STROJNO UČENJE

1. Domaća Zadaća

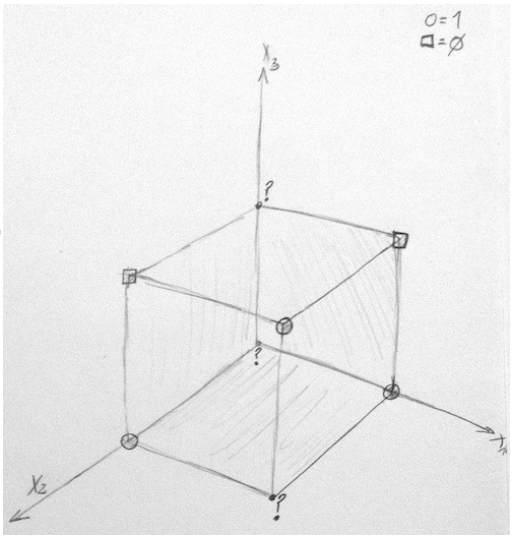
Krešimir Špes
0036419866
ak. god. 2011. / 2012.

---

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 $\mathcal{H}$.

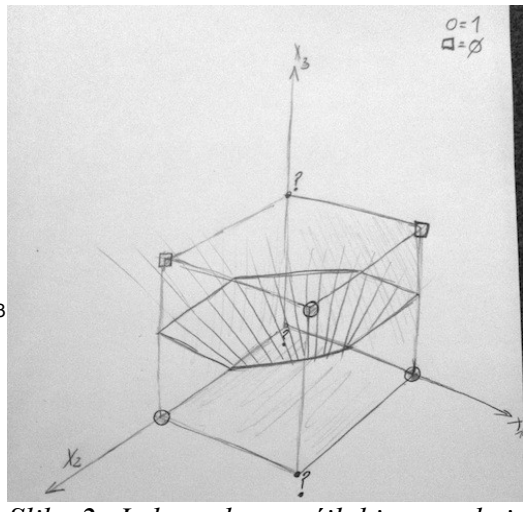
Primjer 4 Vratimo se primjeru 2 i učenju Booleove funkcije od $n = 3$ varijable. Prostor primjera je $\mathcal{X} = \{0,1\}^3$. Uvedimo induktivnu pristranost ograničenjem: neka je model $\mathcal{H}$ ravnina u $\mathbb{R}^3$. Je li ta induktivna pristranost dovoljna da bi se naučila zadana funkcija? 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 $|VS_{\mathcal{H},\mathcal{D}}| = 1$)?

- Zadana induktivna pristranost ograničenjem nije dovoljna da se nauči funkcija jer ovisno o orijentaciji ravnine, vrijednost funkcije u točki (0,0,0) nije jednoznačna.
- Ovisno kako gledamo na |VS| u ovom primjeru, mogli bi reći da je |VS| beskonačno, jer je ravnina zadana u R^3 ali gledano s druge strane od svih takvih ravnina imamo 2 skupine ravnina koje se razlikuju samo u klasifikaciji točke (0,0,0). Pa pošto je funkcija zadana u prostoru $\{0,1\}^3$ a ne u R^3 možemo reći da je **|VS| = 2**. Klasifikacija primjera [0 0 1] je jednoznačna i iznosi -1, klasifikacija [1 1 0] je +1 a ovisno o hipotezi klasifikacija [0 0 0] može biti +1 ili -1.

