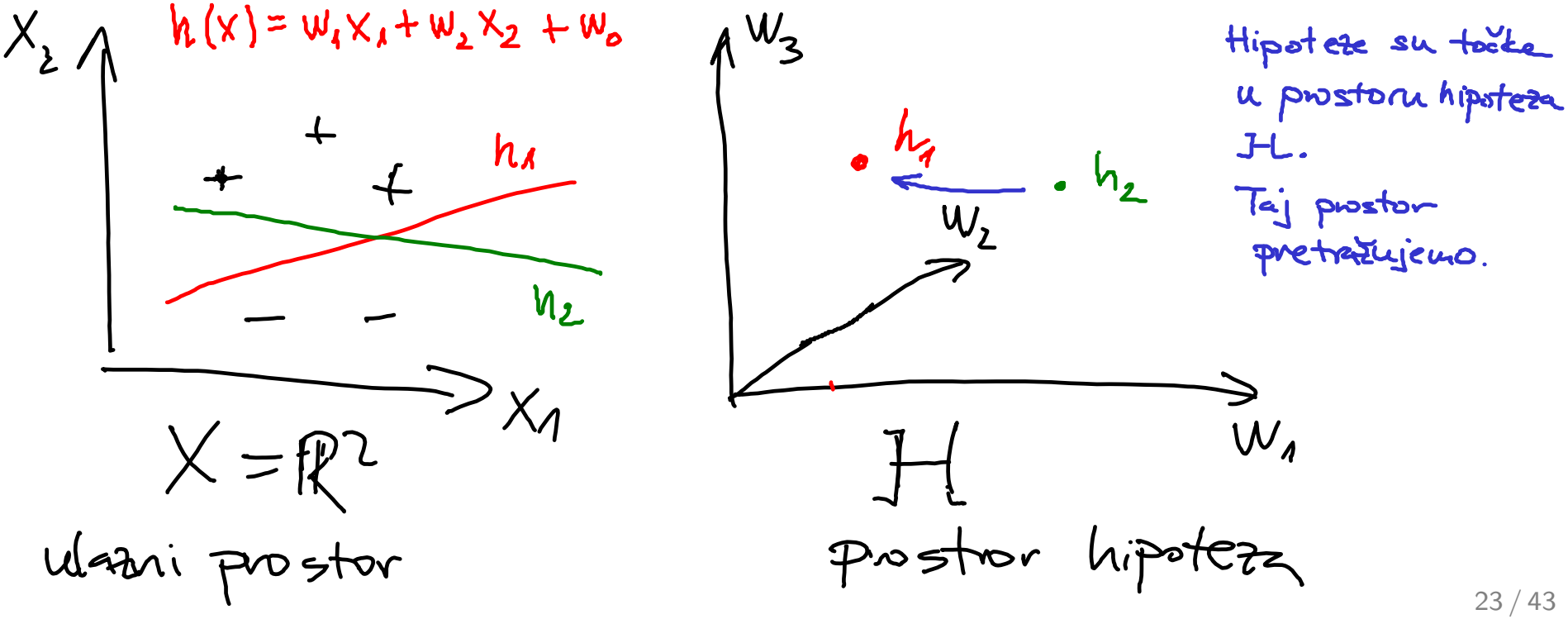
Slijedi deduktivno

Skup pretpostavki koje od indukcije čine dedukciju!

Induktivna pristranost (engl. *inductive bias*)

1. Pristranost ograničenjem (pristranost jezika) – odabiremo model $\mathcal{H}$ koji ograničava skup prikazivih hipoteza
2. Pristranost pretraživanja (pristranost preferencijom) – definiramo način pretraživanja unutar $\mathcal{H}$ → definiramo optimizacijski postupak

