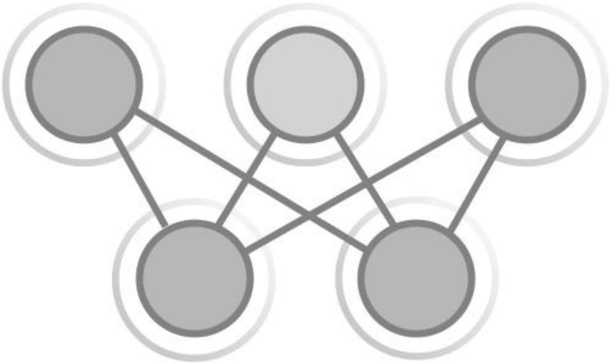
Prof.dr.sc. Bojana Dalbelo Bašić

Fakultet elektrotehnike i računarstva
Zavod za elekroniku, mikroelektroniku, računalne i inteligentne sustave

www.zemris.fer.hr/~bojana
bojana.dalbelo@fer.hr

# Evaluacija hipoteza



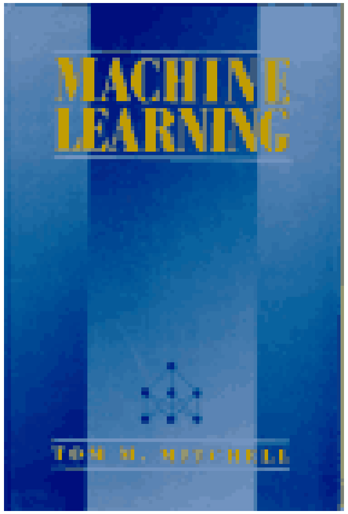


FER
Studeni 2009

# Literatura za predavanje

- *Chapter 5*

  Evaluacija hipoteza



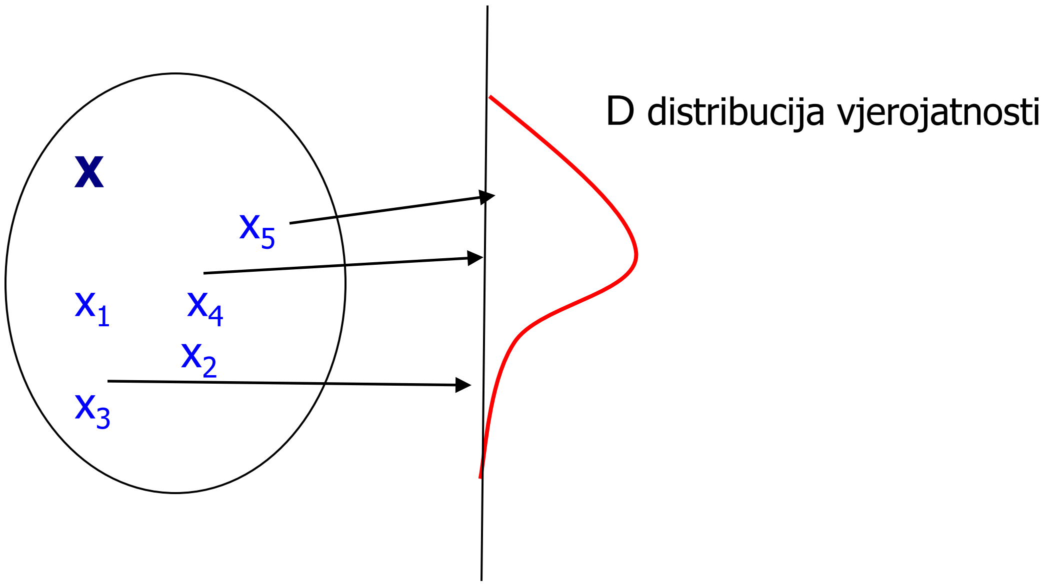
# Evaluacija hipoteza

- Postupak učenja provodi se najčešće na uzorku ograničene veličine.
  - Koliko je precizna i pouzdana naučena hipoteza?

- Važnost evaluacije performansi naučene hipoteze:
  - zbog buduće uporabe hipoteze
  - evaluacija hipoteza komponenta mnogih metoda za učenje (npr.  podrezivanje stabla odluke)

# Procjena pouzdanosti hipoteze

- Zanima nas:
  - točnost (preciznost) klasificiranja budućih primjera
  - pouzdanost te klasifikacije.
- Skup svih mogućih primjera **X** na kojem je definiran ciljni koncept f:




- Različiti primjeri imaju različitu vjerojatnost pojavljivanja.

