Prof.dr.sc. Bojana Dalbelo Bašić

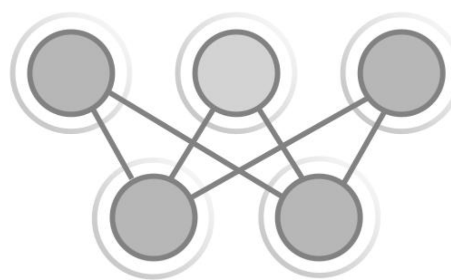
Fakultet elektrotehnike i računarstva
Zavod za elekroniku, mikroelektroniku, računalne i inteligentne sustave

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

Strojno učenje

Bayesova teorija odlučivanja i parametarske metode

Željan Juretić



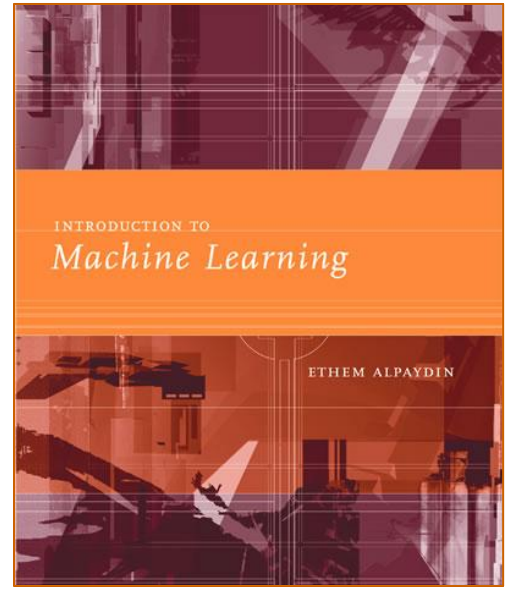
FER
prosinac 2009.

Bayesova teorija odlučivanja (Chapter 3)

- Uvod
- Vjerojatnost i zaključivanje
- Klasifikacija
- Rizici i gubitci
- Funkcije odluke
- Teorija korisnosti
- Vrijednost informacije

Parametarske metode (Chapter 4)

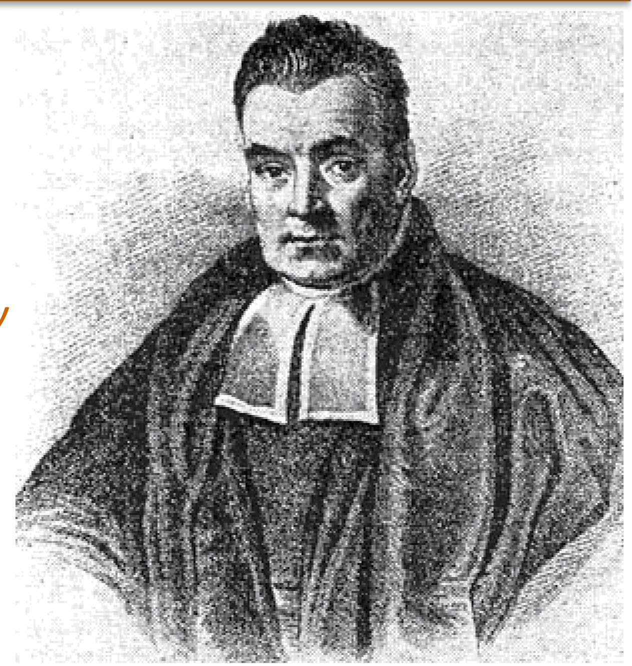
- Uvod
- Kriterij najveće izglednosti *(Maximum Likelihood Estimation)*
- Pristranost i varijanca
- Bayesov estimator
- Parametarska klasifikacija
- Regresija
- Ugađanje složenosti modela: dvojba pristranost/varijanca
- Postupci odabira modela



Bayesova teorija odlučivanja

Ethem Alpaydin,
Introduction to Machine Learning:

Chapter 3: *Bayesian Decision Theory*

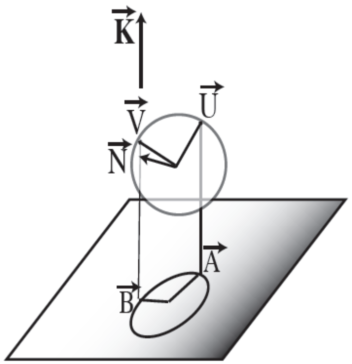


Thomas Bayes, (1702.?-1761.?)