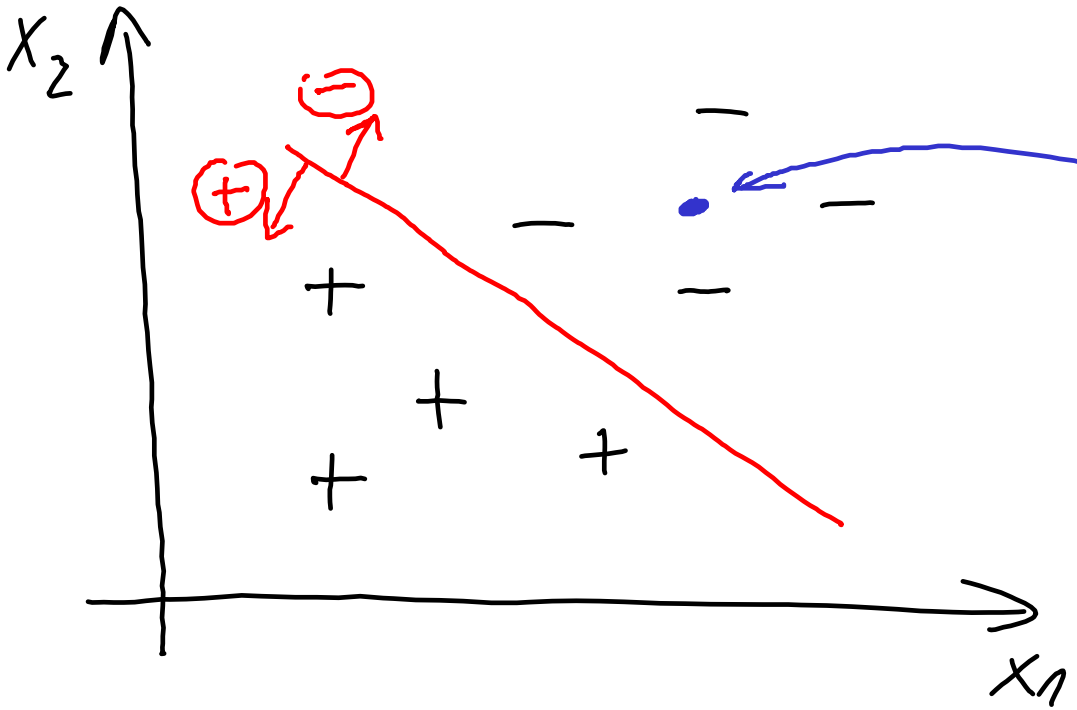
- Većina algoritama kombinira obje pristranosti



Induktivna pristranost – primjer 1

Učenje binarne Booleove funkcije, $\mathcal{X} = \{0, 1\}^2$, $\mathcal{H}$ je skup pravaca u $\mathbb{R}^2$

Pristranost jezika!



Ovo je dosad neviđeni primjer. Model će ga klasificirati kao negativnog

GENERALIZACIJA!

(zahvaljujući induktivnoj pristranosti)

Induktivna pristranost – primjer 2

Učenje ternarne Booleove funkcije, $\mathcal{X} = \{0,1\}^3$, $\mathcal{H}$ je skup ravnina u $\mathbb{R}^3$

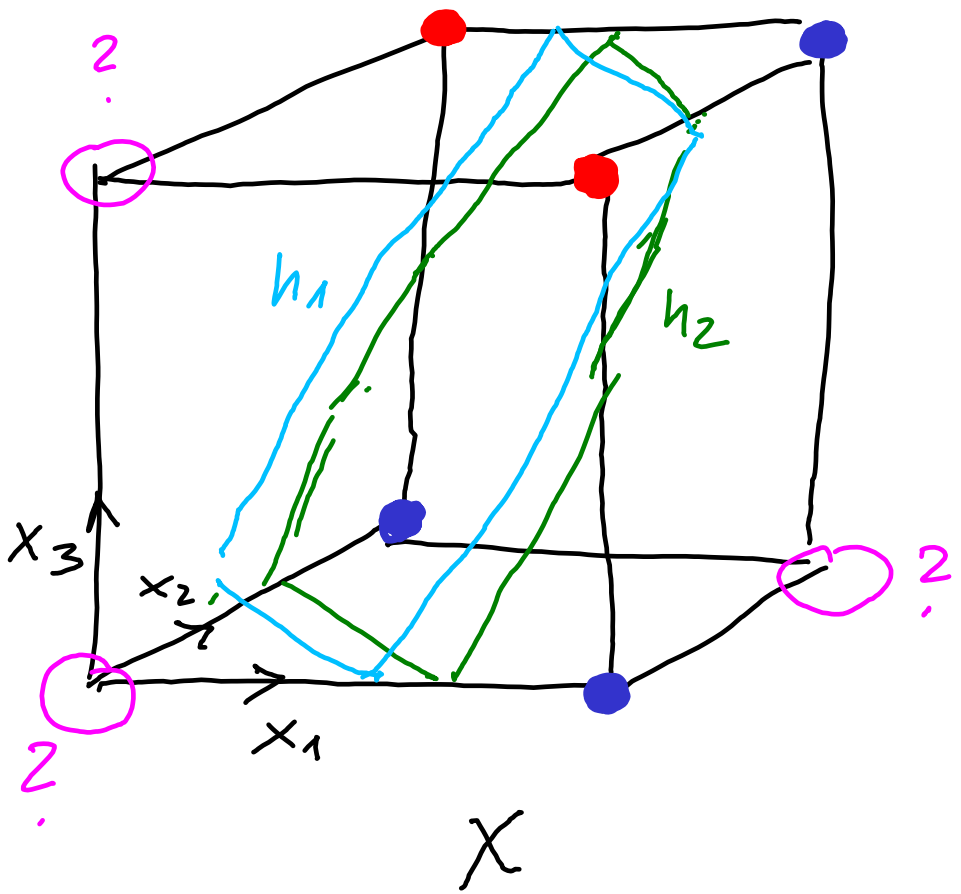
Pristranost jezika!