# Evaluacija hipoteza

- Zadaća učenja: naučiti ciljnu funkciju f: X $\rightarrow$ {0, 1}, pretražujući prostor hipoteza H
- X - skup svih primjera
- **Primjeri za učenje**:
  - Elementi iz X, izabrani slučajno, u skladu s distribucijom D, kojima je pridružena vrijednost ciljnog koncepta f(x).

# Evaluacija hipoteza

- Za danu hipotezu h i skup primjera od n elemenata, izvučen slučajno iz skupa X, **koja je najbolja procjena točnosti hipoteze h** za nove primjere izvučene iz skupa X na temelju distribucije vjerojatnosti D?
-
- Kolika je **vjerojatnost pogreške u toj procjeni točnosti**?

# Pogreška uzorka i stvarna pogreška

- Proporcija (stopa) pogreške hipoteze nad primjerima za učenje – **pogreška uzorka** (*engl. sample error*).
- Proporcija (stopa) pogreške hipoteze nad cjelom populacijom s distribucijom D – **stvarna pogreška** (*engl. true error*).
- Pretpostavimo da imamo neku hipotezu h (npr. stablo odluke).
- Slučajno odaberemo n primjera iz D – to je slučajan uzorak S. Na tom uzorku odredimo stopu pogreške hipoteze h tj. proporciju krivo klasificiranih primjera ( stopa = broj krivo klasificiranih primjera/n). Označimo tu <u>stopu pogreške na uzorku</u> s **$\mathrm{error}_S$(h) i nazovimo je pogreška uzorka.**

# Pogreška uzorka

- *Definicija*
- **Pogreška uzorka** $error_S(h)$ hipoteze h u odnosu na ciljnu funkciju f i uzorak S jest

$$\mathbf{error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))} \ ,$$

gdje je n broj elemenata u uzorku S,

$\delta(f(x), h(x)) = 1$ ako $f(x) \neq h(x)$,

0 inače.

- Stvarna pogreška hipoteze **h** jest vjerojatnost da će **h** krivo klasificirati jedan slučajno izvučen primjer iz distribucije D.

# Stvarna pogreška

- Stvarna pogreška hipoteze **h** jest vjerojatnost da će **h** krivo klasificirati jedan slučajno izvučen primjer iz distribucije D.

*Definicija*

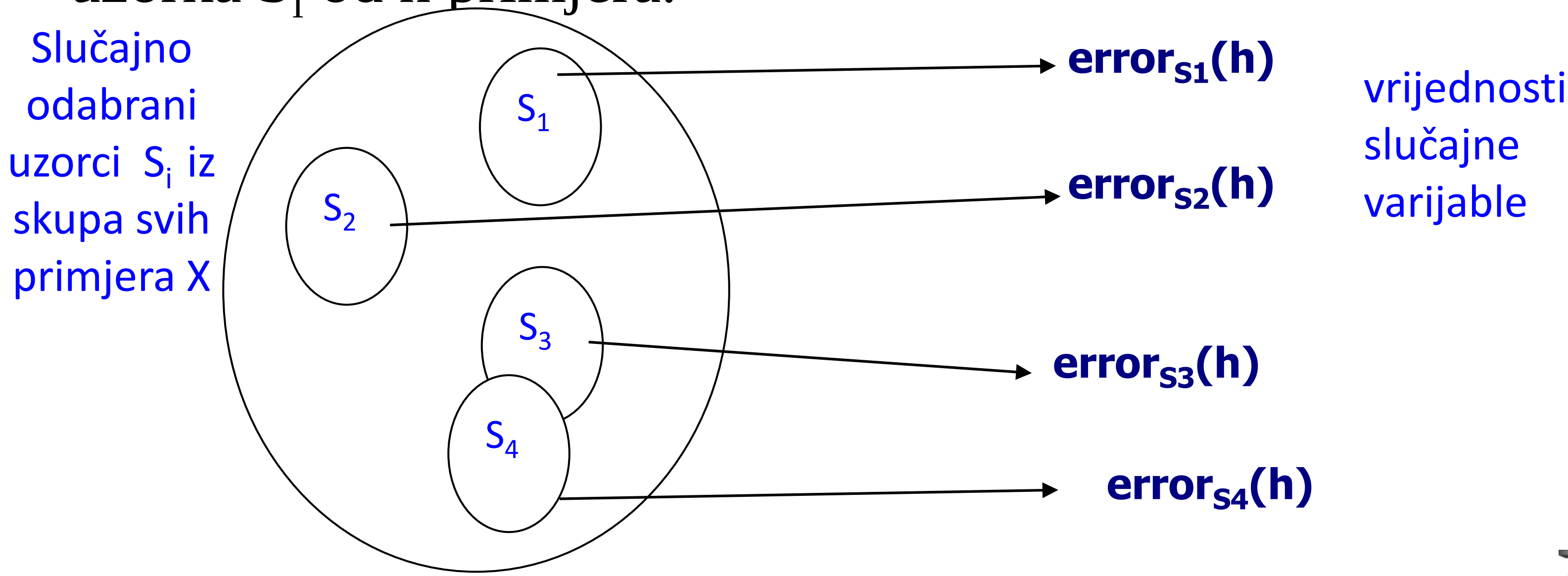
- **Stvarna pogreška** ($error_D(h)$) hipoteze h, u odnosu na ciljnu funkciju f i distribuciju D jest vjerojatnost da će h krivo klasificirati jedan slučajno izvučen primjer iz distribucije D.

