Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu
Strojno učenje, ak. god. 2011/12. – **bilješka za predavanja** (v2.2)

Bilješka 1
Nadzirano učenje

Postupcima nadziranog učenja mogu se rješavati dvije vrste problema: **klasifikacija** i **regresija**. Kod klasifikacije primjeru pridružujemo **klasu** (razred) kojoj taj primjer pripada. Kod regresije primjeru pridružujemo neku kontinuiranu vrijednost. Razlika je dakle u tome je li ciljna varijabla diskretna ili **nominalna** (klasifikacija) ili je kontinuirana (regresija). Razmotrimo najprije klasifikaciju.

1 Osnovni pojmovi

1.1 Primjeri za učenje

Svrha klasifikacije jest odrediti klasu $\mathcal{C}$ kojoj pripada primjer $\mathbf{x}$. Primjer ćemo definirati kao vektor značajki, $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, gdje je n dimenzija vektora. Primjeri se mogu interpretirati kao točke u n-dimenzijskom vektorskom prostoru koji nazivamo **ulazni prostor** (engl. *input space*) ili **prostor primjera** (engl. *instance space*). Neka je $\mathcal{X}$ skup svih mogućih primjera. Pretpostavka svih algoritama strojnog učenja jest da su primjeri iz $\mathcal{X}$ uzorkovani nezavisno i iz iste zajedničke distribucije $P(\mathbf{x}, y)$. Ta se pretpostavka skraćeno označava s **iid** (engl. *independent and identically distributed*).

Kod nadziranog učenja unaprijed nam je poznata **oznaka klase** y kojoj pripada primjer $\mathbf{x}$ iz skupa za učenje. Ako se ograničimo na samo dvije klase, onda $y \in \{0, 1\}$, gdje $y = 1$ označava da je primjer za jednu (od ukupno dvije) klase pozitivan (pripada klasi), a $y = 0$ označava da je primjer za tu klasu negativan (ne pripada toj klasi). Klasifikator koji klasificira u dvije klase nazivamo **binarni klasifikator**. Učenje binarnog klasifikatora istovjetno je učenju Booleove funkcije, koje se u literaturi naziva i **učenje koncepta** (engl. *concept learning*).

Skup primjera za učenje $\mathcal{D}$ sastoji se od parova primjera i pripadnih oznaka, $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, gdje je N ukupan broj primjera za učenje, a i je indeks primjera odnosno njemu pripadne oznake. Skup $\mathcal{D}$ možemo prikazati tablično:

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

1.2 Hipoteza

Zadaća klasifikacijskog algoritma jest inducirati (naučiti) **hipotezu** $h : \mathcal{X} \to \{0, 1\}$ koja određuje pripada li neki primjer $\mathbf{x}$ klasi $\mathcal{C}$ ili ne:

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

Definicija 1 *Kažemo da primjer* $\mathbf{x} \in \mathcal{X}$ **zadovoljava** *hipotezu* $h \in \mathcal{H}$ *akko* $h(\mathbf{x}) = 1$.

Definicija 2 *Kažemo da je hipoteza* h **konzistentna** *s primjerom za učenje* $(\mathbf{x}, y)$ *akko* $h(\mathbf{x}) = y$. *Konzistentnost hipoteze sa svim primjerima za učenje iz* $\mathcal{D}$ *definiramo kao*

$$Consistent(h, \mathcal{D}) \iff \forall (\mathbf{x}, y) \in \mathcal{D}.\ \big(h(\mathbf{x}) = y\big)\,.$$

Često primjere treba klasificirati u više od jedne klase, što nazivamo **višeklasnom klasifikacijom** (engl. *multiclass classification*). Npr. klasifikacija novinskih članaka u rubrike ili klasifikacija rukom pisanih znamenki. U općenitom slučaju postoji K klasa, $\mathcal{C}_j$, gdje $j = 1, \ldots, K$. Tada je najprikladnije oznaku klase primjera $\mathbf{x}^{(i)}$ prikazati kao K-dimenzijski vektor, $\mathbf{y}^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_K^{(i)})^{\mathrm{T}}$ gdje

$$y_j^{(i)} = \begin{cases} 1 & \text{ako } \mathbf{x}^{(i)} \in \mathcal{C}_j \\ 0 & \text{inače} \end{cases}$$

Npr. $\mathbf{y}^{(2)} = (0, 0, 1, 0)^{\mathrm{T}}$ značilo bi da primjer $\mathbf{x}^{(2)}$ pripada klasi $\mathcal{C}_3$.

U još općenitijem slučaju jedan primjer može istovremeno pripadati u više klasa, što nazivamo **klasifikacija s višestrukim oznakama** (engl. *multilabel classification*) ili **klasifikacija jedan-na-više**; klasifikaciju tog tipa nećemo posebno razmatrati budući da se da izvesti kao klasifikacija tipa jedan-na-jedan.

1.3 Model

Hipoteze ne izmišljamo ni iz čega, nego ih odabiremo iz pomno odabranog skupa mogućih hipoteza. Skup mogućih hipoteza $\mathcal{H}$ nazivamo **model** ili **prostor hipoteza** (ova dva pojma koristit ćemo ravnopravno). Npr. ako je prostor primjera dvodimenzionalan, $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$, klasa hipoteze može biti pravac koji razdjeljuje primjere dviju klasa ili to može biti pravokutnik koji obuhvaća sve primjere neke klase. Model $\mathcal{H}$ dakle određuje način prikaza hipoteze.

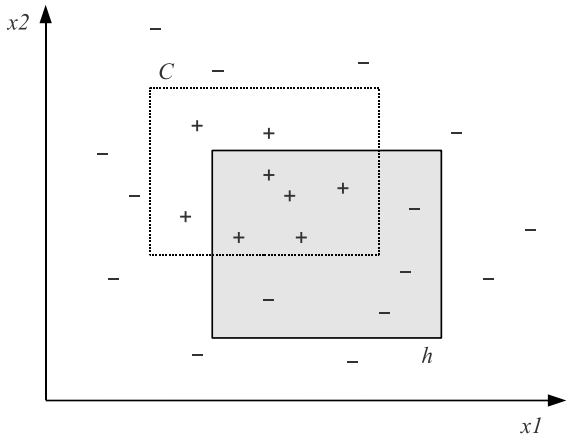
Učenje se zapravo svodi na pretraživanje prostora hipoteza $\mathcal{H}$ i nalaženje najbolje hipoteze $h \in H$. Najbolja hipoteza je ona koja najtočnije klasificira primjere. (Naravno, slijepo pretraživanje prostora $\mathcal{H}$ ne dolazi u obzir budući da je $\mathcal{H}$ obično vrlo velik.)

Koliko dobro hipoteza h klasificira primjere za učenje iskazuje **empirijska pogreška** ili **pogreška učenja** (engl. *training error*). Empirijska pogreška hipoteze h, mjerena na skupu $\mathcal{D}$, jednaka je udjelu primjera iz $\mathcal{D}$ koji nisu ispravno klasificirani:

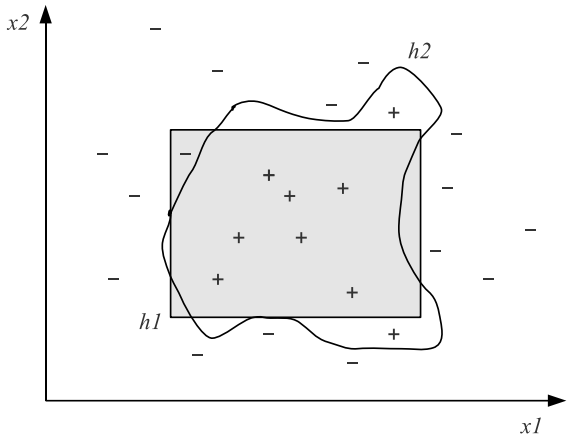
$$E(h|\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{h(\mathbf{x}^{(i)}) \neq y^{(i)}\} = \frac{1}{N} \sum_{i=1}^{N} |h(\mathbf{x}^{(i)}) - y^{(i)}|$$

gdje je $\mathbf{1}\{P\}$ indikatorska funkcija čija je vrijednost 1 ako $P \equiv \top$, a 0 inače. Očito, hipoteza je konzistentna s primjerima za učenje akko $E(h|\mathcal{D}) = 0$. Kao primjer razmotrimo hipotezu sa slike 1: od ukupno 23 primjera, hipoteza pogrešno klasificira njih sedam, pa $E(h|D) = 7/23$. Primjeri koje hipoteza klasificira pozitivno, a zapravo su negativni, zovemo **lažno pozitivni** primjeri (engl. *false positives*, FP). Obrnuto, primjeri koje hipoteza klasifcira negativno, a zapravo su pozitivni, zovemo **lažno negativni** primjeri (engl. *false negatives*, FN).

Idealno, prostor hipoteza $\mathcal{H}$ uključuje klasu $\mathcal{C}$, tj. postoji $h \in \mathcal{H}$ takva da je h konzistentna s primjerima za učenje. 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



Slika 1: Prostor primjera $\mathcal{X} = \mathbb{R}^2$, područje koje zadovoljava hipotezu h te područje koje odgovara klasi $\mathcal{C}$.