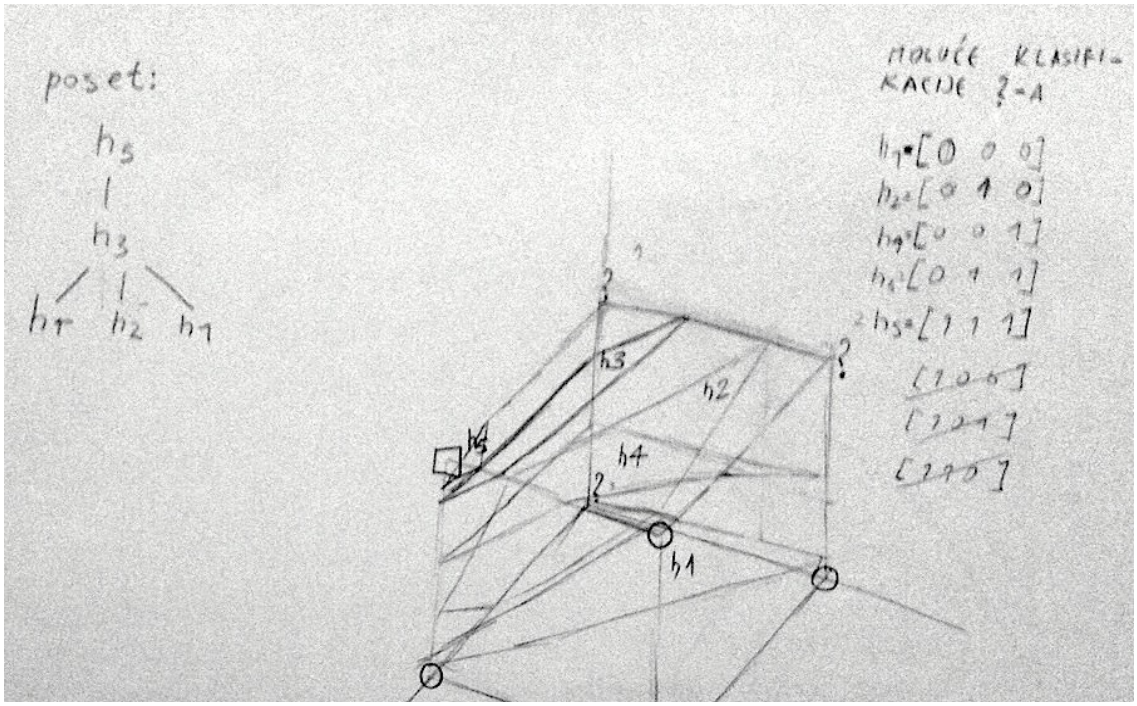


Slika 1: vizualizacija prostora

- Ako se ukloni primjer [1 0 1] onda je prostor inačica veličine **5.** Vizualizacija takvog prostora, svih hipoteza i poset-a je dana na slici na sljedećoj stranici.
- Najspecifičnija hipoteza je hipoteza h_1 koja klasificira primjere [1 0 0], [0 1 0] i [1 1 1] pozitnivno a ostale negativno.
- Najopćenitija hipoteza je h_5 na slici, to je ona koja samo primjer [0 1 1] klasificira negativno a ostale pozitivno.
- Mogli bi smo uvesti takvu pristranost da kažemo da uzimamo sve ravnine koje sijeku x_3 os na pozitivnom dijelu te osi.



Slika 2: Jedna od mogućih hipoteza koja negativno klasificira primjer (0,0,0)



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 $\mathcal{H}_2$.

$h_{kružnice}(x_1, x_2 \mid \theta_x, \theta_y, \theta_r) = 1 \{ (x_1 - \theta_{x1})^2 + (x_2 - \theta_y)^2 \leq \theta_r^2 \}$

u kružnici parametri θ_x, θ_y predstavljaju XY koordinate centra kružnice a θ_r radijus.

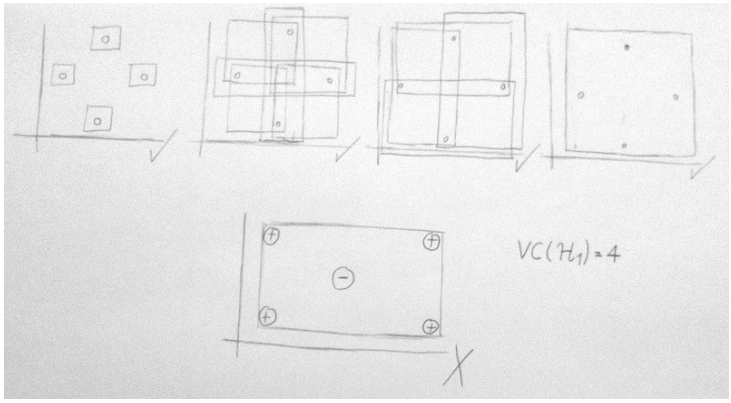
$h_{pravokutnici}(x_1, x_2 \mid \theta_{x1}, \theta_{y1}, \theta_{x2}, \theta_{y2}) = 1 \{ (\theta_{x1} \leq x_1 \leq \theta_{x2}) \wedge (\theta_{y1} \leq x_2 \leq \theta_{y2}) \}$

u pravokutniku θ_{x1}, θ_{y1} predstavljaju gornju lijevu točku pravokutnika a θ_{x2}, θ_{y2} donju desnu.

