Prva domaća zadaća iz strojnog učenja

1. Riješite primjer 4 iz prve bilješke za predavanje. Rješenje, pored ostalog, treba uključivati skicu prostora primjera $\mathcal{X}$ i skicu parcijalnog uređaja hipoteza iz $\mathcal{H}$.

Učenje Booleove funkcije od n varijabli, uz neke poznate primjere:

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

Tablica 1

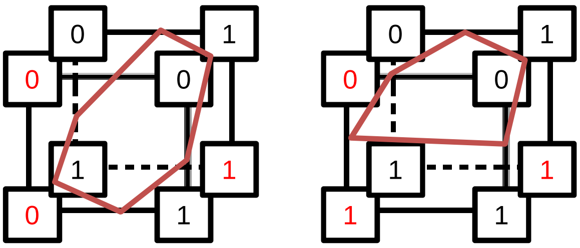
n = 3
N = 5

$$2^{2^n - N} = 8$$

Ako pogledamo primjere u Tablici 1, vidimo da imamo 3 primjera koja ne znamo kako klasificirati. Bez ikakvih drugih podataka, postoji 8 konzistentnih hipoteza za tu funkciju.

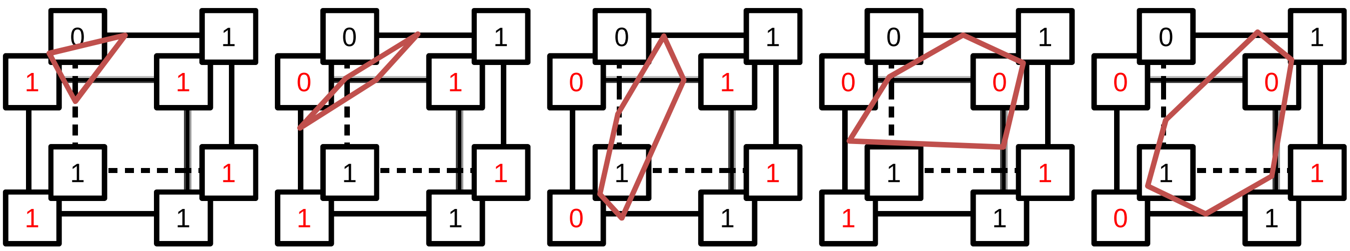
Nakon što uvedemo induktivnu pristranost ograničenjem, da je model $\mathcal{H}$ ravnina u R^3, snizili smo si broj konzistentnih hipoteza na dvije, kao što se vidi na Slici 1.

Jasno se vidi da induktivna pristranost nije dovoljna da bi se naučila zadana funkcija. Veličina prostora inačica je $|VS_{\mathcal{H},\mathcal{D}}| = 2$. Klasifikacije primjera su sljedeće:
$(0\ 0\ 0)^T$ -- 0 ili 1
$(0\ 0\ 1)^T$ -- 0
$(1\ 1\ 0)^T$ -- 1



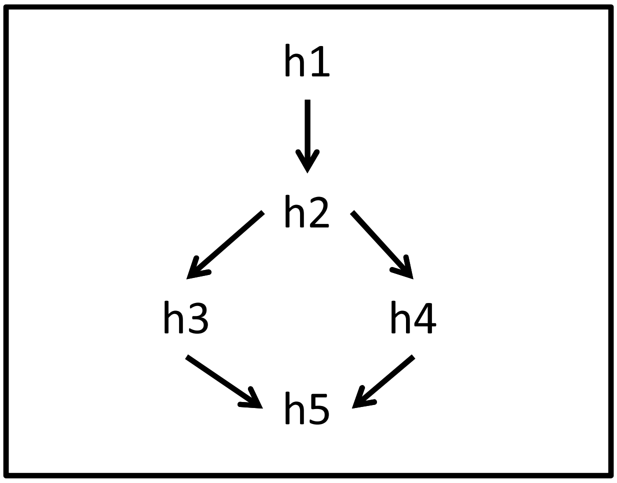
Slika 1. Klasifikacija uz pomoć modela $\mathcal{H}$

Ako se iz skupa $\mathcal{D}$ ukloni primjer $(1\ 0\ 1)^T$, onda imamo još jednu točku koju trebamo klasificirati. Veličina prostora inačica će se povećati i iznosit će $|VS_{\mathcal{H},\mathcal{D}}| = 5$.



Slika 2. Prostor primjera i hipoteze, s lijeva nadesno: h1, h2, h3, h4, h5

Najspecifičnija je hipoteza h5, a najopćenitija hipoteza h1.