Slika 2: Prostor primjera $X = \mathbb{R}^2$ i područje koje zadovoljava hipotezu h_1 (iz jednostavnijeg modela) odnosno hipotezu h_2 (iz složenijeg modela).

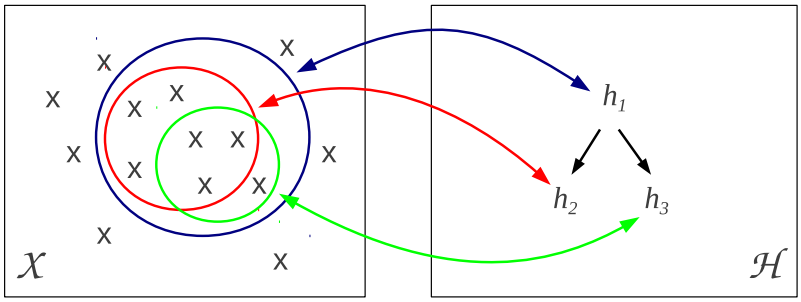
složenosti) da bi naučio klasu $\mathcal{C}$. Razmotrimo kao primjer prostor primjera sa slike 2. Ako je model $\mathcal{H}$ skup pravokutnika poravnatih s osima, onda niti jedna hipoteza iz $\mathcal{H}$ (npr. h_1) ne može naučiti klasu $\mathcal{C}$. U tom slučaju treba nam složeniji model (model u kojemu bismo mogli prikazati npr. hipotezu h_2).

1.4 *Prostor inačica

Moguće je (i zapravo je vrlo često) da za neki skup za učenje $\mathcal{D}$ postoji više (moguće beskonačno mnogo) hipoteza modela $\mathcal{H}$ koje ispravno klasificiraju primjere iz $\mathcal{D}$. Skup takvih hipoteza nazivamo **prostor inačica** (engl. *version space*).

Definicija 3 (Prostor inačica) *Prostor inačica $VS_{\mathcal{H},\mathcal{D}} \subseteq \mathcal{H}$ modela $\mathcal{H}$ jest skup hipoteza koje su konzistentne s primjerima za učenje $\mathcal{D}$:*

$$VS_{\mathcal{H},\mathcal{D}} = \{h \in \mathcal{H} \mid Consistent(h, \mathcal{D})\}\,. \tag{1}$$



Slika 3: Prostor primjera $\mathcal{X} = \mathbb{R}^2$ i odgovarajući uređaj u prostoru hipoteza $\mathcal{H}$.

Prostor inačica može se zapisati kompaktnije ako se u obzir uzmu odnosi između hipoteza, kao što prikazuje slika 3. Za neki odabrani model $\mathcal{H}$, hipoteze $h \in \mathcal{H}$ razlikuju se po svojoj općenitosti, odnosno po tome koliko ih primjera može zadovoljiti.

Definicija 4 (Relacija općenitija-ili-jednaka) *Kažemo da je hipoteza h_1 općenitija-ili-jednaka od hipoteze h_2, i pišemo $h_1 \geqslant_g h_2$, akko svi primjeri koji zadovoljavaju h_2 također zadovoljavaju i h_1:*

$$h_1 \geqslant_g h_2 \iff \forall \mathbf{x} \in \mathcal{X}. \big((h_2(\mathbf{x}) = 1) \Rightarrow (h_1(\mathbf{x}) = 1)\big) .$$

Relacija $\geqslant_g$ je relacija **parcijalnog uređaja** (refleksivna je, antisimetrična i tranzitivna), odnosno $(\mathcal{H}, \geqslant_g)$ je parcijalno uređen skup (*poset*). Skup je parcijalan jer nije svaki par hipoteza iz $\mathcal{H}$ međusobno usporediv (npr. hipoteze h_2 i h_3 sa slike 3 nisu usporedive).

Neka je S skup maksimalno specifičnih hipoteza konzistentnih s primjerima za učenje, a G skup maksimalno općenitih hipoteza konzistentnih s primjerima za učenje. Formalno:

$$\begin{aligned} S &= \big\{s \in \mathcal{H} \mid \mathit{Consistent}(s, \mathcal{D}) \ \wedge \ \forall h \in \mathcal{H}. \big((s \geqslant_g h) \wedge \mathit{Consistent}(h, \mathcal{D}) \Rightarrow (h \geqslant_g s)\big)\big\} \\ G &= \big\{g \in \mathcal{H} \mid \mathit{Consistent}(g, \mathcal{D}) \ \wedge \ \forall h \in \mathcal{H}. \big((h \geqslant_g g) \wedge \mathit{Consistent}(h, \mathcal{D}) \Rightarrow (g \geqslant_g h)\big)\big\} \end{aligned}$$

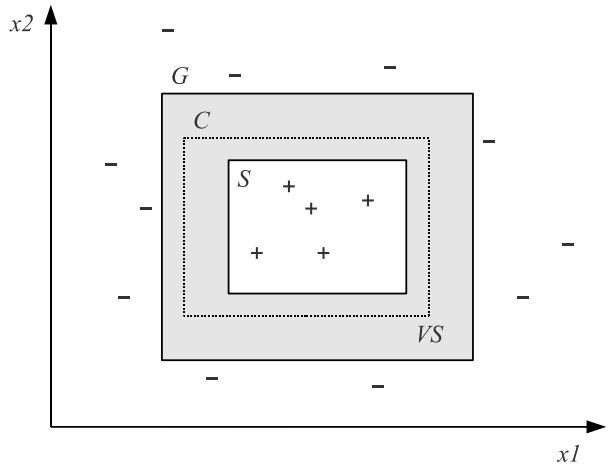
Prostor inačica sada se može sažeto prikazati kao skup svih hipoteza koje su specifičnije od neke hipoteze (ili jednake nekoj hipotezi) iz skupa G i općenitije od neke hipoteze (ili jednake nekoj hipotezi) iz S:

$$VS_{\mathcal{H},\mathcal{D}} = \{h \in \mathcal{H} \mid \exists g \in G, \exists s \in S. (g \geqslant_g h \geqslant_g s)\} . \tag{2}$$

Ova definicija ekvivalentna je definiciji (1). Na prvi pogled možda nije očito da su hipoteze h za koje vrijedi $g \geqslant_g h \geqslant_g s$ ujedno i konzistente s primjerima za učenje $\mathcal{D}$. Razmotrimo zašto je to ipak slučaj. Hipotezu s po definiciji skupa S zadovoljavaju svi pozitivni primjeri. Budući da $h \geqslant_g s$, svi pozitivni primjeri zadovoljavaju i hipotezu h. Slično, hipotezu g po definiciji skupa G ne zadovoljava niti jedan negativan primjer. Budući da $g \geqslant_g h$, to niti jedan negativni primjer ne zadovoljava hipotezu h. Budući da h zadovoljavaju svi pozitivni primjeri, a niti jedan negativan, to je hipoteza h konzistentna s primjerima za učenje.

Slika 4 prikazuje prostor primjera $\mathcal{X} = \mathbb{R}^2$ te pravokutnike koji odgovaraju najopćenitijoj i najspecifičnijoj hipotezi. Zasivljeno područje odgovara hipotezama iz prostora inačica.

Ako $G = S$, prostor inačica sadrži hipoteze koje su jednako općenite (a idealno samo jednu takvu hipotezu), pa možemo reći da je klasa $\mathcal{C}$ potpuno naučena. S druge strane,



Slika 4: Prostor primjera $\mathcal{X} = \mathbb{R}^2$ i prostor inačica.

ako $VS = \varnothing$, onda ne postoji hipoteza $h \in \mathcal{H}$ koja bi bila konzistentna s primjerima $\mathcal{D}$, odnosno model $\mathcal{H}$ nije dovoljnog kapaciteta.

Premda je svaka hipoteza iz prostora inačica konzistentna s primjerima za učenje, ne znači da svaka hipoteza ispravno klasificira još neviđene primjere. Činjenica da prostor inačica postoji posljedica je toga što u skupu za učenje $\mathcal{D}$ nije postojao niti jedan primjer koji bi u ulaznom prostoru pao u područje koje odgovara prostoru inačica. Primjer iz tog područja je svaki onaj koji zadovoljava neku hipotezu iz G, ali ne zadovoljava njoj specifičniju hipotezu iz S, ili obrnuto. Formalno, to je primjer $\mathbf{x} \in \mathcal{X}$ za koji vrijedi

$$\exists g \in G, \exists s \in S.\ (g \geqslant_g s) \wedge \big(g(\mathbf{x}) \neq s(\mathbf{x})\big)\,.$$

Ako bismo u skup $\mathcal{D}$ pridodali primjer $\mathbf{x}$, mogli bismo iz prostora inačica ukloniti sve one hipoteze koje s tim primjerom nisu konzistentne, pa bi se time prostor inačica suzio. Možemo se zapitati: ako bi klasifikacijski algoritam mogao sam birati primjere za učenje (odnosno ako bi mogao generirati upite), koji bi primjer bilo najbolje odabrati u ovakvoj situaciji? Budući da klasifikacija tog primjera još nije poznata, valja pretpostaviti da ona s jednakom vjerojatnošću može biti i pozitivna i negativna. Algoritam bi onda trebao odabrati primjer koji će zadovoljavati polovica hipoteza iz prostora inačica, tj. takav primjer $\mathbf{x} \in \mathcal{X}$ za koji vrijedi

$$\sum_{h \in VS_{\mathcal{H},\mathcal{D}}} h(\mathbf{x}) = \left\lceil \frac{1}{2} |\, VS_{\mathcal{H},\mathcal{D}} | \right\rceil .$$

Takvom strategijom algoritmu će trebati $\mathcal{O}(\log_2 |\, VS_{\mathcal{H},\mathcal{D}}|)$ primjera da potpuno nauči klasu $\mathcal{C}$. Na ovoj se ideji temelji tehnika **aktivnog učenja** (engl. *active learning*): algoritam postavlja upite samo za one primjere koji mu trebaju da bi suzio prostor inačica. To može znatno ubrzati i pojeftiniti izgradnju klasifikatora (osobito onda kada se primjeri označavaju ručno).