$$error_{\mathbf{D}}(h) = P_{\mathbf{x \in D}}[f(x) \neq h(x)].$$

- Želimo znati stvarnu pogrešku.
- Na raspolaganju nam je samo pogreška uzorka.

# Koliko je $error_S(h)$ dobra procjena $error_D(h)$?

- Očekujemo da se pogreška uzorka **$error_{Si}(h)$** razlikuje za različite uzorke $S_i$ od n elemenata.
- Slučajni pokus je pokus čiji ishod nije jednoznačno određen uvjetima eksperimenta.
- U ovom slučaju slučajni pokus je izvlačenje slučajnog uzorka $S_i$ od n primjera.



# Koliko je $error_S(h)$ dobra procjena $error_D(h)$?

- Svakom takvom ishodu slučajnog pokusa tj. jednom uzroku $S_i$ od n elemenata pridružimo proporciju krivo klasificiranih primjera neke hipoteze h.
- **$error_{Si}(h)$** je vrijednost pridružena i-tom slučajnom pokusu – tj. **slučajna varijabla**.
- Pojednostavljeno možemo slučajnu varijablu predstaviti kao funkciju koja pridružuje neku numeričku vrijednost (u našem slučaju $error_{S3}(h)$) ishodu slučajnog pokusa (u našem primjeru slučajno odabrani uzorci od n primjera).

# Binomna distribucija

- Primjer su vjerojatnosti koje se pojavljuju u n uzastopnih međusobno nezavisnih **bacanja novčića.**
- Označimo sa $\pi$ vjerojatnost novčić padne na glavu.
- Napravimo pokus od **n** bacanja novčića i pretpostavimo da je glava pala **r** puta. Razumna procjena za $\pi$ je omjer **r/n**.
- Bacanje novčića jest slučajan pokus – tj. pokus čiji ishod nije jednoznačno određen uvjetima pokusa.
- Zbog toga, ako ponovimo pokus od **n** bacanja novčića očekujemo neku drugu procjenu $\mathbf{r_1/n}$ za pravu vrijednost vjerojatnosti $\pi$.
- Binomna distribucija opisuje vjerojatnost dobivanja točno **r** glava u nizu od **n** bacanja novčića, gdje **r** može biti bilo koja vrijednost od 0 do **n**.

# Binomna distribucija

- Procjena vjerojatnosti $\pi$ pada glave kod slučajnog pokusa bacanja novčića na temelju pokusa od n bacanja novčića ekvivalentna je procjeni prave pogreške $\mathbf{error_D(h)}$ neke dane hipoteze h (npr. stabla odluke) na temelju slučajnog uzorka od n primjera iz D.
- Jedno bacanje novčića odgovara slučajnom odabiru jednog primjera iz D i određivanju da li ga hipoteza h ispravno klasificira.