Slika 3. Parcijalni uređaj hipoteza iz $\mathcal{H}$

Da bismo smanjili prostor inačica mogli bismo uvesti dodatnu pristranost (pristranost pretraživanjem), recimo, iz modela $\mathcal{H}$ odabirali bismo ravnine konzistentne s primjerima za učenje koje imaju najmanju površinu presjeka sa kockom koju razapinju primjeri, tako bismo dobili hipotezu h1.

2. U prostoru primjera $\mathcal{X} = \mathbb{Z}^2$ razmatramo dva modela: $\mathcal{H}_1$ (kružnice s poizvoljno odabranim ishodištem) i $\mathcal{H}_2$ (pravokutnici sa stranicama poravnatima s koordinatnim osima).

   (a) Formalno definirajte $\mathcal{H}_1$ i $\mathcal{H}_2$.

   (b) Vrijedi li $H_1 \cap \mathcal{H}_2 = \emptyset$? Obrazložite odgovor.

   (c) Odredite $VC(\mathcal{H}_1)$ i $VC(H_2)$.

   (d) Odredite koje su moguće vrijednosti za $VC(\mathcal{H}_1 \cup \mathcal{H}_2)$ te obrazložite odgovor.

   (e) Identificirajte dvije najspecifičnije, ali međusobno neusporedive hipoteze iz $\mathcal{H}_1 \cup \mathcal{H}_2$.

a) $\mathcal{H}_1$: $h(x_1, x_2 \mid \theta_{Sx}, \theta_{Sy}, \theta_r) = \mathbf{1}\{(x_1 - \theta_{Sx})^2 + (x_2 - \theta_{Sy})^2 - \theta_r{}^2 \leq 0\}$

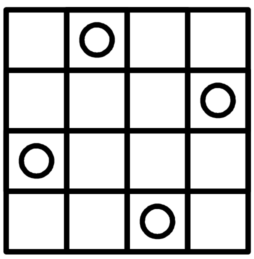
$\mathcal{H}_2$: $h(x_1, x_2 \mid \theta_{x1}, \theta_{y1}, \theta_{x2}, \theta_{y2}) = \mathbf{1}\{(\theta_{x1} \leq x_1 \leq \theta_{x2}) \land (\theta_{y1} \leq x_2 \leq \theta_{y2})\}$

b) Ako bismo $\mathcal{H}_1$ i $\mathcal{H}_2$ promatrali kao skupove kružnica i skupove pravokutnika, onda bi bilo logično da je presjek $\mathcal{H}_1$ i $\mathcal{H}_2$ prazan skup jer ne postoji kružnica koja je ujedno i pravokutnik.

Ali, ako $\mathcal{H}_1$ i $\mathcal{H}_2$ promatramo kao skup hipoteza koje određenom primjeru dodjeljuju 0 ili 1, onda možemo reći da presjek $\mathcal{H}_1$ i $\mathcal{H}_2$ nije neprazan skup, jer postoje primjeri koje će hipoteze iz $\mathcal{H}_1$ i $\mathcal{H}_2$ jednako klasificirati (trivijalan slučaj – samo jedna pozitivna točka).

c) Ako pogledamo raspored od tri točke posložene poput vrhova jednakostraničnog trokuta, možemo pomoću kružnica odvojiti sve kombinacije pozitivnih i negativnih primjera. Ali ako pogledamo raspored od četiri točke, ne možemo kružnicama riješiti XOR problem. Dakle, VC($\mathcal{H}_1$) = 3.

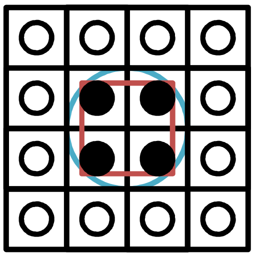
Pogledajmo sad raspored od četiri točke, kao što je prikazan na Slici 4, vidjet ćemo da pomoću pravokutnika poravnatih sa osima možemo odvojiti sve kombinacije. Ali ako povećamo broj točaka na pet, ne možemo više odvojiti sve kombinacije pomoću pravokutnika. Dakle, VC($\mathcal{H}_2$) = 4.



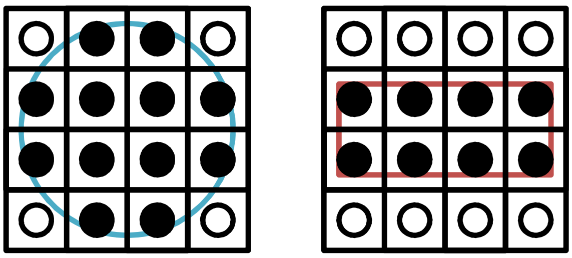
Slika 4. Raspored točaka koji dokazuje da je VC($\mathcal{H}_2$) = 4

d) Da ponovimo, VC($\mathcal{H}_1$) = 3, a VC($\mathcal{H}_2$) = 4. Unija $\mathcal{H}_1$ i $\mathcal{H}_2$ sadrži sve hipoteze iz oba modela. Budući da mi prilikom određivanjaVC dimenzije odabiremo hipotezu koja nam odgovara, možemo za neki slučaj uzimati hipoteze iz $\mathcal{H}_2$, tako da nam je opet u najgorem slučaju VC($\mathcal{H}_1$ U $\mathcal{H}_2$) jednaka VC($\mathcal{H}_2$) = 4. Nisam uspio naći slučaj sa 5 točaka koje bi ($\mathcal{H}_1$ U $\mathcal{H}_2$) mogla razdvojiti, pa tvrdim da je VC($\mathcal{H}_1$ U $\mathcal{H}_2$) = 4.