2 Vapnik-Chervonenkisova dimenzija

Modeli nisu jednakog kapaciteta: neki modeli su fleksibilniji i mogu se bolje prilagoditi podacima, a neki su manje fleksibilni. Razmatranja ovog tipa u domeni su **statističke teorije učenja** (engl. *statistical learning theory*) odnosno **računalne teorije učenja** (engl. *computational learning theory, COLT*). Jedan način iskazivanja kapaciteta modela jest Vapnik-Chervonenkisova dimenzija.

Vapnik-Chervonenkisova dimenzija (VC-dimenzija) iskazuje kapacitet modela $\mathcal{H}$ u smislu broja primjera za klasifikaciju s kojim se model $\mathcal{H}$ može uspješno nositi.

Pretpostavimo da skup podataka za učenje sadrži N primjera (odnosno točaka u ulaznom prostoru). Proizvoljno odaberimo jednu konfiguraciju tih točaka u prostoru. Ako se ograničimo na samo jednu klasu, svaku od tih N točaka moguće je označiti kao pozitivnu ili negativnu. Dakle postoji 2^N mogućih označavanja, odnosno 2^N mogućih problema. Ako za svako označavanje možemo pronaći hipotezu $h \in \mathcal{H}$ takvu da h razdvaja pozitivne primjere od negativnih (tj. da je h konzistentna s $\mathcal{D}$), kažemo da $\mathcal{H}$ **razdjeljuje** (engl. *shatters*) N točaka. Formalno:

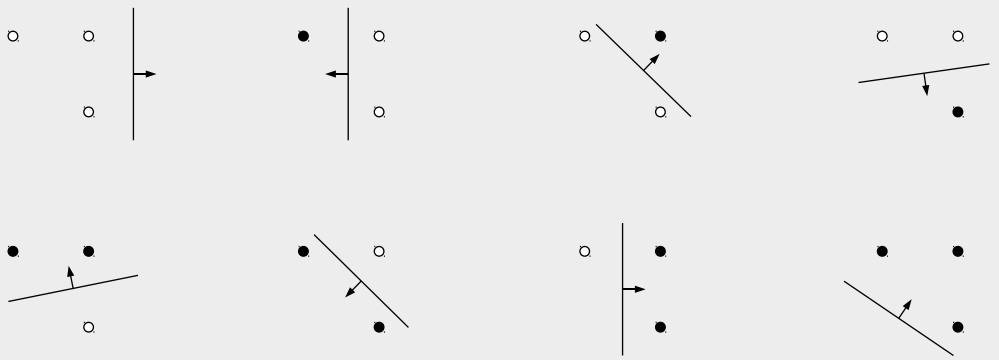
Definicija 5 (Razdjeljivanje primjera) *Neka je funkcija $y : \mathcal{X} \to \{0, 1\}$ funkcija koja primjerima iz $\mathcal{X}$ dodjeljuje oznake klase. Model $\mathcal{H}$ razdjeljuje N primjera akko*

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

Definicija 6 (VC-dimenzija) *VC-dimenzija modela $\mathcal{H}$, označena kao $VC(\mathcal{H})$, jest najveći broj primjera koje model $\mathcal{H}$ može razdijeliti.*

Primijetite da je odabir N primjera u ulaznom prostoru proizvoljan, no jednom kada je on fiksiran, razdvajanje mora biti moguće za svih 2^N označavanja. VC-dimenzija ne ovisi o konkretnom skupu za učenje; to je teoretska mjera složenosti klase hipoteze, a ne složenosti ulaznog skupa podataka. Također valja primijetiti da je moguće $VC(\mathcal{H}) = \infty$.

Primjer 1 (VC-dimenzija pravca) Neka je ulazni prostor $\mathcal{X} = \mathbb{R}^2$ te neka je model $\mathcal{H}$ skup pravaca. Za tri nekolinearne točke svako od 2^3 označavanja moguće je razdvojiti pravcem:



pa zaključujemo $VC(\mathcal{H}) \geqslant 3$. Međutim, za četiri točke, kako god odabrane, ne može se za svako označavanje pronaći razdvajajući pravac. Naime, u potpuno povezanom grafu s četiri vrha, dva brida se uvijek sijeku. Ako se u vrhove jednog takvog brida smjeste dva

pozitivna primjera, a u vrhove drugog dva negativna primjera, onda nije moguće povući pravac koji bi razdvojio pozitivne primjere od negativnih. Stoga vrijedi $VC(\mathcal{H}) < 4$, pa zaključujemo $VC(\mathcal{H}) = 3$.

Može se pokazati da hiperavnina u prostoru $\mathbb{R}^n$ može razdijeliti najviše $n+1$ točaka, pa je dakle VC-dimenzija linearnog klasifikatora u n dimenzijskom prostoru jednaka $n+1$.

VC-dimenzija u stvari daje dosta pesimističnu ocjenu kapaciteta hipoteze. Ispada da linearni klasifikator (klasifikator za koji je $\mathcal{H}$ skup pravaca) može naučiti klasu s najviše tri primjera. Međutim, VC-dimenzija ne uzima u obzir distribuciju primjera: u praksi su slični primjeri bliži jedan drugome u ulaznome prostoru (to naravno ovisi o izboru značajki), pa sva označavanja nisu jednako vjerojatna.

VC-dimenzija je donekle povezana s **brojem parametara** modela: modeli s većom VC-dimenzijom tipično imaju više parametara koje treba optimirati, dok modeli s manjom VC-dimenzijom takvih parametara obično imaju manje. No to nije pravilo; moguće je napraviti model u kojem je više parametara stopljeno u jedan parametar, kao i model koji ima suvišne, međusobno funkcijski zavisne parametre (v. primjer 5).

Zadatak 2.1. Neka je $\mathcal{X} = \mathbb{R}^2$. Koliko iznosi $VC(\mathcal{H})$, ako je model $\mathcal{H}$:

- skup pravokutnika čije su stranice poravnate s osima?
- skup kružnica sa središtem u ishodištu?
- skup elipsa sa središtem u ishodištu?
- skup konveksnih poligona?
- look-up tablica koja pohranjuje primjere?

Napomena: Dokaz za $VC(\mathcal{H}) = N$ potrebno je napraviti u dva koraka. Najprije treba pokazati da $\mathcal{H}$ razdjeljuje N primjera (tj. da za proizvoljno odabranih N točaka za svako moguće označavanje postoji konzistentna hipoteza), a zatim da $\mathcal{H}$ ne razdjeljuje $N+1$ primjera (tj. da za svakih $N+1$ točaka postoji neko označavanje za koje ne postoji konzistentna hipoteza).

3 Induktivna pristranost

Učenje hipoteze interesantan je problem utoliko što je riječ o **loše definiranom problemu** (engl. *ill-posed problem*): primjeri za učenje $\mathcal{D}$ nisu sami po sebi dovoljni da bi se na temelju njih jednoznačno inducirala hipoteza h. Drugim riječima, iz $\mathcal{D}$ ne slijedi (deduktivno) koja će hipoteza dobro klasificirati nove primjere koji se nisu našli u $\mathcal{D}$. Svojstvo hipoteze da odredi (predvidi) klasifikaciju još neviđenih primjera naziva se **generalizacija**.

Primjer 2 (Učenje Booleove funkcije) Razmotrimo kao primjer učenje Booleove funkcije od n varijabli. Ovdje je riječ o klasifikaciji primjera $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ u dvije klase, $y \in \{0, 1\}$. (Digresija: Je li VC-dimenzija linearnog modela dovoljna da se nauči Booleova funkcija od $n = 2$ varijable?) Postoji 2^{2^n} različitih Booleovih funkcija od N varijabli. Ako $|\mathcal{D}| = N$, još uvijek postoji $2^{2^n - N}$ hipoteza koje su konzistentne s primjerima za učenje. Npr., za $n = 3$ i $N = 5$:

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

postoji 8 konzistentih hipoteza. Postavlja se pitanje: koja je od njih ispravna? Problem nije dobro definiran, pa na to pitanje nije moguće odgovoriti. Činjenica je da su sve hipoteze jednako vjerodostojne.