x_1	x_2	x_3	y
0	0	0	?
0	0	1	?
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	?
1	1	1	1

$|D| = 5$

$n = 3$

$|VS| = ?$



(Napomena: h_1 i h_2 su identične hipoteze u $X = \{0,1\}^3$)

Induktivna pristranost – primjer 3

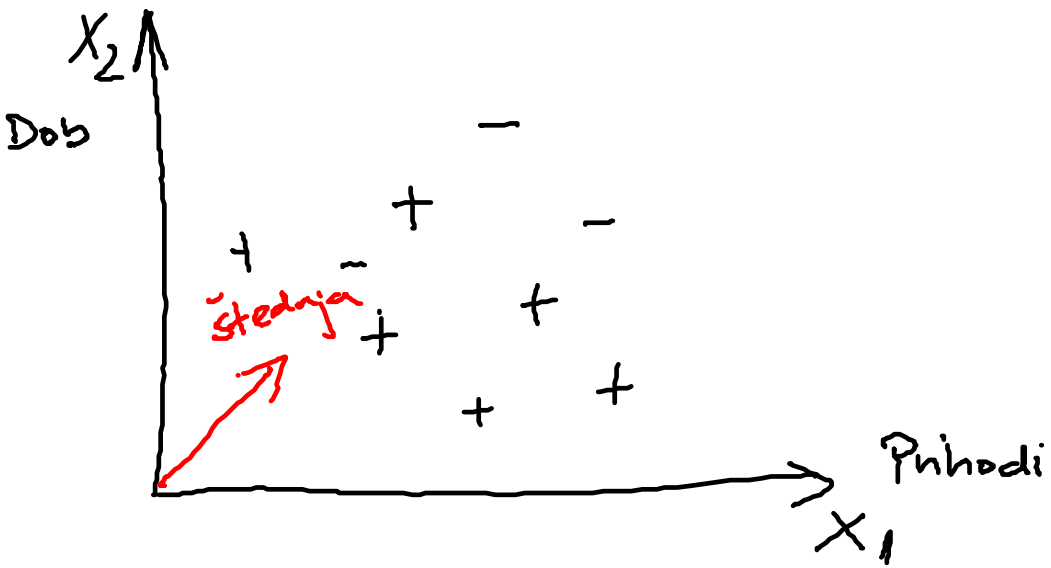
Učenje ternarne Booleove funkcije, $\mathcal{X} = \{0,1\}^3$, $\mathcal{H}$ je skup ravnina u $\mathbb{R}^3$