e) Meni nije jasno tražite li dvije najspecifičnije hipoteze za isti primjer ili za dva različita? Evo vam, na slici 5 dvije hipoteze za isti primjer, s tim da bih rekao da su te dvije hipoteze usporedive jer odjeljuju jednake primjere. A na slici 6 su dvije hipoteze za dva različita primjera, koji su sigurno neusporedivi.



Slika 5. Dvije hipoteze, isti primjer, h1 – kružnica, h2 - pravokutnik



Slika 6. Dvije hipoteze, različiti primjeri, h1 – kružnica, h2 - pravokutnik

3. Na skupu $\mathcal{D}$ od $N = 400$ primjera naučen je linearan klasifikator. Svaki primjer $x^{(i)}$ sastoji se od $n = 10$ značajki. Greška na skupu za učenje je 10%.

(a) Kolika je VC-dimenzija ovog klasifikatora?

Linearni klasifikator je hiperravnina u prostoru R^n, a takva hiperravnina može razdijeliti najviše n+1 točaka. U našem slučaju, imamo R^{10}, dakle, naš linearni klasifikator ima VC dimenziju 11.

$$VC(\mathcal{H}) = 11$$

(b) Izračunajte gornju granicu pogreške klasifikatora uz pouzdanost 95%.

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

$$E^*(h) \leqslant 0.1 + \sqrt{\frac{11 * (\log(2*400/11) + 1) - \log(0.05/4)}{400}}$$

$$E^*(h) \leqslant 0.1 + 0.289 = 0.389$$

(c) Na istom skupu naknadno je isprobano 10 različitih linearnih klasifikatora $(h_1, h_2, \ldots, h_{10})$. Modeli se međuobno razlikuju po broju značajki koje koriste: model h_i koristi samo prvih i značajki. Eksperimentalno su na skupu za učenje dobiveni ovi rezultati:

Klasifikator	Greška (%)
h_1	28.00
h_2	28.00
h_3	28.00
h_4	28.75
h_5	30.25
h_6	30.75
h_7	18.25
h_8	11.75
h_9	11.50
h_{10}	10.00

Korištenjem načela minimizacije strukturnog rizika uz VC-dimenziju (SRMVC) odaberite najbolji klasifikator.

Klasifikator	$E^*(h)$
h_1	0.431
h_2	0.455
h_3	0.474
h_4	0.499
h_5	0.530
h_6	0.549
h_7	0.437
h_8	0.384
h_9	0.393
h_{10}	0.389

Tablica 2

Načelo minimizacije strukturnog rizika pomoću VC-dimenzije (SRMVC): Korištenjem formule za procjenu gornje granice pogreške E*(h), najmanja gornja granica pogreške se dobije za h_8, što nas dovodi do zaključka da je za ovaj problem najbolji linearni klasifikator h_8, klasifikator koji gleda samo prvih 8 značajki.

(d) Je li u ovom slučaju opravdano korištenje načela minimizacije strukturnog rizika za pronalazak najboljeg klasifikatora umjesto npr. metode unakrsne provjere?

Imamo 400 primjera, a moja intuicija mi govori da je to malen broj za učiti linearni klasifikator, dakle ne bih htio smanjiti broj primjera za učenje time što bih neke primjere odvojio za unakrsnu provjeru, smatram da je opravdano korištenje načela minimizacije strukturnog rizika za pronalazak najboljeg klasifikatora.

4. Odabrali smo model $\mathcal{H}$ koji ima hiperparametar α kojim se može ugađati složenost modela. Za odabrani α naučili smo hipotezu koja minimizira empirijsku pogrešku. Unakrsnom provjerom ustanovili smo da je pogreška generalizacije znatno veća od empirijske pogreške. Je li naš odabir parametra α optimalan? Obrazložite odgovor.

Unakrsnom provjerom smo dobili da je pogreška generalizacije znantno veća od empirijske pogreške. To znači da nam je model prenaučen (presložen), jer lošije klasificira primjere na kojima nije učio. Naravno, naš odabir hiperparametra α nije optimalan.