Očito, učenje, a tako i generalizacija, nisu mogući bez dodatnih pretpostavki. Skup naših (apriornih) pretpostavki koje omogućavaju induktivno učenje nazivamo **induktivna pristranost** (engl. *inductive bias*). Činjenicu da je učenje bez pristranosti uzaludno izražava i jedan od teorema nazvan *No Free Lunch Theorem*.[1]

Induktivnu pristranost odabiremo na način da: (1) definiramo model $\mathcal{H}$, čime je određeno koje hipoteze uopće dolaze u obzir te (2) definiramo način na koji se iz prostora $\mathcal{H}$ odabire točno jedna hipoteza h. U tom smislu razlikujemo dvije vrste induktivne pristranosti:

(1) **Pristranost ograničavanjem** ili **pristranost jezika** (engl. *restriction bias, language bias*) – odabiremo model $\mathcal{H}$ i time ograničavamo skup hipoteza koji se mogu prikazati tim modelom;

(2) **Pristranost preferencijom** ili **pristranost pretraživanja** (engl. *preference bias, search bias*) – definiramo način pretraživanja hipoteza unutar $\mathcal{H}$ i na taj način zapravo dajemo prednost jednim hipotezama u odnosu na druge.

Većina algoritama učenja kombinira obje vrste induktivne pristranosti. Npr. induktivna pristranost može biti: (1) $\mathcal{H}$ je skup pravokutnika te (2) odabiremo pravokutnik najmanje površine konzistentan s primjerima za učenje (tj. najspecifičniju hipotezu).

Premda izgleda da je pristranost preferencijom – budući da ne ograničava skup pretraživih hipoteza – bolja vrsta pristranosti, u praksi se ipak odlučujemo za neki ograničen model $\mathcal{H}$. Jednostavniji modeli imaju niz pogodnosti (v. odjeljak 6).

Induktivnu pristranost moguće je definirati i nešto formalnije. Ako znamo da hipotezu nije moguće odrediti bez induktivne pristranosti, onda induktivnu pristranost možemo shvatiti kao dodatnu informaciju koja nam omogućava da na temelju nepotpune informacije iz primjera za učenje ipak možemo zaključiti o kojoj je točno hipotezi riječ. Rečeno drugačije: induktivna pristranost je skup pretpostavki temeljem kojih klasifikacija (novog) primjera slijedi **deduktivno**. Formalno:

Definicija 7 (Induktivna pristranost) *Neka je $\mathcal{L}$ algoritam za učenje, neka je $h_{\mathcal{L}}$ hipoteza inducirana pomoću $\mathcal{L}$ na skupu primjera $\mathcal{D}$ i neka je $h_{\mathcal{L}}(\mathbf{x})$ klasifikacija primjera*

[1] http://www.no-free-lunch.org

$\mathbf{x} \in \mathcal{X}$ *temeljem te hipoteze. Induktivna pristranost od* $\mathcal{L}$ *je bilo koji skup minimalnih pretpostavki* $\mathcal{B}$ *takvih da*

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

Induktivnu pristranost tako možemo shvatiti kao vezu između induktivnog zaključivanja i deduktivnog zaključivanja – ono što je nedostajalo indukciji da bi bila dedukcija.

Primjer 3 Razmotrimo ponovno učenje Booleove funkcije, ali najprije funkcije od dvije varijable, $n = 2$. Ukupno mogućih funkcija je $2^{2^n} = 16$. Ako kao model $\mathcal{H}$ odaberemo pravac, uveli smo induktivnu pristranost ograničenjem. Naime, za $n = 2$ broj primjera je $2^n = 4$, no kako je $VC(\mathcal{H}) = 3$, to kapacitet modela $\mathcal{H}$ sigurno nije dovoljan da bi se naučila baš svaka Booleova funkcija dvije varijable. Odabirom pravca kao modela $\mathcal{H}$, ograničili smo se na 14 od 16 hipoteza odnosno Booleovih funkcija (koje hipoteze nisu u $\mathcal{H}$)?

Primjer 4 Vratimo se primjeru 2 i učenju Booleove funkcije od $n = 3$ varijable. Prostor primjera je $\mathcal{X} = \{0, 1\}^3$. Uvedimo induktivnu pristranost ograničenjem: neka je model $\mathcal{H}$ ravnina u $\mathbb{R}^3$. Je li ta induktivna pristranost dovoljna da bi se naučila zadana funkcija? Koja je veličina prostora inačica te koja je klasifikacija preostala tri primjera? Što ako se iz skupa $\mathcal{D}$ ukloni primjer $(1, 0, 1)^{\mathrm{T}}$? Kolika je tada veličina prostora inačica i kako izgleda poset $(VS_{\mathcal{H},\mathcal{D}}, \geqslant_g)$? Koja je najspecifičnija, a koja najopćenitija hipoteza? Koju bismo dodatnu pristranost mogli uvesti (pristranost ograničenjem) tako da klasifikacija novog primjera ipak slijedi deduktivno (tj. da $|VS_{\mathcal{H},\mathcal{D}}| = 1$)?

4 Problem šuma

Šum je neželjena anomalija u podacima. Mogući uzroci šuma su:

- nepreciznost pri mjerenju značajki,
- pogreške u označavanju (engl. *teacher noise*),
- postojanje skrivenih značajki (latentnih varijabli),
- nejasne granice klasa (subjektivnost).

U prisustvu šuma ne postoji jednostavna granica između pozitivnih i negativnih primjera, čak i onda kada je problem zapravo inherentno jednostavan. Zbog toga jednostavni modeli (npr. modeli s niskom VC-dimenzijom) ne mogu ostvariti $E(h|\mathcal{D}) = 0$. Problem je to što u načelu šum nije moguće razdvojiti od pravih podataka. Iznimka su pojedinačni primjeri koji po nekoj vrijednosti znatno odskaču od većine drugih primjera (engl. *outliers*).

5 Regresija

Kod regresije ciljna vrijednost y je kontinuirana, $y \in \mathbb{R}$. Na temelju primjera $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$ potrebno je naučiti nepoznatu funkciju $f : \mathcal{X} \to \mathbb{R}$ tako da, idealno, $y^{(i)} = f(\mathbf{x}^{(i)})$. Učenje funkcije možemo tumačiti kao interpolaciju između točaka $\mathbf{x}^{(i)}$, odnosno ekstrapolaciju izvan točaka $\mathbf{x}^{(i)}$. Međutim, zbog prisustva šuma, zapravo učimo funkciju $y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon$, gdje je ε slučajni šum.

Regresijom nad skupom $\mathcal{D}$ dobivamo funkciju (hipotezu) h kao aproksimaciju funkcije f. Empirijska pogreška hipoteze h na skupu za učenje $\mathcal{D}$ najjednostavnije se može definirati kao

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

Pogrešku mjerimo kao zbroj kvadratnih odstupanja predviđene vrijednosti $h(\mathbf{x})$ i stvarne vrijednosti y. Pogrešku bismo mogli iskazati i na neki drugi način (razlog zašto smo odabrali baš ovaj oblik bit će jasan kasnije). Faktor $\frac{1}{2}$ uvršten je da bi pojednostavio račun.

Sada se najprije treba odlučiti za model $\mathcal{H}$ koji će biti dovoljnog kapaciteta da minimizira empirijsku pogrešku. Izaberimo linearan model:

$$h(\mathbf{x}) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_0 = \sum_{i=1}^{n} w_ix_i + w_0 = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

gdje su w_i parametri koje treba naučiti na temelju skupa primjera $\mathcal{D}$. Budući da vrijednost $h(\mathbf{x})$ linearno ovisi o ulaznim vrijednostima $\mathbf{x}$, ovu vrstu regresije nazivamo **linearna regresija**. Pretpostavimo, zbog jednostavnosti, da je prostor primjera $\mathcal{X}$ jednodimenzijski, $\mathcal{X} = \mathbb{R}$. Tada je hipoteza h definirana jednadžbom pravca

$$h(x) = w_1x + w_0$$

dok je empirijska pogreška

$$E(h|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - (w_1x^{(i)} + w_0)\right)^2.$$

Naš je cilj pronaći hipotezu h koja minimizira empirijsku pogrešku. Budući da je empirijska pogreška dana kao zbroj kvadrata pogrešaka koje nastaju na pojedinačnim primjerima, riječ je o postupku **najmanjih kvadrata** (engl. *least squares*). Oblik hipoteze fiksiran je modelom, pa pronaći optimalnu hipotezu zapravo znači pronaći parametre w_0 i w_1 takve da $\nabla_{w_0,w_1}E(h|\mathcal{D}) = 0$.