Uvod

- Donošenje odluka pri nesigurnosti ima dugu povijest: Zvijezde, kristalne kugle, vidovnjaci, ...
- Teorija vjerojatnosti je stara tek nekoliko stotina godina, a nastala je kao pokušaj analiziranja igara na sreću: Gerolamo Cardano, Pierre de Fermat, Blaise Pascal, (17. stoljeće).
- U današnje vrijeme spoj statistike i računarske znanosti omogućuje nam zaključivanje na temelju dostupnih podataka *(data mining, automatic classification, itd.)*

- Vidljive i nevidljive varijable (engl. *observable and unobservable*)
- Rezultat pokusa bacanja novčića: $\in \{ pismo, glava \}$
- Slučajna varijabla: $X \in \{1, 0\}$
- Bernoullijev pokus: $P\{X = 1\} = p_0^x (1 - p_0)^{(1-x)}$
- Uzorak: $X = \{x^t\}_{t=1}^N$
- Procjena: $p_0 = \dfrac{\#\{pisama\}}{\#\{bacanja\}} = \dfrac{\sum_t x^t}{N}$



- Predviđanje ishoda sljedećeg bacanja:

$$\begin{cases} pismo, \text{ ako } p_0 > \dfrac{1}{2} \\ glava, \text{ inače} \end{cases}$$

Simulacija

Klasifikacija

- Analiza kreditne sposobnosti građana
- Ulazi su visina prihoda i ušteđevina građana, a izlaz je odluka da li dodjela kredita predstavlja nizak ili visok rizik
- Ulaz: $\boldsymbol{x} = \left[x_1, x_2\right]^T$ Izlaz: $C \in \{0,1\}$
- Predviđanje:

$$\begin{cases} C = 1, \text{ ako } P\left(C = 1 \mid x_1, x_2\right) > 0.5 \\ C = 0, \text{ inače} \end{cases}$$

ili ekvivalentno

$$\begin{cases} C = 1, \text{ ako } P\left(C = 1 \mid x_1, x_2\right) > P\left(C = 0 \mid x_1, x_2\right) \\ C = 0, \text{ inače} \end{cases}$$

Bayesovo pravilo

A priori vjerojatnost

Izglednost (likelihood)

A posteriori vjerojatnost

$$P(C \mid \boldsymbol{x}) = \frac{P(C)\,p(\boldsymbol{x} \mid C)}{p(\boldsymbol{x})}$$

Činjenica (evidence)

Vrijedi ako su hipoteze C_i međusobno isključive, a zbroj njihovih vjerojatnosti iznosi 1

$$P(C=0) + P(C=1) = 1$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x} \mid C=1)P(C=1) + p(\boldsymbol{x} \mid C=0)P(C=0)$$

$$P(C=0 \mid \boldsymbol{x}) + P(C=1 \mid \boldsymbol{x}) = 1$$

Zadatak

Bayesovo pravilo: K>2 razreda

$$P(C_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_i) P(C_i)}{p(\boldsymbol{x})}$$

$$= \frac{p(\boldsymbol{x} \mid C_i) P(C_i)}{\sum_{k=1}^{K} p(\boldsymbol{x} \mid C_k) P(C_k)}$$

Pri čemu mora vrijediti: $P(C_i) \geq 0 \text{ i } \sum_{i=1}^{K} C_i = 1$

Odabiremo: $C_i \text{ ako } P(C_i \mid \boldsymbol{x}) = \max_k P(C_k \mid \boldsymbol{x})$

Rizici i gubici

- Akcija α_i predstavlja odluku da ulaz pridružimo razredu C_i
- Gubitak nastao poduzimanjem akcije α_i ako ulaz pripada razredu C_k označit ćemo sa: λ_{ik}

- Očekivani rizik (engl. *Risk*):

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \boldsymbol{x})$$

Odabiremo: α_i ako $\mathrm{R}(\alpha_i \mid \boldsymbol{x}) = \min_k R(\alpha_k \mid \boldsymbol{x})$

Rizici i gubici: 0/1 gubitak

$$\lambda_{ik} = \begin{cases} 0, & \text{ako } i = k \\ 1, & \text{ako } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i \mid \boldsymbol{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \boldsymbol{x}) \\ &= \sum_{k \neq i} P(C_k \mid \boldsymbol{x}) \\ &= 1 - P(C_i \mid \boldsymbol{x}) \end{aligned}$$

Ako želimo minimalni rizik, odabrat ćemo najvjerojatniji slučaj

Rizici i gubici: odbacivanje

$$\lambda_{ik} = \begin{cases} 0, \text{ ako } i = k \\ \lambda, \text{ ako } i = K+1, \ 0 < \lambda < 1 \\ 1, \text{ inače} \end{cases}$$

$$R(\alpha_{K+1} \mid \boldsymbol{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \boldsymbol{x}) = \lambda \underbrace{\sum_{k=1}^{K} P(C_k \mid \boldsymbol{x})}_{=1} = \lambda$$

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{k \neq i} P(C_k \mid \boldsymbol{x}) = 1 - P(C_i \mid \boldsymbol{x})$$

Odabiremo:

$$\begin{cases} C_i, \text{ ako } P(C_i \mid \boldsymbol{x}) > P(C_k \mid \boldsymbol{x}) \forall k \neq i \ \text{ i } \ P(C_i \mid \boldsymbol{x}) > 1 - \lambda \\ odbaci, \text{ inače} \end{cases}$$

- Što je s rubnim slučajevima kada je $\lambda = 0$, odnosno $\lambda \geq 1$?