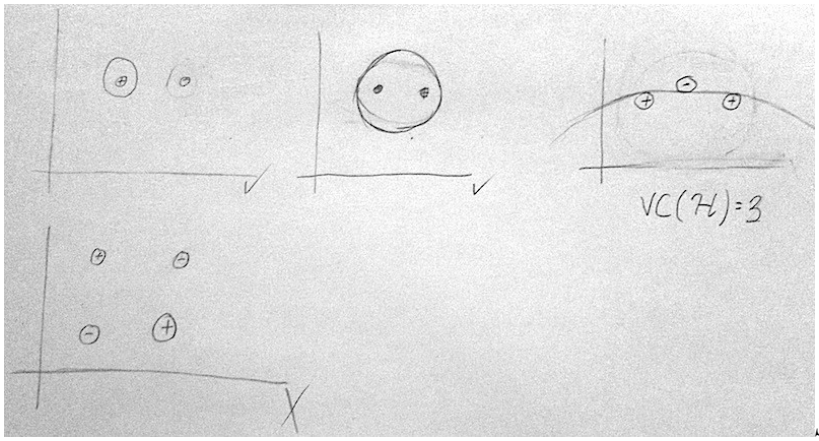
(b) Vrijedi li $H_1 \cap \mathcal{H}_2 = \varnothing$? Obrazložite odgovor.

Vrijedi. Zato što su elementi iz H_1 i H_2 međusobno neusporedivi. Po definiciji presjeka skupova, u presjek ulaze oni elementi koji su zajednički u oba skupa. A pošto se elementi iz ta 2 skupa ne mogu uspoređivati, presjek je prazan skup. Unija npr. ne bi bila prazan skup.

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



Slika 4: Pravokutnik

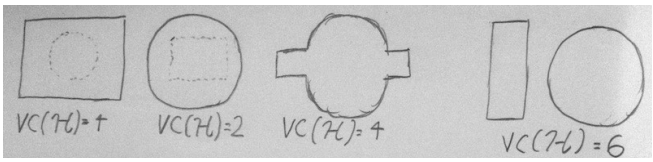


Slika 5: Krug

VC dimenzija pravokutnika sa stranicama poravnatim s koordinatnim osima jest **4** što sam nastojao ilustrirati na papiru. Prve 4 slike pokazuju da za sve kombinacije 4 pozitivnih i negativnih primjera možemo pronaći hipotezu konzistentnu s primjerima. Dok za 5 primjera dana je jedna kombinacija primjera koji se nemogu obuhvatiti pravokutnikom što dokazuje da VC dimenzija nemože biti veća od 4.

VC dimenzija za krug za proizvoljnim centrom je **3** što se može vidjeti iz gornje ilustracije.