Pronađimo najprije minimum s obzirom na parametar w_0:

$$\frac{\partial}{\partial w_0}\Big[\frac{1}{2}\sum_{i}^{N}\left(y^{(i)} - (w_1x^{(i)} + w_0)\right)^2\Big] = \tag{3}$$

$$\sum_{i}^{N}(-y^{(i)} + w_1x^{(i)} + w_0) = -\sum_{i}^{N}y^{(i)} + w_1\sum_{i}^{N}x^{(i)} + Nw_0 = 0\,.$$

Rješavanjem za w_0 i uvrštenjem $\bar{x} = \sum_i^N x^{(i)}/N$ i $\bar{y} = \sum_i^N y^{(i)}/N$ dobivamo

$$w_0 = \frac{1}{N}\Big(\sum_i^N y^{(i)} - w_1 \sum_i^N x^{(i)}\Big) = \frac{1}{N}\sum_i^N y^{(i)} - w_1 \frac{1}{N}\sum_i^N x^{(i)} = \bar{y} - w_1 \bar{x}\,. \tag{4}$$

Za w_1 dobivamo

$$\frac{\partial}{\partial w_1}\Big[\frac{1}{2}\sum_i^N \big(y^{(i)} - (w_1 x^{(i)} + w_0)\big)^2\Big] = \tag{5}$$

$$\sum_i^N -x^{(i)}(y^{(i)} - w_1 x^{(i)} - w_0) = -\sum_i^N x^{(i)} y^{(i)} + w_1 \sum_i^N (x^{(i)})^2 + w_0 \sum_i^N x^{(i)} = 0\,.$$

Uvrštenjem (4) u (5) te zamjenom $\sum_i^N x^{(i)} = N\bar{x}$ dobivamo

$$-\sum_i^N x^{(i)} y^{(i)} + w_1 \sum_i^N (x^{(i)})^2 + (\bar{y} - w_1\bar{x}) N\bar{x} = w_1\Big(\sum_i^N (x^{(i)})^2 - N\bar{x}^2\Big) + N\bar{x}\bar{y} - \sum_i^N x^{(i)} y^{(i)} = 0\,.$$

iz čega slijedi

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

U ovom slučaju dakle postoji analitičko rješenje kojim se nalazi optimalna hipoteza. Također kažemo da rješenje postoji u **zatvorenoj formi** (engl. *closed-form solution*). U slučajevima kada analitičko rješenje ne postoji, učenje ćemo morati provesti **iterativnim metodama**.

Ako je linearan model prejednostavan, empirijska pogreška bit će i nakon optimizacije prevelika. U tom slučaju možemo odabrati složeniji model, npr. polinom drugoga reda

$$h(x) = w_2 x^2 + w_1 x + w_0\,.$$

Ovo je primjer **polinomijalne regresije**, za koju također postoji rješenje u zatvorenoj formi. (Napomena: Polinomijalna regresija specifičan je slučaj tzv. poopćenog linearnog modela. Taj model zovemo linearnim jer je linearan u parametrima $\mathbf{w}$, premda nije linearan u ulazima $\mathbf{x}$).

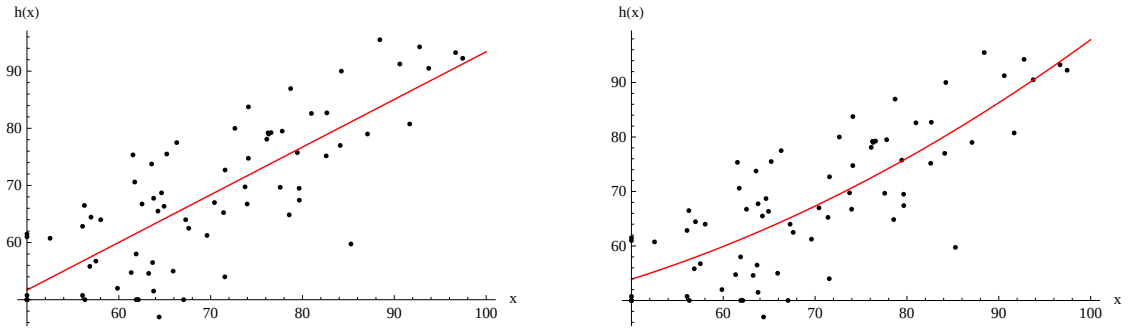
Na slici 5 prikazan je primjer linearne i polinomijalne regresije za $\mathcal{X} = \mathbb{R}$.

6 Odabir modela

Zaključili smo da je učenje bez induktivne pristranosti uzaludno te da je nužno odlučiti se za neki model $\mathcal{H}$. Taj postupak nazivamo **odabir modela** (engl. *model selection*). Budući da se odabir svodi na optimizaciju tzv. **hiperparametara** nekog fiksiranog modela, često se koristi i naziv **optimizacija modela**.

Postavlja se pitanje: kako odabrati idealan model odnosno idealnu induktivnu pristranost? Očito je da što je kapacitet modela veći, to je manja pogreška $E(h|\mathcal{D})$. No važno je imati na umu da svrha hipoteze nije da dobro klasificira primjere za učenje – klasifikacija tih primjera već nam je poznata. Želimo da hipoteza ispravno klasificira nove, buduće primjere, tj. da ima svojstvo **generalizacije**. Hipoteza je bezvrijedna ako nema takvu prediktivnu moć.

U načelu preferiramo što jednostavnije modele. Za to postoji niz razloga:



(a) (b)

Slika 5: Regresija u $\mathcal{X} = \mathbb{R}$: (a) linearna regresija i (b) polinomijalna regresija drugog stupnja.

1. Jednostavan (ali ne prejednostavan) model bolje generalizira;
2. Jednostavan model je lakše koristiti (manja računalna složenost);
3. Jednostavan model je lakše naučiti (složeniji modeli imaju više parametara koje treba optimirati);
4. Jednostavan model lakše je tumačiti te iz njega ekstrahirati znanje (npr. pravila).

Preferencija jednostavnijih modela nad složenijim modelima ima svoje uporište u filozofiji znanosti, gdje je to načelo poznato kao **Occamova britva** (engl. *Occam's razor*) ili načelo **parsimonije**.[2]

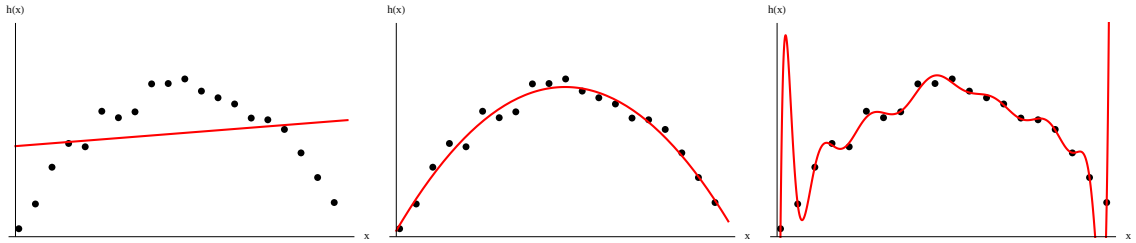
U praksi je problematično odlučiti što je jednostavno, a što ipak prejednostavno. Kako bi hipoteza što bolje generalizirala, trebamo odabrati model $\mathcal{H}$ čija složenost najbolje odgovara složenosti stvarne funkcije koju nastojimo naučiti. To zapravo znači da moramo odabrati odgovarajuću vrstu induktivne pristranosti. Ovdje susrećemo dvije krajnosti:

- **Prenaučenost** (engl. *overfitting*)[3] – Ako je model $\mathcal{H}$ previše složen (prevelikog kapaciteta) u odnosu na stvarnu funkciju, hipoteze $h \in \mathcal{H}$ previše su prilagodljive, pa podaci iz $\mathcal{D}$ nisu dovoljni da ih ograniče. Dobivamo "preraskošne" hipoteze koje pretpostavljaju više nego što postoji u stvarnim podatcima. Npr. ako model od dva pravokutnika pokušavamo naučiti na primjerima koji zapravo potječu od samo jednog pravokutnika, ili ako polinomom petog stupnja pokušamo modelirati točke koje su zapravo linearno zavisne. Pored toga, ako je model $\mathcal{H}$ previše složen, hipoteze će se prilagoditi šumu u podacima. U oba slučaja gubimo svojstvo generalizacije.

  Ako je model suviše složen, hipoteze će se vrlo varirati u ovisnosti o skupu primjera za učenje $\mathcal{D}$ – male promjene u skupu $\mathcal{D}$ dovest će do velikih oscilacija u hipotezi. Zbog toga kažemo da složeni modeli imaju visoku **varijancu**. Modeli s visokom varijancom skloni su prenaučenosti.

[2] William od Ockhama (1288–1348), engleski franjevac i skolastičar. Njegovo načelo ontološke parsimonije iskazuje izreka *"Entia non sunt multiplicanda sine necessitate"* (entitete ne treba umnožavati bez potrebe), a u suvremenoj inačici: *"Keep it short and simple"* (KISS).

[3] Također: *pretreniranost*.



(a) (b) (c)

Slika 6: Regresija funkcije $f(x) = -x^2 + \varepsilon$: (a) podnaučenost (linearna regresija), (b) optimalan model (regresija polinomom drugog stupnja), (c) prenaučenost (regresija polinomom 15. stupnja).

- **Podnaučenost** (engl. *underfitting*)[4] – Ako je model $\mathcal{H}$ prejednostavan (premalog kapaciteta) u odnosu na stvarnu funkciju, hipoteza se ne može dovoljno prilagoditi podatcima, pa onda loše opisuje i podatke iz samog skupa za učenje $\mathcal{D}$. Ako hipoteza ne može ispravno klasificirati podatke za učenje, izgledno je da će još lošije klasificirati nove primjere, tj. takva će hipoteza loše generalizirati.