Funkcije odluke

Odabiremo:

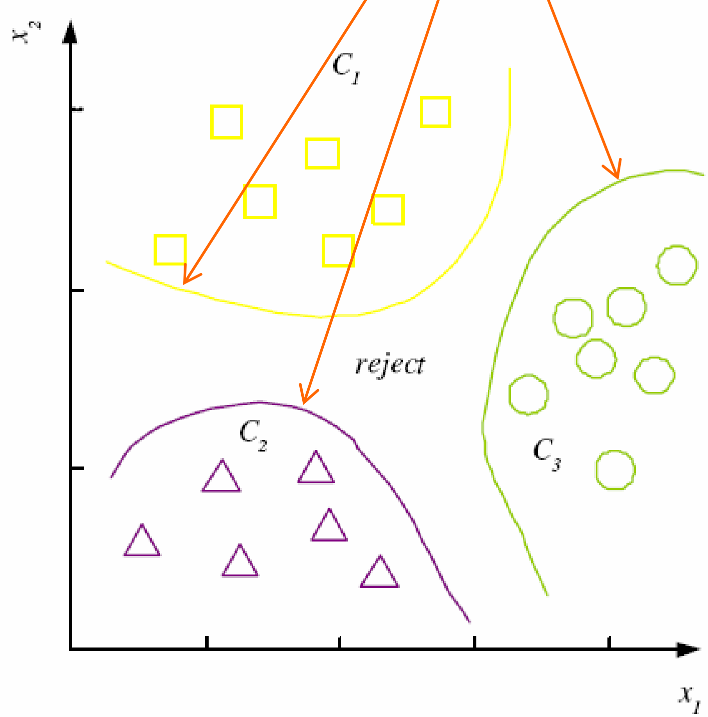
C_i ako $g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})$

$$g_i(\boldsymbol{x}) = \begin{cases} -R(\alpha_i \mid \boldsymbol{x}) \\ P(C_i \mid \boldsymbol{x}) \\ p(\boldsymbol{x} \mid C_i) P(C_i) \end{cases}$$

K područja odluke $R_1, \ldots, R_K$

$$R_i = \left\{ \boldsymbol{x} \mid g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x}) \right\}$$

$g_i(\boldsymbol{x}), \quad i = 1, 2, 3, \ldots, K$



Teorija korisnosti

- Vjerojatnost stanja S_k uz poznat vektor značajki $\boldsymbol{x}$: $P(S_k \mid \boldsymbol{x})$
- Korisnost akcije α_i kada je stvarno stanje S_k : U_{ik}

- Očekivana korisnost (engl. *Expected Utility*):

$$EU(\alpha_i \mid \boldsymbol{x}) = \sum_k U_{ik} P(S_k \mid \boldsymbol{x})$$

Odabiremo: α_i ako $EU(\alpha_i \mid \boldsymbol{x}) = \max_j EU(\alpha_j \mid \boldsymbol{x})$

Vrijednost informacije

- Očekivana korisnost ako koristimo samo vektor značajki $\boldsymbol{x}$:

$$EU(\boldsymbol{x}) = \max_i \sum_k U_{ik} P(S_k \mid \boldsymbol{x})$$

- Očekivana korisnost ako uz vektor značajki $\boldsymbol{x}$ koristimo novu značajku z :

$$EU(\boldsymbol{x}, z) = \max_i \sum_k U_{ik} P(S_k \mid \boldsymbol{x}, z)$$

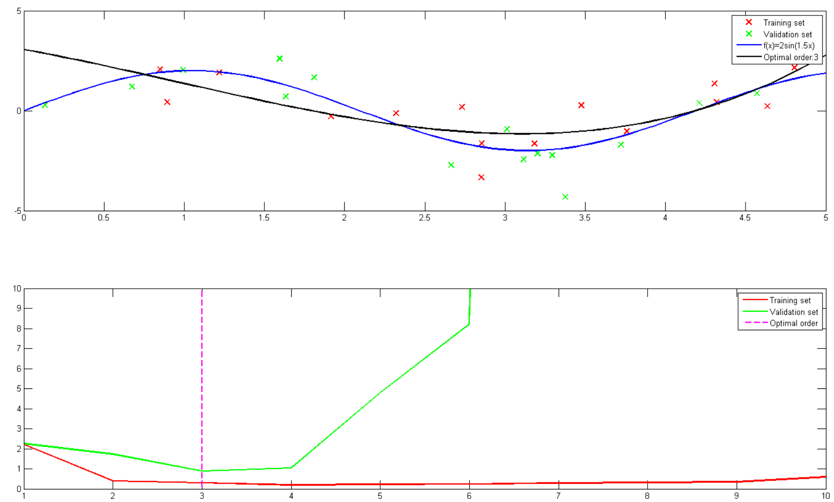
- Kažemo da je značajka z korisna ako vrijedi:

$$EU(\boldsymbol{x}, z) > EU(\boldsymbol{x})$$

Parametarske metode

Ethem Alpaydin, *Introduction to Machine Learning*:

Chapter 4: *Parametric Methods*



Uvod

- Statistika je bilo koja vrijednost izračunata iz danog uzorka.
- Kod statističkog zaključivanja odluke donosimo koristeći informacije koje su nam dostupne iz uzorka.
- Parametarski pristup podrazumijeva da uzorak dolazi iz određene distribucije kojoj se podvrgavaju i još neviđeni podaci.
- Prednost parametarskih metoda je da podatke možemo opisati malim brojem parametara.
- Dovoljna statistika (engl. *Sufficient statistics*) je skup parametara koje nužno moramo poznavati kako bi u potpunosti opisali distribuciju, npr: srednja vrijednost i varijanca čine dovoljnu statistiku Gaussove distribucije $\mathcal{N}(\mu, \sigma^2)$.
- Parametre distribucije procjenjujemo iz danog uzorka te na temelju procijenjene distribucije donosimo potrebne zaključke.

Parametarska procjena

- Smatramo da su značajke iz uzorka nezavisne i identično distribuirane (engl. *Independent identically distributed – iid*) $X = \left\{ x^t \right\}_{t=1}^{N}$
- Pretpostavljamo da su x^t instance dobivene iz neke poznate familije gustoće vjerojatnosti čija je dovoljna statistika skup parametara θ .

$$x^t \sim p(x|\Theta)$$

- Želimo pronaći θ koja bi uzorkovanje x^t iz distribucije $p\left(x\,|\,\theta\right)$ učinilo najvjerojatnijim što je moguće.
- Dovoljan skup parametara θ procjenjujemo na temelju danog uzorka.

Kriterij najveće izglednosti

- Kriterij najveće izglednosti (engl. *Maximum Likelihood Estimation – MLE*) je metoda kojom nalazimo dovoljan skup parametara θ koja bi uzorkovanje x^t iz distribucije $p(x|\theta)$ učinilo najvjerojatnijim što je moguće.
- ***Izglednost** (engl. *Likelihood*) od θ uz dani uzorak X :

$$l(\theta|X) \equiv p(X|\theta) = \prod_{t=1}^{N} p\left(x^t|\theta\right)$$

Zašto logaritmiramo?

- **Logaritamska izglednost** (engl. *Log likelihood*):

$$L(\theta|X) = \log l(\theta|X) = \sum_{t=1}^{N} \log p\left(x^t|\theta\right)$$

- **Procjenitelj najveće izglednosti** (engl. *Maximum likelihood estimator - MLE*):

$$\theta^* = \arg\max_{\theta} L(\theta|X)$$

*U hrvatskoj literaturi koristi se još i naziv *vjerodostojnost*

MLE – Bernoullijeva razdioba

- **Bernoullijeva razdioba**
- Dva stanja, neuspjeh/uspjeh, $x \in \{0,1\}$

$$p(x) = p_0^x (1 - p_0)^{(1-x)}$$

$$\begin{aligned}
L(p_0 \mid X) &= \log l(p_0 \mid X) \\
&= \log \prod_{t=1}^{N} p_0^{(x^t)} (1 - p_0)^{(1-x^t)} \\
&= \sum_{t=1}^{N} x^t \log p_0 + \left(N - \sum_{t}^{N} x^t \right) \log (1 - p_0)
\end{aligned}$$

- **MLE:** $\hat{p} = \dfrac{dL}{dp} = 0 \quad \Rightarrow \quad \hat{p} = \dfrac{\sum_t x^t}{N}$

Dokaz

MLE – Polinomna razdioba

- **Polinomna razdioba**
- Generalizacija Bernoullijeve razdiobe, K>2 stanja, $x \in \{0,1\}$

$$p(x_1, x_2, \ldots, x_K) = \prod_{i=1}^{K} p_i^{x_i}$$

- Radimo ***N*** nezavisnih pokusa s ishodima $X = \{x^t\}_{t=1}^{N}$ pri čemu vrijedi:

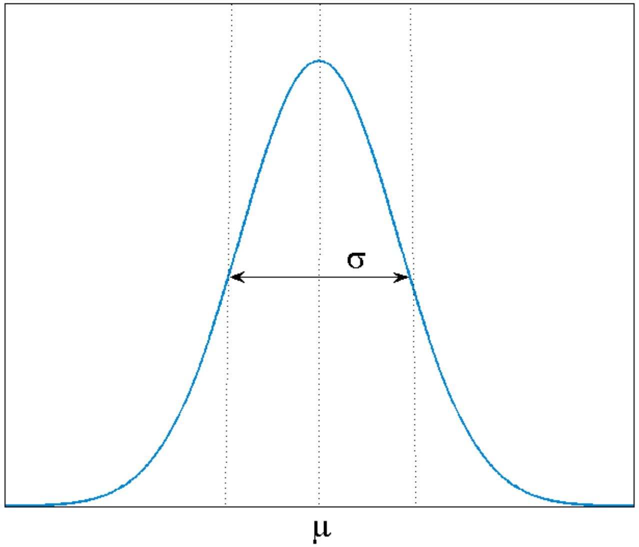
$$x_i^t = \begin{cases} 1, & \text{ako je ishod pokusa } t \text{ stanje } i \\ 0, & \text{inače} \end{cases} \qquad \sum_i x_i^t = 1$$

