UNIZG FER, **Strojno učenje**, ak. god. 2011/12.

# Prva domaća zadaća

*Zadaću treba predati do **17.11.2011.** do **17.00 sati**.*

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 $\mathcal{H}$.

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 $H_1 \cap \mathcal{H}_2 = \varnothing$? 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 $\mathcal{H}_1 \cup \mathcal{H}_2$.

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?

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

   (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.

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

4. Odabrali smo model $\mathcal{H}$ koji ima hiperparametar $\alpha$ kojim se može ugađati složenost modela. Za odabrani $\alpha$ 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 $\alpha$ optimalan? Obrazložite odgovor.

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

   (a) Potrebno je skinuti podatke o 398 automobila s adrese:
   `http://archive.ics.uci.edu/ml/datasets/Auto+MPG`
   Ciljna varijabla jest energetska učinkovitost vozila (*miles-per-gallon*) i nju promatramo kao zavisnu od ostalih varijabli koje opisuju broj cilindara (*cylinders*), zapreminu motora (*displacement*), snagu (*horsepower*), težinu (*weight*), ubrzanje (*acceleration*), godinu modela (*model year*), podrijetlo (*origin*[1]). Također su dana i imena automobila, no njih nećemo koristiti kao varijable modela. Svaku nedostajuću vrijednost potrebno je zamijeniti sa srednjom vrijednošću ostalih redaka u dotičnome stupcu.

   (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.

   (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?

   (d) Za modele $h_3$ i $h'_3$ pronađite tri automobila čiji podatci najviše odstupaju od modela i pokušajte objasniti zašto je tome tako.

   (e) Naučite tri modela ($h''_1$, $h''_2$, $h''_3$) s istim varijablama te skupovima za učenje i provjeru koje ste koristili pri učenju modela $h_1$, $h_2$ odnosno $h_3$ iz podzadatka 5b. Neka novi modeli ($h''_1$, $h''_2$, $h''_3$) budu polinomi drugog stupnja (dakle riječ je o kvadratnoj regresiji), definirani na sljedeći način:

$$h(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \sum_{\substack{i,j \\ i \leqslant j}} \alpha_{ij} x_i x_j + \sum_i \beta_i x_i + \gamma \,.$$

   Provjerite je li u svim slučajevima linearni model ima bolju generalizaciju od kvadratnog.

[1] SAD=1, Europa=2, Japan=3.

[2] Za navedene alate postoji dobra podrška (tj. *help* stranice) u kojoj se detaljno opisuju procedure za linearno regresijsko modeliranje (engl. *multiple linear regression*).

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.

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

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

   (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.

## Dodatak: Učitavanje i obrada podataka u sustavu MATLAB

Ovaj isječak koda prikazuje jedan način učitavanja i obrade podataka u okruženju MATLAB/Octave:

```
% Učitaj podatke u jedan string.
url = 'http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data';
tekst_podatci = urlread(url);

% Funkcija textscan nije prisutna u programu octave starijem od verzije 3.4.
% U slučaju da radite sa starijom verzijom možete koristiti kombinaciju
% funkcija fopen i fscanf.
% Ovaj poziv funkcije textscan radi sljedeće:
% - pretvori string u matricu po zadanoj specifikaciji,
% - sve nepoznate vrijednosti (označene znakom ?) zamijeni vrijednosti NaN,
% - više stupaca spoji u jednu matricu.
tmp = textscan(tekst_podatci,'%f %f %f %f %f %f %f %f %*q','TreatAsEmpty','?', ...
               'CollectOutput',1);
% Tražena matrica
m = tmp{1};
```

Popunjavanje vrijednosti koje nedostaju:

```
% NaN vrijednosti nalaze se samo u četvrtom stupcu (horsepower).
redak_ima_nan = isnan(m(:,4));
m(redak_ima_nan,4) = mean(m(~redak_ima_nan,4));
```

Primjer računanja linearne regresije na temelju snage, težine i godine modela:

```
% ciljna vrijednost
mpg = m(:,1);

% varijable snaga, težina i godina modela
X = m(:,[4,5,7]);

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

% predviđanje modela:
[ones(size(X)) X] * k
```

Funkcija `regress` nalazi parametre $a_1, a_2, \ldots, a_n$ u funkciji $a_1x_1 + a_2x_2 + \cdots + a_nx_n$. Mi želimo da funkcija ima oblik $a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$, pa zato svim primjerima dodajemo jednu jedinicu kao prvu dimenziju (tj. postavljamo $x_0 = 1$ svakom primjeru).