  Jednostavan model ima malu varijancu, budući da je rigidniji. S druge strane, u jednostavan model ugrađeno je više pretpostavki, stoga kažemo da jednostavan model ima veću **pristranost**. (Ovdje se misli na pristranost u statističkome smislu: koliko izlaz klasifikatora odstupa od ciljne vrijednosti. Što je induktivna pristranost *bolja*, tj. što model više odgovara podatcima, to je statistička pristranost *manja*.) Modeli s velikom pristranošću skloni su podnaučenosti.

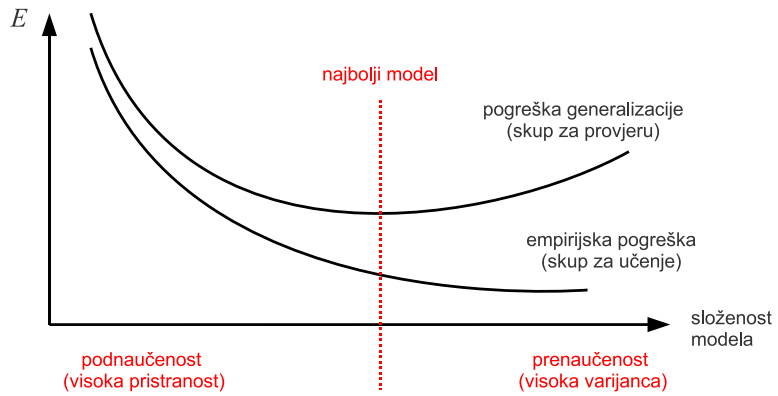
Pojavu podnaučenosti i prenaučenosti kod regresije ilustrira slika 6. Podatci su generirani funkcijom $f(x) = -x^2$, uz dodatak slučajnog šuma. Podnaučeni model ne uspijeva modelirati nelinearnost, dok prenaučen model modelira oscilira te ne uspijeva pratiti opći trend. U ovom slučaju optimalan model je polinom drugog stupnja.

Model $\mathcal{H}$ ne smije dakle biti previše složen, kako bi se izbjegla prenaučenost, ali opet ne smije biti suviše jednostavan, kako ne bi bio podnaučen. Odabir modela često se formulira kao **dvojba između pristranosti i varijance** (engl. *bias-variance dilemma*): optimalan model je onaj koji minimizira i pristranost i varijancu, i tako ostvaruje najbolju generalizaciju. Ovo je iskazano tzv. **pretpostavkom 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.*

Pretpostavka iskazuje da je generalizacija moguća, pod trima uvjetima: (1) da nije došlo do podnaučenosti, (2) da nije došlo do prenaučenosti te – što se često zaboravlja – (3) da su novi primjeri slični onima na kojima je model bio učen. Pretpostavka je dakle da su primjeri za učenje iz iste distribucije kao i primjeri na kojima će se klasifikator koristiti, tj. da su primjeri za učenje reprezentativni za problem koji rješavamo.

[4] Također: *podtreniranost*.



Slika 7: Empirijska pogreška i pogreška generalizacije u ovisnosti o složenosti modela.

6.1 Unakrsna provjera

Postavlja se pitanje: kako utvrditi je li model prenaučen ili podnaučen? Jednostavan način da se to kvantitativno ocijeni jest **unakrsna provjera**[5] (engl. *cross-validation*). Kod unakrsne provjere skup podataka razdvajmo na dva djela: **skup za učenje** (engl. *training set*) i **skup za provjeru** (engl. *validation set*). Model učimo na skupu za učenje, a njegovu generalizacijsku sposobnost provjeravamo na skupu za provjeru. Budući da klasifikator nije učen na primjerima iz skupa za provjeru, na ovaj način možemo vrlo dobro procijeniti kako će se klasifikator ponašati na neviđenim primjerima. Što bolje hipoteza klasificira primjere iz skupa za provjeru, to je bolja njezina generalizacijska sposobnost. Pogreška hipoteze mjerena na skupu koji nije korišten za učenje (engl. *off-training-set error*) naziva se **pogreška generalizacije**. Ako se unakrsna provjera ne koristi za odabir modela, već za utvrđivanje konačne pogreške već odabranog modela, onda se skup na kojemu se mjeri pogreška generalizacije naziva **ispitni skup** (engl. *test set*).

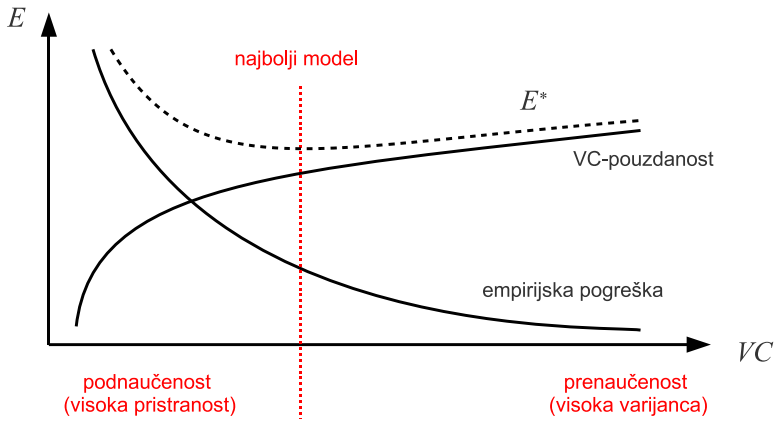
Skup za učenje, skup za provjeru i ispitni skup trebaju biti međusobno disjunktni. Naime, skup za učenje ne smije sadržavati iste primjere kao i skup za provjeru ili ispitni skup jer u protivnom ne možemo odrediti pogrešku generalizacije. Također, ispitni skup ne smije sadržavati iste primjere kao i skup za provjeru jer je skup za provjeru efektivno korišten za izgradnju modela, pa izmjerena pogreška generalizacije opet ne bi bila realna.

Tipično ponašanje empirijske pogreške (mjerene na skupu za učenje) odnosno pogreške generalizacije (mjerene na skupu za provjeru) u ovisnosti o složenosti modela prikazano je slikom 7. Za prenaučenost (visoku varijancu) je tipično da je greška na skupu za provjeru znatno veća od greške na skupu za učenje. Kod podnaučenosti (visoke pristranosti) greške su podjednako loše na oba skupa. Razumijevanje ovih odnosa bitno je pri dijagnosticiranju rada klasifikatora odnosno regresijskog postupka.

6.2 Drugi načini odabira modela

Unakrsna provjera najčešće je korišten način odabira modela. Ako međutim skup za provjeru nije dostupan (npr. skup za učenje je premalen da bi se razdijelio), model treba odabrati na neki drugi način. Neke od mogućnosti, temeljene na računalnoj teoriji učenja, jesu:

[5] Također: *unakrsna validacija*, *križna validacija*.



Slika 8: Procjena pogreške generalizacije i odabir modela temeljem načela SRMVC.

- Načelo minimizacije strukturnog rizika (engl. *Structural risk minimization*, SRM);
- Akaikeov informacijski kriterij (engl. *Akaike information criterion*, AIC);
- Bayesov informacijski kriterij (engl. *Bayesian information criterion*, BIC);
- Načelo minimalnog opisa (engl. *Minimum description length*, MDL).

Ilustrirajmo ukratko pristup temeljen na načelu minimizacije strukturnog rizika pomoću VC-dimenzije (engl. *structural risk minimization with VC-dimension*, SRMVC). Pretpostavimo da se moramo odlučiti između modela $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$. Modele ćemo poredati tako da $VC(\mathcal{H}_1) \leqslant VC(H_2) \leqslant \cdots \leqslant VC(\mathcal{H}_n)$. Označimo sa E^* očekivanu pogrešku generalizacije. Važan rezultat is statističke teorije učenja jest da je pogreška E^* s vjerojatnošću $1-\eta$ takva da vrijedi

$$E^*(h) \leqslant E(h|D) + \underbrace{\sqrt{\frac{VC(\mathcal{H})\big(\log(2N/\,VC(\mathcal{H})) + 1\big) - \log(\eta/4)}{N}}}_{\text{VC-pouzdanost}}.$$

Vrijednost E^* je zapravo procjena gornje granice pogreške (engl. *upper bound error estimate*) izračunata na temelju empirijske pogreške $E(h|D)$ i tzv. VC-pouzdanosti (pribrojnik čija vrijednost ovisi o VC-dimenziji modela $VC(\mathcal{H})$ i broju primjera za učenje N). S porastom VC-dimenzije, empirijska pogreška $E(h|D)$ će se smanjivati, ali će VC-pouzdanost rasti (slika 8). Optimalan model je onaj kod kojeg je zbroj ta dva pribrojnika minimalan.

7 Komponente algoritma nadziranog učenja

Navedimo na jednom mjestu što je sve uključeno u postupak nadziranog učenja. Klasifikacijski problem definiran je (doduše nepotpuno) skupom primjera za učenje, $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$. Primjeri su nezavisno i identično distribuirani (iid), što znači da poredak primjera nije bitan i da su primjeri uzorkovani iz iste zajedničke distribucije $P(\mathbf{x}, y)$. Naš je cilj pronaći hipotezu h koja što bolje aproksimira vrijednosti $y^{(i)}$. Pritom moramo definirati sljedeće tri komponente:

1. **Model** ili prostor hipoteza. Definicija modela kao skupa hipoteza $\mathcal{H}$ nije dovoljno operativna. U praksi je lakše ako model predočimo kao funkciju $h(\cdot|\theta)$ definiranu do na neke parametre θ, a hipotezu kao jednu konkretnu instancu te funkcije. Model je dakle skup funkcija h koje su parametrizirane s θ:

$$\mathcal{H} = \{h(\mathbf{x}|\theta)\}_\theta$$

dok je hipoteza jedna konkretna funkcija h s fiksiranim vrijednostima parametra θ.