x_1	x_2	x_3	y
0	0	0	?
0	0	1	?
0	1	0	1
0	1	1	0
1	0	0	1
1	1	0	?
1	1	1	1

→ za DZ

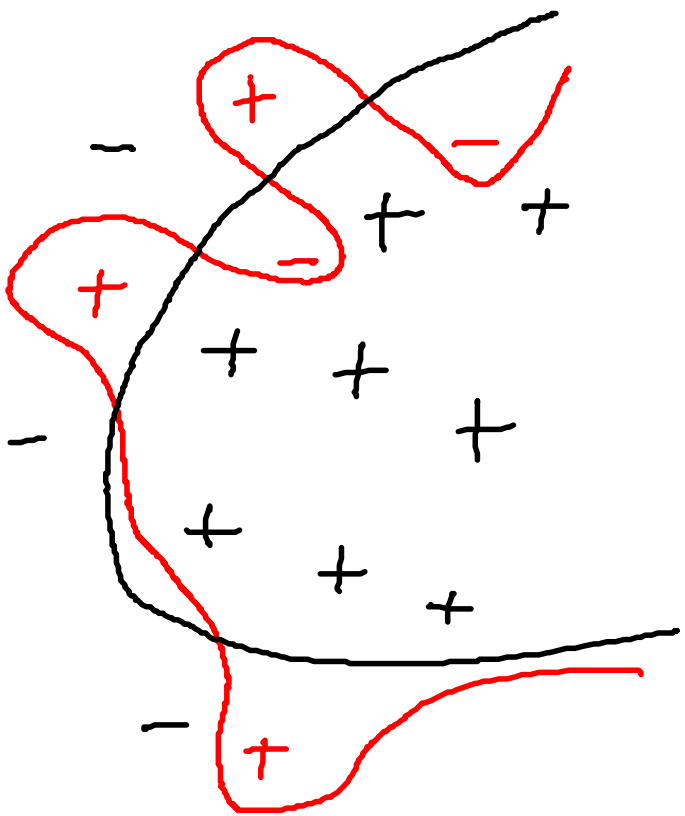
Problem šuma

- Šum je neželjena anomalija u podacima
- Mogući uzroci:
  1. nepreciznost pri mjerenju značajki,
  2. pogreške u označavanju (engl. *teacher noise*),
  3. postojanje skrivenih značajki (latentnih varijabli),
  4. nejasne granice klasa (subjektivnost).
- Zbog šuma je granica između pozitivnih i negativnih primjera složena!



Problem šuma

- Jednostavni modeli ne mogu doseći $E(h|\mathcal{D}) = 0$
- Složeni modeli uče šum, a ne pravu klasifikaciju
- Šum u principu nije moguće odvojiti od pravih podataka
- Moguće za: vrijednosti koje odskaču (engl. *outliers*)



Danas. . .

1. Osnovni pojmovi
2. Vapnik-Chervonenkisova dimenzija
3. Induktivna pristranost
4. Primjer: regresija
5. Odabir modela

Regresija

- $y \in \mathbb{R}$
- Na temelju primjera $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$ učimo hipotezu (funkciju) h koja aproksimira nepoznatu funkciju $f : \mathcal{X} \to \mathbb{R}$
- Idealno, $y^{(i)} = f(\mathbf{x}^{(i)})$, ali zbog šuma $y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon$

- Empirijska pogreška hipoteze:

$$E(h|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - h(\mathbf{x}^{(i)})\right)^2$$

Kvadratno odstupanje

Regresija

- Linearan model:

$$h_1(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_0 = \sum_{i=1}^{n} w_i x_i + w_0 = \mathbf{w}^{\mathrm{T}} \mathbf{x} + w_0$$

- Neka $\mathcal{X} = \mathbb{R}$. Onda je model:

$$h_1(x) = w_1 x + w_0$$

a empirijska pogreška:

$$E(h_1 | \mathcal{D}) = \frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - (w_1 x^{(i)} + w_0) \right)^2$$

- Naš cilj je pronaći h_1 koja minimizira $E(h_1|\mathcal{D})$
  $\Rightarrow$ Postupak najmanjih kvadrata (engl. *least squares*)

Regresija

- Ovo je optimizacijski problem: treba pronaći parametre w_0 i w_1 koji minimiziraju $E(h|\mathcal{D})$. U ovom slučaju postoji analitičko rješenje

$$\nabla_{w_0,w_1} E(h_1|\mathcal{D}) = 0$$

$$\frac{\partial}{\partial w_0}\left[\frac{1}{2}\sum_i^N \left(y^{(i)} - (w_1 x^{(i)} + w_0)\right)^2\right] = 0$$

$$\vdots$$

$$w_0 = \bar{y} - w_1\bar{x}$$

$$\frac{\partial}{\partial w_1}\left[\frac{1}{2}\sum_i^N \left(y^{(i)} - (w_1 x^{(i)} + w_0)\right)^2\right] = 0$$

$$\vdots$$

$$w_1 = \frac{\sum_i^N x^{(i)} y^{(i)} - N\bar{x}\bar{y}}{\sum_i^N (x^{(i)})^2 - N\bar{x}^2}$$

Matematiku prodite za DZ!

Rješenje je u zatvorenoj formi (closed form solution)

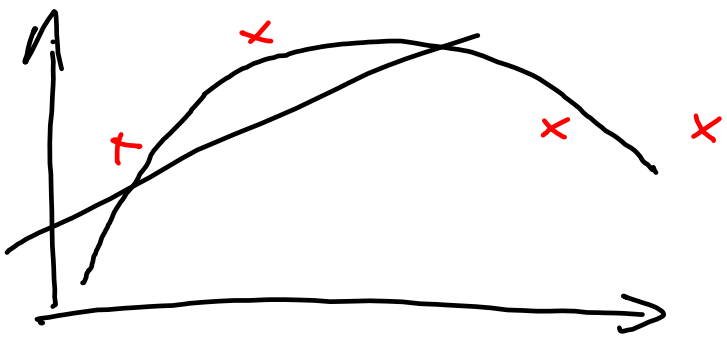
Regresija

- Mogli smo odabrati i složeniji model, npr. polinom drugog stupnja:

$$h_2(x) = w_2 x^2 + w_1 x + w_0$$

- Ovo je i dalje linearna regresija, i dalje ima analitičko rješenje

$$E(h_1|\mathcal{D}) \geq E(h_2|\mathcal{D})$$



Model 2. stupnja je složeniji, pa se bolje prilagođava primjercima.