**Bacanje novčića** ← **Slučajan pokus** → **Hipoteza h klasificira primjere**

Zanima nas ishod:

*pala je glava*

Zanima nas ishod:

*hipoteza h krivo klasificira primjer*

| | |
|---|---|
| **$\pi$**<br>- vjerojatnost pada glave pri jednom bacanju novčića | **$error_D(h)$**<br>- vjerojatnost da će jedan slučajno izvučen primjer biti krivo klasificiran |
| **r**<br>- broj bacanja novčića u nizu od n čiji je ishod bio pad glave novčića | **R**<br>- broj krivo klasificiranih primjera u uzorku od n slučajno odabranih primjera. |
| **r/n**<br>- proporcija pada glave u nizu od n bacanja novčića | **$error_S(h)$**<br>- proporcija krivo klasificiranih primjera u slučajnom uzorku S on n elemenata |

# Procjenitelj

- Slučajna varijabla $error_S(h)$ je varijabla binomne distribucije.
- Ako broj krivo klasificiranih primjera u uzorku S od **n** elemenata označimo s **r** možemo pisati da je

$$\mathbf{error_S(h) = r/n,\ i}$$

$$\mathbf{error_D(h) = \pi,}$$

  gdje je $\pi$ vjerojatnost pogrešne klasifikacije jednog primjera.
- Vjerojatnost $\pi$ nam nije poznata. Pomoću $\mathbf{error_S(h)}$ procjenjujemo $\mathbf{error_D(h) = \pi}$.
- $\mathbf{error_S(h)}$ je **procjenitelj** (*engl. estimator*) za pravu pogrešku.
- Procjenitelj je bilo koja slučajna varijabla pomoću koje procjenjujemo parametar populacije iz koje je izvučen slučajan uzorak.

# Pristranost procjenitelja

- **Pristranost procjenitelja** Y za neki proizvoljan parametar **p** je

$$\mathbf{E[Y] - p.}$$

- Ako je razlika E[Y] – p = 0 kažemo da je **procjenitelj nepristran**.
  - Napomena: E[Y] je očekivanje slučajne varijable Y

*Primjer:*

  - Srednja vrijednost uzorka  je nepristran procjenitelj prave srednje vrijednosti populacije $\mu$.
  - $\mathbf{error_S(h)}$ je nepristran procjenitelj za $\mathbf{error_D(h)}$.
  - Proporcija uzorka **r/n** je nepristran procjenitelj prave proporcije $\pi$.

# Varijanca procjenitelja

- ***Važna napomena:*** Pristranost procjenitelja ne smije se zamijeniti s induktivnom pristranošću. Pristranost procjenitelja je numerička vrijednost dok je induktivna pristranost skup pretpostavki.
- Drugo važno svojstvo procjenitelja je njegova **varijanca**. Ako imamo izbor od više procjenitelja, razumno je izabrati onog s najmanjom varijancom.
- **Varijanca procjenitelja Y**, tj. slučajne varijable Y je

$$V(Y) = E[(Y-EY)^2].$$

- Takav izbor će voditi najmanjoj očekivanoj kvadratnoj pogrešci između procjene i prave vrijednosti parametra.

# Standardna devijacija procjenitelja

- **Standardna devijacija procjenitelja naziva se standardna pogreška** (SE) (standardna devijacija je korijen od varijance).
- *Primjer:*
  - Varijanca slučajne varijable  je $\sigma^2/n$. Standardna devijacija (tj. standardna pogreška) je $\frac{\sigma}{\sqrt{n}}$ . (Centralni granični teorem!)
- *Primjer:*
  - Proporcija je binomna slučajna varijabla pomnožena s konstantom 1/n.
  - Varijanca proporcije P, $V(P) = V(B/n) = 1/n^2 V(B) = \pi(1-\pi)/n$.
  - Standardna pogreška proporcije P je SE(P) = $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$.

# Standardna devijacija procjenitelja

- *Primjer:*
  - Uzorak od n=40 slučajno odabranih primjera. Hipoteza h krivo klasificira r=12 primjera.
  - $error_S(h)$ = r/n = 12/40 =0.3, pa je nepristrana procjena za $error_D(h)$ jednaka 0.3.
  - Standardna pogreška od $error_S(h)$ jednaka je $\sqrt{\frac{\pi(1-\pi)}{n}}$ .
- S obzirom da pravi $\pi$ nemamo, zamjenjujemo ga s njegovim procjeniteljem r/n pa je

$$\sigma_{error_S(h)} = \frac{\sigma_{binomne}}{n} = \sqrt{\frac{0.3(1-0.3)}{40}} = 0.07$$

# Intervalne procjene proporcije

- Uzorak S sadrži n primjera izvučenih slučajno i nezavisno iz populacije s distribucijom D.

  $n > 30$

  hipoteza h krivo klasificira r od n primjera, tj.