Primjer 5 Razmotrimo neke primjere modela. Model pravokutnika u $\mathcal{X} = \mathbb{R}^2$ možemo definirati kao

$$h(x_1, x_2|\theta_{x1}, \theta_{y1}, \theta_{x2}, \theta_{y2}) = \mathbf{1}\{(\theta_{x1} \leqslant x_1 \leqslant \theta_{x2}) \wedge (\theta_{y1} \leqslant x_2 \leqslant \theta_{y2})\}$$

a jedna konkretna hipoteza mogla bi biti $h(x_1, x_2|0, 2, 1, 8)$.

Neki drugi primjeri modela:

- dvodimenzijski linearan model: $h(x_1, x_2|\theta_0, \theta_1, \theta_2) = \mathbf{1}\{\theta_1 x_1 + \theta_2 x_2 + \theta_0 > 0\}$
- kružnica u ravnini: $h(x_1, x_2|\theta) = \mathbf{1}\{x_1^2 + x_2^2 - \theta < 0\}$
- n-dimenzijski linearan model: $h(\mathbf{x}|\boldsymbol{\theta}, \theta) = \mathbf{1}\{\boldsymbol{\theta}^\mathrm{T}\mathbf{x} + \theta > 0\}$
- hiperkugla: $h(\mathbf{x}|\theta) = \mathbf{1}\{\mathbf{x}^\mathrm{T}\mathbf{x} - \theta < 0\}$
- linearna regresija: $h(\mathbf{x}|\boldsymbol{\theta}, \theta_0) = \boldsymbol{\theta}^\mathrm{T}\mathbf{x} + \theta_0$

Zanimljivo je usporediti modele $\mathbf{1}\{\theta_1 x_1 + \theta_2 x_2 + 1 > 0\}$ i $\mathbf{1}\{\theta_1 x_1 + \theta_2 x_2 + \theta_0 > 0\}$. Premda ovi modeli imaju različit broj parametara, oni su ekvivalentni u smislu da odgovaraju identičnim skupovima hipoteza. Naime, za zadani pravac, parametar θ_0 funkcijski zavisi od parametara θ_1 i θ_2. Iz ovoga je vidljivo da veći broj parametara ne mora nužno značiti i veću VC-dimenziju modela.

2. **Funkcija gubitka** (engl. *loss function*) L, koja za dane parametre modela θ izračunava razliku između ciljne vrijednosti $y^{(i)}$i njezine aproksimacije $h(\mathbf{x}^{(i)}|\theta)$. Kod regresije, funkcija gubitka tipično je definirana kao kvadratno odstupanje:

$$L\big(y^{(i)}, h(\mathbf{x}^{(i)}|\theta)\big) = (h(\mathbf{x}^{(i)}|\theta) - y^{(i)})^2\,.$$

Kod klasifikacije, funkcija gubitka tipično je definirana kao

$$L\big(y^{(i)}, h(\mathbf{x}^{(i)}|\theta)\big) = \mathbf{1}\{h(\mathbf{x}^{(i)}|\theta) \neq y^{(i)}\} = |h(\mathbf{x}^{(i)}|\theta) - y^{(i)}|\,. \tag{6}$$

Empirijska pogreška definira se kao očekivanje funkcije gubitka nad primjerima iz skupa za učenje, uz pretpostavku uniforme distribucije primjera (tj. $P(\mathbf{x}) = 1/N$):

$$E(\theta|\mathcal{D}) = \mathbb{E}_{\mathcal{D},\theta}[L] = \frac{1}{N}\sum_{i=1}^{N} L\big(y^{(i)}, h(\mathbf{x}^{(i)}|\theta)\big)\,. \tag{7}$$

Faktor $1/N$ u (7) u načelu je nebitan jer ne utječe na to za koje vrijednosti θ funkcija E doseže minimum (kod regresije se tipično koristi faktor $1/2$ umjesto $1/N$).

U nekim slučajevima pogreške klasifikacije nemaju jednaku težinu, tj. gubitci su asimetrični. Npr., kod klasifikacije zloćudnog tumora lažno pozitivne slučajeve preferiramo nad lažno negativnim slučajevima, dok je kod filtriranja neželjenih poruka ili ocjene kreditne sposobnosti situacija upravo obrnuta. U takvim slučajevima funkcija gubitka definira se pomoću **matrice gubitka** (engl. *loss matrix*). Matrica gubitka $L = [L_{kj}]$ definira gubitak L_{kj} koji nastaje uslijed klasifikacije primjera $\mathbf{x}$, koji zapravo pripada klasi $\mathcal{C}_k$, u klasu $\mathcal{C}_j$ (vrijedi $L_{kk} = 0$). Poseban slučaj je **matrica gubitka nula-jedan** (engl. *zero-one loss matrix*), za koju $L_{kj} = \mathbf{1}\{k \neq j\}$, i koja odgovara gubitku definiranom sa (6).

Primjer 6 (Matrica gubitka) Matrica gubitka za klasifikaciju zloćudnog tumora mogla bi biti definirana kao

$$L = \begin{array}{c} \\ \text{rak} \\ \neg\,\text{rak} \end{array} \begin{array}{c} \begin{array}{cc} \text{rak} & \neg\,\text{rak} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

gdje retci predstavljaju stvarne klase, a stupci odabrane klase. U ovom slučaju lažno negativna klasifikacija nanosi 1000 puta veći gubitak od lažno pozitivne klasifikacije.

3. **Optimizacijski postupak** kojim nalazimo vrijednosti θ^* za koje je empirijska pogreška najmanja, tj.

$$\theta^* = \operatorname*{argmin}_{\theta} E(\theta|\mathcal{D})$$

gdje argmin daje vrijednost argumenta koji minimizira funkciju. U nekim slučajevima (npr. kod linearnog modela regresije) minimum je moguće odrediti analitički, odnosno rješenje je u zatvorenoj formi. Kada to nije moguće, potrebno je koristiti iterativne metode optimizacije. Učenje se dakle svodi na optimizacijski problem, koji se može rješavati raznim numeričkim optimizacijskim postupcima. (Ovu optimizaciju ne treba miješati s optimizacijom složenosti modela, koja je zapravo optimizacija optimizacije, tj. *metaoptimizacija*.)

Algoritmi strojnog učenja međusobno se dakle razlikuju po (1) modelu, (2) funkciji gubitka i (3) optimizacijskom postupku. Ta tri odabira definiraju ujedno i induktivnu pristranost algoritma.

8 Pristupi nadziranom učenju

8.1 Generativni i diskriminativni modeli

Osnovna podjela klasifikacijskih postupaka jest na generativne i diskriminativne modele. Razlika se svodi na to kako modeliramo pripadnost primjera klasi. Analogna podjela vrijedi i za modele regresije.

Generativni modeli. Ovi modeli pretpostavljaju da je vjerojatnost da primjer $\mathbf{x}$ pripada klasi $\mathcal{C}_j$ proporcionalna zajedničkoj vjerojatnosti primjera $\mathbf{x}$ i klase $\mathcal{C}_j$, tj.:

$$P(\mathcal{C}_j|\mathbf{x}) \propto P(\mathbf{x}, \mathcal{C}_j)\,.$$

Generativni modeli modeliraju zajedničku vjerojatnost $P(\mathbf{x}, \mathcal{C}_j)$. Temeljem te vjerojatnosti može se, primjenom Bayesovog pravila, izračunati posteriorna vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$, tj. vjerojatnost da primjer $\mathbf{x}$ pripada klasi $\mathcal{C}_j$. Ovakav pristup, koji modelira zajedničku razdiobu primjera $\mathbf{x}$ i klasa $\mathcal{C}_j$, nazivamo generativnim jer modelira postupak generiranja podataka. Model se također može koristiti za generiranje sintetičkih primjera u ulaznome prostoru, uzorkovanjem iz zajedničke distribucije $P(\mathbf{x}, \mathcal{C}_j)$.

Tipični generativni modeli su Bayesov klasifikator, mješavina Gaussovih distribucija, latentna Dirichletova alokacija, Bayesove mreže i skriveni Markovljev model (HMM).

Diskriminativni modeli. Diskriminativni modeli ne modeliraju zajedničku vjerojatnost $P(\mathbf{x}, \mathcal{C}_j)$, već izravno modeliraju pripadnost primjera $\mathbf{x}$ klasi $\mathcal{C}_j$. Diskriminativni modeli mogu biti probabilistički ili neprobabilistički. Probabilistički diskriminativni modeli izravno modeliraju posteriornu vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$. Neprobabilistički diskriminativni modeli modeliraju funkciju $h(\mathbf{x})$ koja primjeru $\mathbf{x}$ izravno dodjeljuje oznaku klase $\mathcal{C}_j$. Ovaj pristup nije probabilistički jer ne barata s vjerojatnostima.