- Koji model odabrati? Ovisi!

Odabir modela

- Moramo odabrati model $\mathcal{H}$ (učenje bez pristranosti je uzaludno)!
- Često radimo odabir modela unutar neke familije modela – npr. odabir stupnja polinoma

- Stupanj polinoma je hiperparametar modela (w_i su parametri)
- Odabir modela = optimizacija modela, odabir parametara

Odabir modela

- Što veći kapacitet modela $\mathcal{H}$, to je manja pogreška $E(h|\mathcal{D})$, $h \in \mathcal{H}$
- Ali model mora moći generalizirati!
- Preferiramo jednostavne modele
  - bolja generalizacija
  - lakše učenje/uporaba
  - lakše tumačenje

  $\Rightarrow$ Occamova britva
- Trebamo odabrati model koji točno odgovara pravoj složenosti funkcije koju nastojimo naučiti

Odabir modela

1. Prenaučenost – $\mathcal{H}$ je previše složen u odnosu na stvarnu funkciju
   $\Rightarrow$ loša generalizacija!

2. Podnaučenost – $\mathcal{H}$ je prejednostavan u odnosu na stvarnu funkciju
   $\Rightarrow$ loša klasifikacija viđenih i neviđenih primjera!

Odabir modela

- Jednostavan model ima visoku pristranost (engl. *high bias*)
- Složen model ima visoku varijancu (engl. *high variance*)

- Odabir modela $\Rightarrow$ dvojba između pristranosti i varijance
- Optimalan model minimizira i pristranost i varijancu

Odabir modela

Pretpostavka induktivnog učenja

Ako je (1) pogreška hipoteze na dovoljno velikom skupu primjera za učenje mala i (2) ako model nije suviše složen, hipoteza će dobro klasificirati i nove, (3) slične primjere.

Unakrsna provjera (engl. *cross-validation*)

- Metoda za procjenu sposobnosti generalizacije modela
- Skup primjera dijelimo na skup za učenje i skup za ispitivanje
- Model učimo na skupu za učenje, a ispitujemo na skupu za ispitivanje
- Primjeri iz skupa za ispitivanje model dosad nije vidio, pa dobivamo dobru (pravednu) procjenu pogreške generalizacije

Unakrsna provjera

- Što ako želimo optimirati parametre modela?
- Ne možemo to raditi na skupu za provjeru!
- Trebamo još jedan skup: skup za provjeru (engl. *validation set*)

$$\mathcal{D} = \mathcal{D}_U \cup \mathcal{D}_P \cup \mathcal{D}_I$$

$$\mathcal{D}_U \cap \mathcal{D}_P = D_U \cap \mathcal{D}_I = \mathcal{D}_P \cap \mathcal{D}_I = \emptyset$$

Unakrsna provjera

- Empirijska pogreška $E(h|\mathcal{D}_U)$ pada sa složenošću modela
- Pogreška generalizacije $E(h|\mathcal{D}_P)$ tipično pada pa raste
- Optimalan model je onaj koji minimizira $E(h|\mathcal{D}_P)$

Sažetak

- **Hipoteza** je funkcija koja klasificira primjere, a **model** je skup hipoteza
- Različiti modeli imaju različite **složenosti**. Jedna teorijska mjera složenosti modela je **Vapnik-Chervonenkisova dimenzija**
- Učenje nije moguće bez **induktivne pristranosti**, koja može biti **pristranost ograničenjem** ili **pristranost pretraživanja**
- Učenje se svodi na **optimizaciju parametara** modela. Kod regresije postoji **analitičko rješenje** za taj problem
- Model koji je **prenaučen** ili **podnaučen** loše generalizira
- Odabir modela svodi se na **optimiranje hiperparametara** modela
- **Unakrsnom provjerom** može se procijeniti **pogreška generalizacije** i odabrati optimalan model

Sljedeća tema: Nadzirano učenje (nastavak)