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



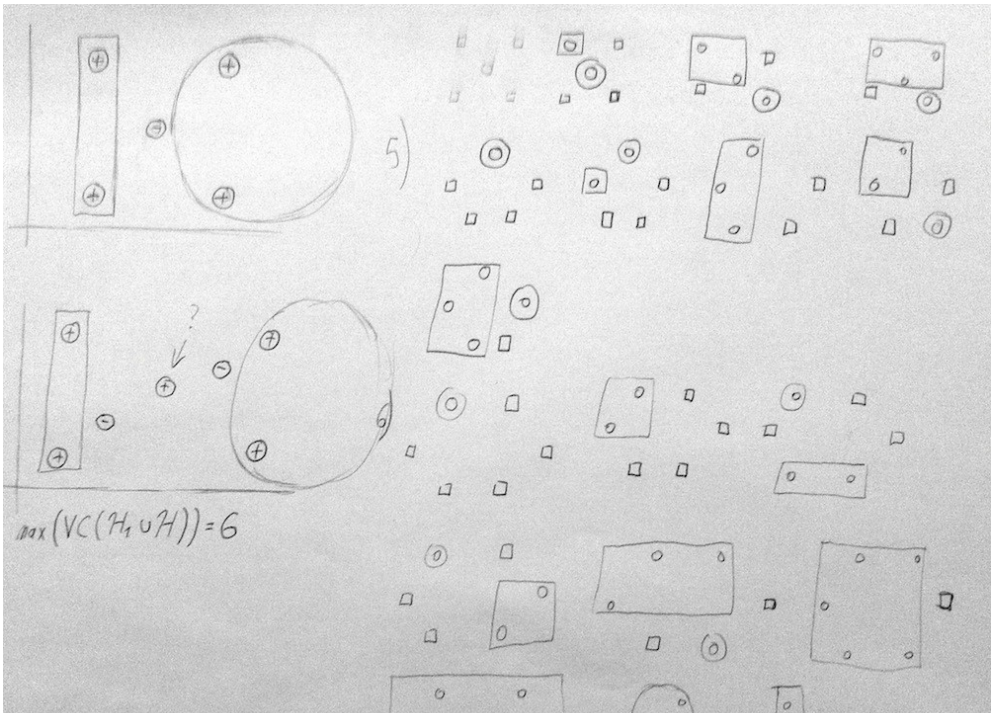
Ovisno o tome kako je izvedena unija identificiramo 4 slučaja ilustriranih na slici.

- Za prvi slučaj u kojem je krug unutar pravokutnika, VC dimenzija je jednaka VC dimenziji pravokutnika jer krug ne sudjeluje u modelu.
- Analogno za drugi slučaj VC dimenzija je jednaka VC dimenziji kruga (ispravak na slici piše 2 umjesto 3)
- Za treći slučaj VC dimenzija je ista kao i VC dimenzija pravokutnika jer ne uspijeva

razlomiti istu kombinaciju od 5 točaka kao i primjer dan u prethodnoj ilustraciji VC dimenzije za pravokutnik

- Dok za zadnji slučaj možemo vidjeti da je VC dimnenzija pravokutnika i kruga koji se ne dodiruju jednaka 6. Mogao bi se izvući zaključak da je takva VC dimenzija jednaka zbroju VC dimenzija pojedine komponente ali to nije uvijek slučaj. Na primjer, VC dimenzija unije kružnice u ishodištu i kvadrata u ishodištu je i dalje 1.

Na sljedećoj ilustraciji možemo vidjeti reprezentativne primjere kombinacije pozitivnih i negativnih uzoraka za 5,6 i 7 uzoraka:



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

Hipoteza sa Pravokutnikom površine > 0 i krugom površine 0 te Hipoteza s pravokutnikom površine 0 i krugom površine > 0.

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?

VC dimenzija linearnog klasifikatora je n + 1 gdje je n dimenzionalnost decizijske hiperravnine. Dimenzionalnost ovog klasifikatora je 10, stoga VC dimenzija mu je **11**.

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

E*(h) = 49.5398241139%

(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:

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

Računanjem E* (h) dobivamo sljedeće rezultate:

Hipoteza	**E(h | D)**	**E* (h)**
h_1	28.00	49.4271765107
h_2	28.00	52.5662496892
h_3	28.00	55.1915869806
h_4	28,75	58.2279136435
h_5	30,25	61.7694940415
h_6	30,75	64.1239934998
h_7	18,25	53.3296907493
h_8	**11,75**	**48.4135073976**
h_9	11,5	49.6452136517
h_{10}	10	49.5398241139

SMRVC medotom smo odredili da je h_8 najbolji klasifikator.

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

Nije, jer imamo dovoljno primjera za odabir klasifikatora metodom unakrsne provjere.

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.

Nije, odabrali smo model koji je složeniji od stvarne distribucije uzoraka koja se pokušava pronaći. Dogodila se prenaučenost.

Zato što se model previše prilagodio primjerima, moć generalizacije se znatno smanjila.

5. Ovaj zadatak izvodite u okružju Matlab ili Octave. Opis učitavanja i obrade potadaka u tim sustavima dan je u Dodatku.