- Polinomnu razdiobu možemo promatrati kao ***K*** odvojenih Bernoullijevih pokusa stoga vrijedi:
- **MLE:** $\hat{p} = \dfrac{\sum_t x^t}{N}$

MLE – Gaussova (normalna) razdioba

- **Gaussova razdioba**

$$p(x) = \mathcal{N}\left(\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



- **MLE** za μ i σ^2 :

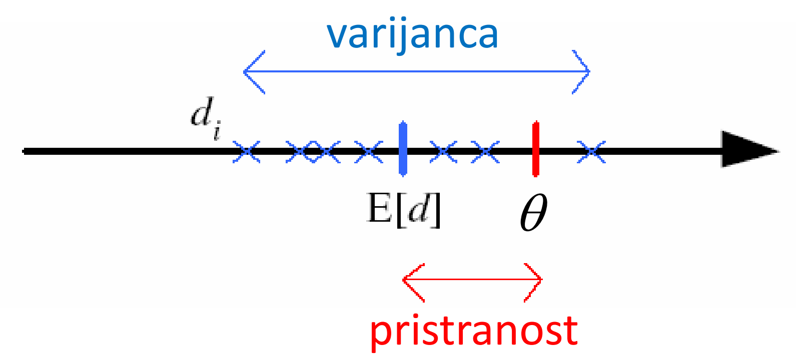
$$m = \frac{\sum_t x^t}{N} \qquad \sigma^2 = \frac{\sum_t \left(x^t - m\right)^2}{N}$$

- **Konvencija:** Malim grčkim slovima označavamo parametre populacije, a malim slovima latinice njihove procjenitelje dobivene iz uzoraka.

Dokaz

Pristranost i varijanca

- Nepoznati parametar θ
- Procjenitelj parametra θ na uzorku X_i je $d_i = d(X_i)$



- **Pristranost:** $b_\theta(d) = E[d] - \theta$
- **Varijanca:** $E\left[\left(d - E[d]\right)^2\right]$
- **Srednja kvadratna pogreška:**

$$r(d,\theta) = E\left[(d-\theta)^2\right]$$

$$= \left(E[d]-\theta\right)^2 + E\left[\left(d - E[d]\right)^2\right]$$

$$= \color{red}{Pristranost^2} + \color{blue}{Varijanca}$$

$$Varijanca = \sigma^2$$

Bayesov estimator

- Ponekad i prije promatranja uzoraka imamo *a priori* informacije o mogućem intervalu vrijednosti koje neki parametar može poprimiti. Takva informacija je izuzetno korisna i uvijek ju treba iskoristiti, posebno u slučajevima kad je skup uzoraka malen.

- PRIMJER: Ekspert iz domene problema nam je rekao da se vrijednost parametra λ u 90% slučajeva nalazi u intervalu [5, 9], (simetrični interval oko 7). Također znamo da traženi parametar ima normalnu razdiobu.

$$P\left\{-1.64 < \frac{\theta-\mu}{\sigma} < 1.64\right\} = 0.9$$

$$P\{\mu - 1.64\sigma < \theta < \mu + 1.64\sigma\} = 0.9$$

- Zaključujemo: $p(\theta) \sim \mathcal{N}\left(7,(2/1.64)^2\right)$

Bayesov estimator

- Promatramo parametar θ kao slučajnu varijablu s pripadnom gustoćom vjerojatnosti $p(\theta)$

- **Bayesovo pravilo:**

$$p(\theta | X) = \frac{p(X|\theta)\,p(\theta)}{p(X)} = \frac{p(X|\theta)\,p(\theta)}{\int p(X|\theta')\,p(\theta')\,d\theta'}$$

- **Procjena gustoće vjerojatnosti kod uzorka *x*:**

$$\begin{aligned} p(x|X) &= \int p(x,\theta|X)\,d\theta \\ &= \int p(x|\theta,X)\,p(\theta|X)\,d\theta \\ &= \int p(x|\theta)\,p(\theta|X)\,d\theta \end{aligned}$$

Bayesov estimator

- Promatramo parametar θ kao slučajnu varijablu s pripadnom gustoćom vjerojatnosti $p(\theta)$

- **Maximum a Posteriori (MAP):** $$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid X) = \arg\max_{\theta} \frac{p(X \mid \theta)\, p(\theta)}{p(X)} = \arg\max_{\theta} p(X \mid \theta)\, p(\theta)$$

Konstantan izraz, ne zavisi od skupa parametara, θ

- **Maximum Likelihood (ML):** $$\theta_{ML} = \arg\max_{\theta} p(X \mid \theta)$$

- **Bayesov estimator:** $$\theta_{Bayes} = E[\theta \mid X] = \int \theta\, p(\theta \mid X)\, d\theta$$

Bayesov estimator - primjer

- $x^t \sim \mathcal{N}(\theta, \sigma_0^2)$ i $\theta \sim \mathcal{N}(\mu, \sigma^2)$

Ako je je $p(\theta \mid X)$ gustoća normalne razdiobe tada vrijedi $\theta_{MAP} = \theta_{Bayes}$

- $\theta_{MAP} = \theta_{Bayes} =$

$$E[\theta \mid X] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

- U ovom slučaju Bayesov estimator je težinski prosjek od apriorne srednje vrijednosti, μ i srednje vrijednosti procjenjene iz uzoraka, m.