  $error_S(h) = r/n$.
- S 95% pouzdanošću stvarna pogreška leži u intervalu:

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

ili općenito

$$error_S(h) \pm u_p\sqrt{\frac{error_S(h)(1-error_S(h))}{n}} ,$$

gdje izbor $u_p$ određuje nivo pouzdanosti:

| Nivo pouzdanosti | 50% | 68% | 80% | 90% | 95% | 98% | 99% |
|---|---|---|---|---|---|---|---|
| $u_p$ | 0.67 | 1 | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

# Intervalne procjene proporcije

- *Primjer:*

  Uzorak S sadrži n = 40 primjera.

  Hipoteza h pogrešno klasificira r = 12 primjera iz S.
- Proporcija pogreške hipoteze h na uzorku S dana je sa:

  $$error_S(h) = r/n = 12/40 = 0.30.$$
- $error_S(h)$ je točkovna procjena za pravu proporciju pogreške hipoteze h, tj. za $error_D(h)$.
- Ako izvučemo drugi slučajni uzorak od 40 elemenata, $error_S(h)$ će poprimiti neku drugu vrijednost zbog slučajnih odstupanja. Svaki uzorak daje drugu točkovnu procjenu.
- Kvalitetniju procjenu prave pogreške hipoteze h možemo dobiti ako odredimo interval pouzdanosti za pravu pogrešku.

# Intervalne procjene proporcije

- Odaberimo 95%-tni interval pouzdanosti. Iz tablica za jediničnu normalnu distribuciju U očitavamo da je pripadna vrijednost $u_p$ za pouzdanost od 95% jednaka $u_p = 1.96$.
  - Standardna pogreška proporcije je $\sqrt{\frac{(.3)(.7)}{40}} = 0.072$
  - Donja granica intervala pouzdanosti = .30 - (1.96)(.072) = .16
    Gornja granica intervala pouzdanosti = .30 + (1.96)(.072) = .44.

  - Prava se vrijednost pogreške hipoteze h nalazi u intervalu $0.16 \leq \pi \leq 0.44$ i to možemo reći s 95% pouzdanosti.

# Intervalne procjene razlike proporcija dviju hipoteza

- Pretpostavimo da **uspoređujemo stope pogrešaka** dvije **hipoteze $h_1$ i $h_2$** za neku ciljnu funkciju s diskretnim vrijednostima. Hipoteza $h_1$ se provjerava na uzorku $S_1$ koji sadrži $n_1$ primjera slučajno izvučenih iz populacije, dok se $h_2$ provjerava na uzorku $S_2$ koji sadrži $n_2$ primjera slučajno izvučenih iz populacije (osnovnog skupa s primjerima s nekom distribucijom D).
- Pretpostavimo da želimo usporediti te dvije hipoteze tj. odrediti razliku u pogreškama koje čine te dvije hipoteze.

$$\boldsymbol{\delta = error_D(h_1) - error_D(h_2)}$$

- $\delta$ je <u>parametar populacije</u> koji želimo procijeniti.
- Procjenitelj je razlika proporcija pogreške hipoteza $h_1$ i $h_2$ određena na temelju uzoraka $S_1$ i $S_2$.

$$d = error_{S1}(h_1) - error_{S2}(h_2).$$

# Intervalne procjene razlike proporcija dviju hipoteza

- Može se dokazati da je $E[d] = \delta$, što znači da je d nepristran procjenitelj za parametar $\delta$.
- Potrebno je poznavati distribuciju uzorkovanja d tj. pripadnu distribuciju vjerojatnosti za procjenitelj d.
- Znamo od prije da su za velike $n_i$, ($n_i \geq 30$), slučajne varijable $error_{S1}(h_1)$ i $error_{S2}(h_2)$ distribuirane približno normalno (Moivre-Laplaceov teorem). Također vrijedi da je razlika dviju normalno distribuiranih slučajnih varijabli - opet normalno distribuirana slučajna varijabla s očekivanjem $\delta$. Može se pokazati da je standardna pogreška od d jednaka

$$SE(d) = \sigma_d = \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$$

# Intervalne procjene razlike proporcija dviju hipoteza

- Kada znamo da je procjenitelj d distribuiran normalno s parametrima: očekivanjem $\delta$ i standardnom pogreškom $\sigma_d$, lako načinimo interval pouzdanosti za pravu razliku pogrešaka $\delta$ koju čine hipoteze $h_1$ i $h_2$.

$$\boldsymbol{\delta = d \pm u_p \sigma_d}\text{, s pouzdanošću ovisnom od } u_p$$

- U specijalnom slučaju možemo hipoteze $h_1$ i $h_2$ testirati na istom uzorku S i tada će d imati nešto manju varijancu.