Što bi trebali napraviti da poboljšamo model?

Primjere bi trebali podijeliti na tri skupa, skup za učenje, skup za provjeru i testni skup. Pomoću unakrsne provjere na skupu za provjeru trebali bi odrediti hiperparametar α koji će minimizirati pogrešku modela, te bi zatim pomoću unakrsne provjere na testnom skupu [nadamo se] vidjeli da je pogreška generalizacije smanjena.

5. b) Za računanje pogreški koristio sam sljedeću formulu:

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

Za h1 i h2 dobivam empirijsku grešku oko 8.8 i pogrešku generalizacije oko 9.6, a za h3 emprijiska greška je oko 5.8 te pogreška generalizacije oko 6.6. To bi značilo da modeli h1,h2 i h3 dobro generaliziraju [pogreška generalizacije nije višestruko puta veća], i da složeniji model h3 bolje predviđa rezultate.

c) Za h1' i h2' sam dobio empirijsku grešku oko 5.5 i pogrešku generalizacije oko 17, a za h3' empirijska greška je oko 3 te pogreška generalizacije oko 15. Kad se uspoređuje s h1, h2 i h3, vidimo da crtane hipoteze imaju manju empirijsku grešku i veći pogrešku generalizacije, što bi značilo da su si američki auti sličniji, odnosno više se razlikuju od europskih i japanskih.

d) U oba slučaja sam dobio da su auti s najvećim odstupanjem od modela:

46.6	4	86.00	65.00	2110.	17.9	80	3	"mazda glc“
44.3	4	90.00	48.00	2085.	21.7	80	2	"vw rabbit c (diesel)“
43.4	4	90.00	48.00	2335.	23.7	80	2	"vw dasher (diesel)“

Ne razumijem se u aute, ali mislim da oni najviše odstupaju jer imaju jako visoku prvu vrijednost u odnosu na druge aute (mislim da su ovo 3 auta sa najvećim mpg vrijednostima).

e) h1'' i h2'' imaju empirijsku grešku oko 7.5 i pogrešku generalizacije oko 8.5, h3'' ima empirijsku pogrešku oko 3.5 i pogrešku generalizacije oko 5. Kod mene je kvadratni model bolji za sve tri hipoteze.

6. (a) U zadatku 5 koristili ste linearni regresijski model. Svaki algoritam strojnog učenja sastoji se se od tri osnovne komponente. Identificirajte i objasnite te komponente na slučaju linearnog regresijskog modela iz zadatka 5.

Tri komponente od kojih se sastoji svaki algoritam strojnog učenja su: **Model**, **Funkcija gubitka**, **Optimizacijski postupak**.

Model – kod nas broj i odabir značajki koje ćemo koristiti u linearnom klasifikatoru.

Optimizacijski postupak – kod nas funkcija regress, s kojom smo učili modele h1, h2, h3, h1', h2', h3', h1'', h2'', h3''

Funkvija gubitka – kvadratno odstupanje dobivenih izlaza od očekivanih izlaza .

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

```
% izračunaj koeficijente linearne regresije
k = regress(mpg, [ones(size(X,1),1) X]);
```

(b) Objasnite koja je induktivna pristranost tog modela i koje je vrste.

Ovdje imamo induktivnu pristranost ograničenjem, odredili smo da će nam modeli biti linearni klasifikatori.

(c) Je li linearni regresijski model koji ste koristili u zadatku 5 parametarski ili neparametarski pristup strojnom učenju? Obrazložite odgovor.

Linearni regresijski model je parametarski, broj parametara ne ovisi o broju primjera, nego o odabiru modela.

(d) Obrazložite u kojim sitacijama preferiramo koristiti matricu gubitka koja nije tipa nula-jedan. Izmislite neki primjer u kojem bi takva matrica gubitka bila od koristi.

Matricu gubitka koja nije tipa nula-jedan, preferiramo koristiti kada nam nije svejedno je li pogreška klasifikatora bila lažno pozitivan slučaj ili lažno negativan. Evo DVA primjera:
Ako radimo klasifikaciju emailova na spam i nespam, bilo bi jako štetno da nam se nespam mail klasificira kao spam, tj. lažno pozitivni slučajevi su gori.
Ako radimo klasifikaciju medicinskih kartona, tko ima a tko nema rak, puno gore je da nam klasifikator kaže da osoba koja zapravo ima rak da nema rak, tj. lažno negativni slučajevi su gori.
Spam u inboxu nije jednako loše kao nespam u junk folderu[možemo ručno obrisati spam], i lažna pozitivna dijagnoza raka nije jednako loša kao lažno negativna dijagnoza raka [ako je netko lažno pozitivan, ići će na dodatne pretrage, lažno negativan će umrijeti].