- Što možemo zaključiti iz dobivenog izraza:
  1) kad se povećava broj uzoraka, N?
  2) kad je varijanca σ^2 mala?
- Kakvo je fizikalno tumačenje toga?

$$g_i(x) = p(x \mid C_i) P(C_i)$$

ili ekvivalentno

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2} \right]$$

$$g_i(x) = -\frac{1}{2} \log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Parametarska klasifikacija

- Neka nam je dan skup primjera za učenje: $\mathcal{X} = \{x^t, r^t\}_{t=1}^{N}$

$$x \in \mathbb{R} \qquad\qquad r_i^t = \begin{cases} 1, \text{ ako } x^t \in C_i \\ 0, \text{ ako } x^t \in C_j,\ j \neq i \end{cases}$$

- ML procjenitelji su:

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \qquad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \qquad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

- Funkcija odluke sada postaje:

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{\left(x - m_i\right)^2}{2 s_i^2} + \log \hat{P}(C_i)$$

Parametarska klasifikacija

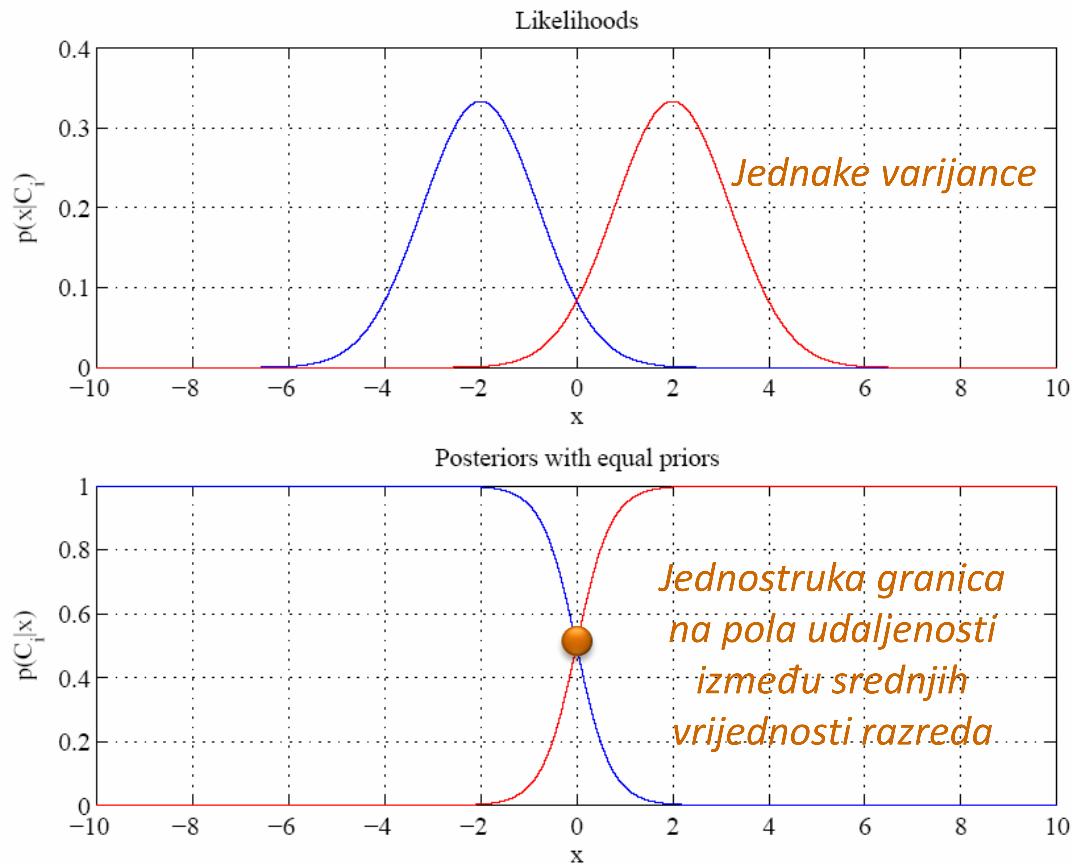
Funkcija izglednosti $p(x|C_i)$ i posteriori vjerojatnost $p(C_i|x)$ za slučaj kada su:

- *a priori* vjerojatnosti za dva razreda jednake
- varijance su jednake

$$g_1(x) = g_2(x)$$

$$(x-m_1)^2 = (x-m_2)^2$$

$$x = \frac{m_1+m_2}{2}$$



$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

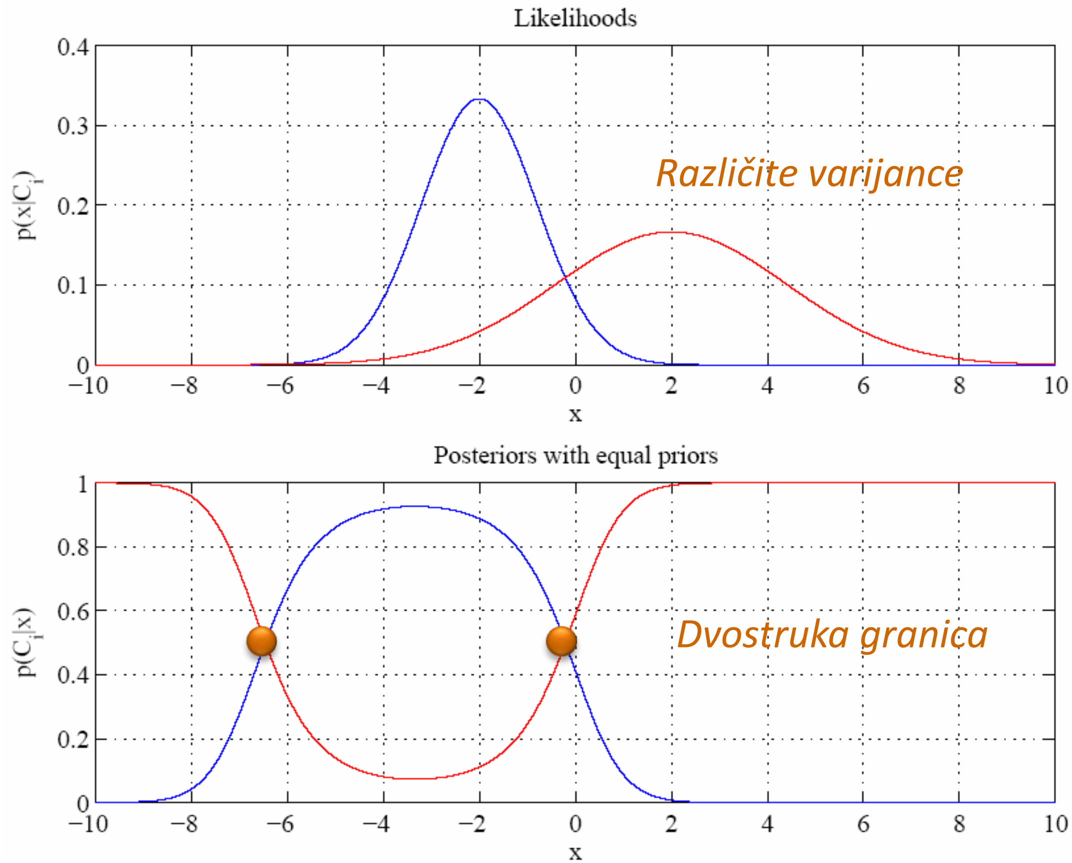
Parametarska klasifikacija

Funkcija izglednosti $p(x \mid C_i)$ i posteriori vjerojatnost $p(C_i \mid x)$ za slučaj kada su:

- *a priori* vjerojatnosti za dva razreda jednake
- varijance su različite

Što se događa ako su *a priori* vjerojatnosti različite?

- Prag odluke se pomiče prema srednjoj vrijednosti manje izglednijeg razreda.



$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

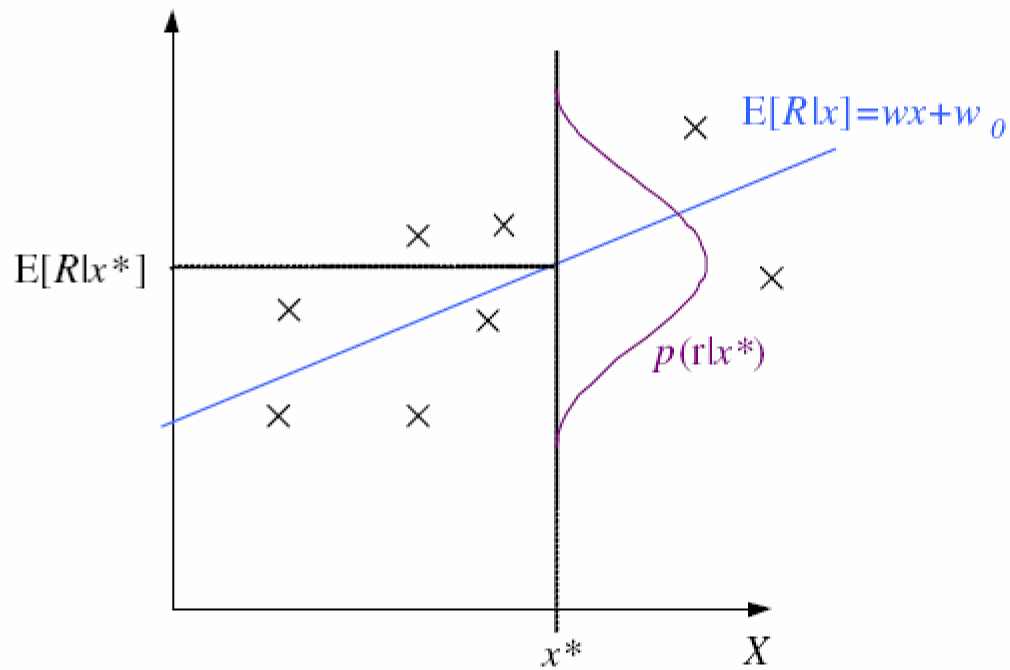
Regresija

$$r = f(x) + \varepsilon$$

procjenitelj: $g(x \mid \theta)$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$$