(b) U okružuju Matlab ili Octave potrebno je načiniti linearni regresijski model nad više varijabli.[2] Učitajte podatke i podijelite ih na skup za učenje (slučajnih 249 zapisa) i skup za provjeru (preostalih 149 zapisa). Najprije naučite model (h_1) koji predviđa energetsku učinkovitost vozila na temelju snage i težine vozila. Zatim naučite model (h_2) koji koristi tri varijable – snagu, težinu i ubrzanje. Treći model (h_3) neka koristi broj cilindara, zapreminu motora, snagu, težinu, ubrzanje i godinu modela. Za svaki od tri modela izračunajte empirijsku pogrešku i pogrešku generalizacije. Komentirajte rezultate. U izvještaju navedite izračunate pogreške, ispis kôda i komentare.

Korištena je formula za empirijsku grešku iz 1. bilješke uz dodatak da je broj podjeljen sa brojem primjera kako bi pogreške bile uprosječene i time usporedive. U sljedećoj tablici dan je primjer izvođenja matlab koda ispod. Pošto je odabir elemenata skupa nasumičan, rezultati malo variraju ovisno o pokretanju programa.

h_x	h_1	h_2	h_3
Empirijska pogreška	8.3096	8.3003	5.6486
Pogreška generalizacije	32.8789	32.9156	25.9725

Iz tih rezultata možemo vidjeti da se h2 jako malo razlikuje od h1 što upućuje na to da značajka ubrzanja ne utječe skoro nikako na potrošnju goriva, dok h3 koristi više značajki i daje potpuniju sliku, tj. Bolje procjenjuje neviđene primjere.

MATLAB Kod za 5b zadatak:

```
% Ucitaj podatke u jedan string.

url = 'file:///Users/kreso/Dropbox/FER/SU/dz1/zad5/input.txt'; % prebacio sam u lokalni
fajl radi brzine
tekst_podatci = urlread(url);
tmp = textscan(tekst_podatci,'%f %f %f %f %f %f %f %f %*q','TreatAsEmpty','?', ...
'CollectOutput',1); % Trazena matrica
m = tmp{1};
redak_ima_nan = isnan(m(:,4));
m(redak_ima_nan,4) = mean(m(~redak_ima_nan,4));
%disp(m) %debug

X = m;
% podjeli skup primjera na 2 skupa
[m, n] = size(X);
A = [];
mpgA = [];
for i=1:249
     index = random('unid', m);
     A(i,:) = X(index,:);
     X(index,:) = []; % izbrisi red
     m = m-1;
end
mpg = X(:,1);
mpgA = A(:,1);


% prosirujemo X sa jedinicama zbog x0 parametra hiperravnine
X = [ones(size(X,1),1) X];
A = [ones(size(A,1),1) A];
S1 = A(:,[1,5,6]); % varijable snaga, tezina, indeksi pomaknuti za 1 zbog ones-a
S2 = A(:,[1,5,6,7]); % varijable snaga, tezina, akceleracija
S3 = A(:,[1,3,4,5,6,7,8]); % varijable sve osim proizvo?a?a
X1 = X(:,[1,5,6]); % testni skupovi
X2 = X(:,[1,5,6,7]);
X3 = X(:,[1,3,4,5,6,7,8]);

% izracunaj koeficijente linearne regresije
k1 = regress(mpgA, S1);
k2 = regress(mpgA, S2);
k3 = regress(mpgA, S3);
% predvidanje modela:
r1 = S1 * k1;
r2 = S2 * k2;
r3 = S3 * k3;
test1 = X1 * k1;
test2 = X2 * k2;
test3 = X3 * k3;

[size_mpgA size2] = size(mpgA);
[size_mpg size3] = size(mpg);
[sum(abs(   r1-mpgA).*abs(   r1-mpgA))/(2*size_mpgA) % empirijska pogreska h1
 sum(abs(test1-mpg ).*abs(test1-mpg ))/(2*size_mpg )] % pogreska generalizacije h1

[sum(abs(   r2-mpgA).*abs(   r2-mpgA))/(2*size_mpgA) % empirijska pogreska h2
 sum(abs(test2-mpg ).*abs(test2-mpg ))/(2*size_mpg )] % pogreska generalizacije h2

[sum(abs(   r3-mpgA).*abs(   r3-mpgA))/(2*size_mpgA) % empirijska pogreska h3
 sum(abs(test3-mpg ).*abs(test3-mpg ))/(2*size_mpg )] % pogreska generalizacije h3
```