Logistička regresija tipičan je primjer probabilističkog diskriminativnog modela. Većina diskriminativnih modela nisu probabilistički, međutim u mnogo slučajeva izlaz modela može se koristiti kao indikacija pouzdanost klasifikacijske. Tipični primjeri neprobabilističkih diskriminativnih modela jesu perceptron, višeslojni perceptron, stroj s potpornim vektorima (SVM), stabla odluke, k-najbližih susjeda i linearna diskriminantna analiza (LDA).

Generativni modeli ponekad se u literaturi nazivaju **zajednički modeli** (engl. *joint models*), dok se probabilistički diskriminativni modeli nazivaju i **uvjetni modeli** (engl. *conditional models*). Diskriminativni modeli (probabilistički i neprobabilistički) ponekad se nazivaju **metode temeljene na granici** (engl. *boundary-oriented methods*). Neprobabilistički diskriminativni modeli nazivaju se i **diskriminacijske funkcije**.

Kod probabilističkih pristupa klasifikacija se odvija kroz dvije faze: fazu **zaključivanja** i fazu **odlučivanja**. U fazi zaključivanja (odnosno učenja)[6] koristimo primjere za učenje kako bismo izračunali $P(\mathcal{C}_j|\mathbf{x})$ (kod generativnih modela na jedan, a kod diskriminativnih na drugi način). U fazi odlučivanja koristimo posteriornu vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$ kako bismo klasificirali nove primjere. Kod neprobabilističkih diskriminativnih modela zaključivanje i odlučivanje stopljeni su u jedan korak.

Generativni modeli imaju niz pogodnosti:

- Ocjena pouzdanosti klasifikacije – izlaz klasifikatora može se tumačiti kao vjerojatnost ili pouzdanost da primjer $\mathbf{x}$ pripada klasi $\mathcal{C}_j$;
- Interpretabilnost rezultata – generativni modeli nude vrlo intuitivnu interpretaciju podataka temeljenu na teoriji vjerojatnosti;

[6] *Zaključivanje* (engl. *inference*) u statističkome smislu odnosi se na izgradnju modela koji opisuje podatke, što je postupak koji je, u kontekstu strojnog učenja, istovjetan učenju.

- Ugradnja pozadinskog znanja – u generativne modele lako je ugraditi pozadinsko znanje stručnjaka i takvo znanje kombinirati sa znanjem dobivenim na temelju podataka;
- Odbijanje klasifikacije – ako je za neki primjer $\mathbf{x}$ izlaz klasifikatora manji od unaprijed zadanog praga, klasifikator može odbiti klasificirati primjer $\mathbf{x}$ i tako smanjiti broj pogrešnih klasifikacija. (Primjeri koje klasifikator odbije klasificirati mogu se proslijediti na ručnu klasifikaciju.);
- Nalaženje vrijednosti koje odskaču – marginalizacijom vjerojatnosti $P(\mathbf{x}, \mathcal{C}_j)$ možemo odrediti vjerojatnost primjera $P(\mathbf{x})$ i tako detektirati vrijednosti koje odskaču (engl. *outliers*);
- Minimizacija rizika – u slučajevima kada pogreške klasifikacije nemaju jednaku težinu (tj. kada matrica gubitka $[L_{kj}]$ nije tipa nula-jedan), probabilistički model može donositi optimalne odluke u smislu minimizacije rizika.

Probabilistički generativni modeli imaju naravno i neke nedostatke u odnosu na diskriminativne modele. Glavni nedostatci su:

- Broj primjera – modeliranje zajedničke vjerojatnosti $P(\mathbf{x}, \mathcal{C}_j)$ iziskuje velik broj primjera, a da bi procjena bila pouzdana. To je osobit problem kada je ulazni prostor visoke dimenzije.
- Nepotrebna složenost modeliranja – ako je naš cilj klasifikacija, a ne generiranje primjera, onda je nepotrebno modelirati zajedničku vjerojatnost $P(\mathbf{x}, \mathcal{C}_j)$, koja može biti nepotrebno složena. U tom slučaju dovoljno je izravno modelirati samo posteriornu vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$, kao što to čine diskriminativni modeli.

8.2 Parametarski i neparametarski modeli

Nezavisno od gornje podjele, s obzirom na odnos između broja primjera za učenje i broja parametara modela, nadzirane postupke možemo podijeliti na parametarske i neparametarske modele.

Parametarski modeli. Kod parametarskih modela složenost modela ne ovisi o broju primjera za učenje. Konkretno, probabilistički parametarski postupci pretpostavljaju da se podaci pokoravaju nekoj teorijskoj razdiobi (npr. Gaussovoj razdiobi). Učenje se svodi na nalaženje parametara pretpostavljene distribucije, broj kojih ne ovisi o broju primjera.

Neparametarski modeli. Kod neparametarskih modela broj parametara, a time i složenost modela, raste s brojem primjera za učenje. Ovdje ne pretpostavljamo da se podatci pokoravaju nekoj teoretskoj distribuciji. Treba napomenuti da neparametarski modeli (nazivu unatoč) imaju parametre, ali da to nisu parametri neke pretpostavljene distribucije.

Generativni modeli, poput Bayesovog klasifikatora i skrivenog Markovljevog modela, jesu parametarski modeli jer se kod tih postupaka učenje svodi na optimizaciju fiksnog broja parametara neke pretpostavljene distribucije. Linearna i polinomijalna regresija su parametarski modeli budući da pretpostavljaju oblik funkcije koju aproksimiraju i da

Tablica 1: Podjela nadziranih pristupa klasifikaciji

	Generativni	**Diskriminativni**
Parametarski	• Bayesov klasifikator • Bayesove mreže • latentna Dirichletova alokacija • skriven Markovljev model	• logistička regresija • perceptron • stroj s potpornim vektorima (linearan) • linearna diskriminantna analiza
Neparametarski		• višeslojni perceptron • k-najbližih susjeda • stabla odluke • klasifikacijska pravila

je broj parametra unaprijed zadan. Diskriminativni modeli, poput perceptrona, stroja s potpornim vektorima (SVM) i linearne diskriminantne analize također su parametarski modeli budući da imaju fiksiran broj parametara i da je složenost modela fiksna. S druge strane, višeslojni perceptron, stabla odluke, Parzenovi prozori (regresija) i algoritam k-najbližih-susjeda tipični su primjeri neparametarskih modela koji ne pretpostavljaju nikakvu distribuciju primjera i kod kojih broj parametara odnosno složenost modela raste s porastom brojem primjera za učenje. Ovi su odnosi za klasifikacijske postupke sažeto prikazani u tablici 1.

Parametarski modeli očito imaju jače pretpostavke o podatcima. Ako su te pretpostavke točne (npr. ako koristimo Gaussovu razdiobu primjera, a primjeri se doista pokoravaju toj razdiobi), onda su u pravilu parametarski modeli bolji od neparametarskih. Međutim, ako se stvarni podatci ne pokoravaju pretpostavljenoj teorijskoj razdiobi, pogreška klasifikacije bit će razmjerno velika.

8.3 Linearni i nelinearni modeli

Konačno, nadzirane modele možemo podijeliti s obzirom na granicu kojom ti modeli u ulaznom prostoru razdvajaju pozitivne primjere od negativnih (u slučaju klasifikacije), odnosno s obzirom na krivulju kojom aproksimiraju funkciju (u slučaju regresije).

Linearni modeli. Povlače linearnu granicu između primjera dviju klasa, odnosno funkciju aproksimiraju linearnim modelom. U dvodimenzijskom prostoru granica je pravac, a u trodimenzijskom granica je ravnina. Općenito, za ulazni prostor $\mathcal{X}$ dimenzije n, granica je $(n-1)$-dimenzijska hiperravnina.

Primjeri linearnih modela su naivan (diskretan) Bayesov klasifikator, logistička regresija, perceptron, stroj s potpornim vektorima (SVM), linearna diskriminantna analiza i linearna regresija.

Nelinearni modeli. Povlače nelinearnu granicu između primjera dviju klasa, odnosno funkciju aproksimiraju nelinarnom hiperravninom.

Primjeri nelinearnih modela su algoritam k-najbližih susjeda, stabla odluke, višeslojni

perceptron, SVM s jezgrenim funkcijama (kao i sve druge metode proširene nelinearnim jezgrenim funkcijama) i polinomijalna regresija.

Očito je da nelinearni modeli imaju veći kapacitet (veću VC-dimenziju) i da su zbog toga sposobni rješiti klasifikacijske probleme koji nisu rješivi linearnim modelima (tipičan primjer je problem *isključivo-ili*). U praksi većina interesantnih klasifikacijskih problema nije linearno razdvojiva. To je osobito izraženo kada je broj primjera znatno veći od dimenzije ulaznog prostora, $N \gg n$, budući da je tada ulazni prostor vrlo gusto naseljen. Suprotno, ako primjera nije mnogo više nego što je dimenzija ulaznog prostora, ulazni prostor neće biti gusto naseljen i veća je vjerojatnost da je problem linearno razdvojiv. Ova činjenica predstavlja motivaciju za jezgrene metode (engl. *kernel methods*), kod kojih se linearna razdvojivost ostvaruje povećanjem dimenzije prostora primjera.