(c) Načinite novu podjelu podataka, i to tako da je skup za učenje sačinjen samo od zapisa o američkim automobilima, a skup za provjeru od preostalih zapisa. Na novom skupu za učenje naučite tri modela (h'_1, h'_2, h'_3) s istim varijablama kao i kod modela h_1, h_2 odnosno h_3. Izračunajte pogreške i komentirajte ih. Usporedite h_1 s h'_1, h_2 s h'_2 te h_3 s h'_3. Što se iz toga može zaključiti?

h'_x	h'_1	h'_2	h'_3
Empirijska pogreška	5.6743	5.5600	3.0756
Pogreška generalizacije	41.7614	41.7608	33.7411

Empirijske greške su proporcionalne prethodnom zadatku što je bilo i za očekvati.

No, kao što se može vidjeti, empirijska pogreška je u prosjeku manja u odnos na prethodni zadatak jer je distribucija primjera za učenje uniformnija, dok je pogreška generalizacije veća zbog toga što klasifikator nije naučen na reprezentativnim primjerima za funkciju.

Slično kao da naučite funkciju sin(x)/x na primjerima unutar prve periode, onda je za očekivati da će regresija odstupati više nego da ste naučili funkciju na šire distribuiranom skupu primjera.

U sljedećem odjeljku dan je kod za ovaj zadatak, samo onaj dio koji se odnosi na raspodjelu skupa jer je to jedina promjena u odnosu na prošli zadatak:

```
X = m;
% podjeli skup primjera na 2 skupa ovisno o mjestu proizvodnje
[m, n] = size(X);
A = [];
mpgA = [];
n=1;
index=1;
for i=1:m
     if X(index,8) == 1;
        A(n,:) = X(index,:);
        X(index,:) = []; % izbrisi red
        n = n+1;
     else
         index=index+1;
     end
end
```

(d) Za modele h_3 i h_3' pronađite tri automobila čiji podatci najviše odstupaju od modela i pokušajte objasniti zašto je tome tako.

za h_1 najviše odstupaju:

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

za h_2 najviše odstupaju:

46.6	4	86.00	65.00	2110.	17.9	80	3	"mazda glc"
43.4	4	90.00	48.00	2335.	23.7	80	2	"vw dasher (diesel)"
44.0	4	97.00	52.00	2130.	24.6	82	2	"vw pickup"

Može se pretpostaviti da ovi auti odstupaju od ostalih jer fali neka značajka koja bi bolje opisala potrošnju tih autiju i objasnila toliko odstupanje. Npr. otpor zraka. Recimo “wv pickup” je veliki auto i po logici nebi trebao imati toliko velik MPG, ali možda je dobro aerodinamično složen ili nešto drugo što omogućuje efikasniju potrošnju.

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.

1. **Odabir modela:** u 5. zadatku smo model odabrali tako da smo isprobali više različitih linearnih klasifikatora koji se međusobno razlikuju u dimenzionalnosti, tj. U tome koje kombinacije značajki uzimaju za klasifikaciju. Tim eksperimentom smo zaključili da je za taj slučaj najbolja hipoteza h_3 .
2. **Definiranje funkcije gubitka:** odlučili smo se za Mean-Square-Error funkciju gubitka te jediničnu matricu gubitka.
3. **Optimizacijski postupak:** optimizirali smo parametre decizijske hiperravnine pomoću matlabove regress funkcije

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

Induktivna pristranost ograničenjem je dana u obliku modela hipperravnine.

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

Parametarski jer postoje parametri modela koje trebamo optimizirati te složenost modela ne raste s brojem primjera za učenje.

(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.

Kada želimo minimizirati rizik. Primjeri s predavanja koje vrijedi navesti jest detekcija raka kod pacijenata di se penalizira situacija kada pacijent ima rak a sustav to ne otkrije te filtriranje spam-a gdje se penalizira označavanje emaila kao spam koji to nije.

Neki drugi primjer: klasifikacija razine pića u bocama na pokretnoj traci u industriji. Penalizira se ako se boca klasificira kao unutar dopuštene razine a to nije. Dakle bolje je imati lažno negativne klasifikacije koje će radnici na traci ručno klasificirati nego da kupac dobije bocu koja ima manje tekućine nego što je platio.