$$L(\theta \mid X) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

$$p(x, r) = p(r \mid x)\, p(x)$$

Srednja kvadratna pogreška

$$L(\theta \mid X) = \log \prod_{t=1}^{N} p\left(r^t \mid x^t\right) + \log \prod_{t=1}^{N} p\left(x^t\right)$$

$$L(\theta \mid X) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{2\sigma^2}\right]$$

$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2\right]$$

$$= -N \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{\sigma^2}\frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$E(\theta \mid X) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

Linearna regresija

$$g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0 \qquad E(\theta \mid X) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$\frac{\partial E(\theta \mid X)}{\partial w_0}: \quad \sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\frac{\partial E(\theta \mid X)}{\partial w_1}: \quad \sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

- Zapis u matričnom obliku: $\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$

Polinomska regresija

- Generalizacija linearne regresije:

$$g\left(x^t \mid w_k, \ldots, w_2, w_1, w_0\right) = w_k \left(x^t\right)^k + \ldots + w_2 \left(x^t\right)^2 + w_1 x^t + w_0$$

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & \left(x^1\right)^2 & \cdots & \left(x^1\right)^k \\ 1 & x^2 & \left(x^2\right)^2 & \cdots & \left(x^2\right)^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x^N & \left(x^N\right)^2 & \cdots & \left(x^N\right)^k \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

- Zapis u matričnom obliku: $\mathbf{w} = \left(\mathbf{D}^T \mathbf{D}\right)^{-1} \mathbf{D}^T \mathbf{r}$

Mjere pogreške

- **Srednja kvadratna pogreška:**

$$E(\theta \mid X) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

- **Relativna kvadratna pogreška:**

$$E(\theta \mid X) = \frac{\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t - \overline{r}\right]^2}$$

- **Apsolutna pogreška:**

$$E(\theta \mid X) = \sum_{t=1}^{N}\left|r^t - g\left(x^t \mid \theta\right)\right|$$

Ugađanje složenosti modela

- **Očekivana kvadratna pogreška:**

$$E\left[\left(r-g(x)\right)^2 \mid x\right] = \underbrace{E\left[\left(r-E[r \mid x]\right)^2 \mid x\right]}_{šum} + \underbrace{\left(E[r \mid x]-g(x)\right)^2}_{kvadratna\ pogreška}$$

Šum ne ovisi o izboru estimatora. To je dio pogreške kojeg nikad ne možemo ukloniti

Kvadratna pogreška ovisi o izboru procjenitelju i o skupu primjera za učenje

- **Očekivana vrijednost *(prosjek nad svim uzorcima X)*:**

$$E_x\left[\left(E[r \mid x]-g(x)\right)^2 \mid x\right] = \underbrace{\left(E[r \mid x]-E_x\left[g(x)\right]\right)^2}_{pristranost} + \underbrace{E_x\left[\left(g(x)-E_x\left[g(x)\right]\right)^2\right]}_{varijanca}$$

Procjena pristranosti i varijance

- M skupova primjra za učenje, $X_i = \{x_i^t, r_i^t\}, i = 1, \ldots, M$ koristimo kako bi podesili regresijske funkcije u obliku polinoma $g_i(x), i = 1, \ldots, M$.

$$Pristranost^2(g) = \frac{1}{N}\sum_t \left[\bar{g}(x^t) - f(x^t)\right]^2$$

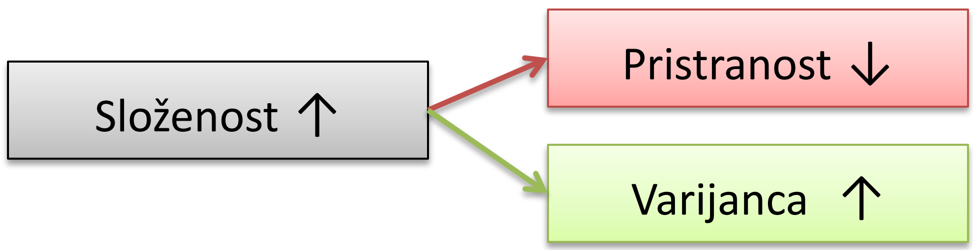
$$Varijanca(g) = \frac{1}{NM}\sum_t \sum_i \left[g_i(x^t) - \bar{g}(x^t)\right]^2$$

- Pri čemu $\bar{g}(x)$ računamo kao:

$$\bar{g}(x) = \frac{1}{M}\sum_t g_i(x)$$

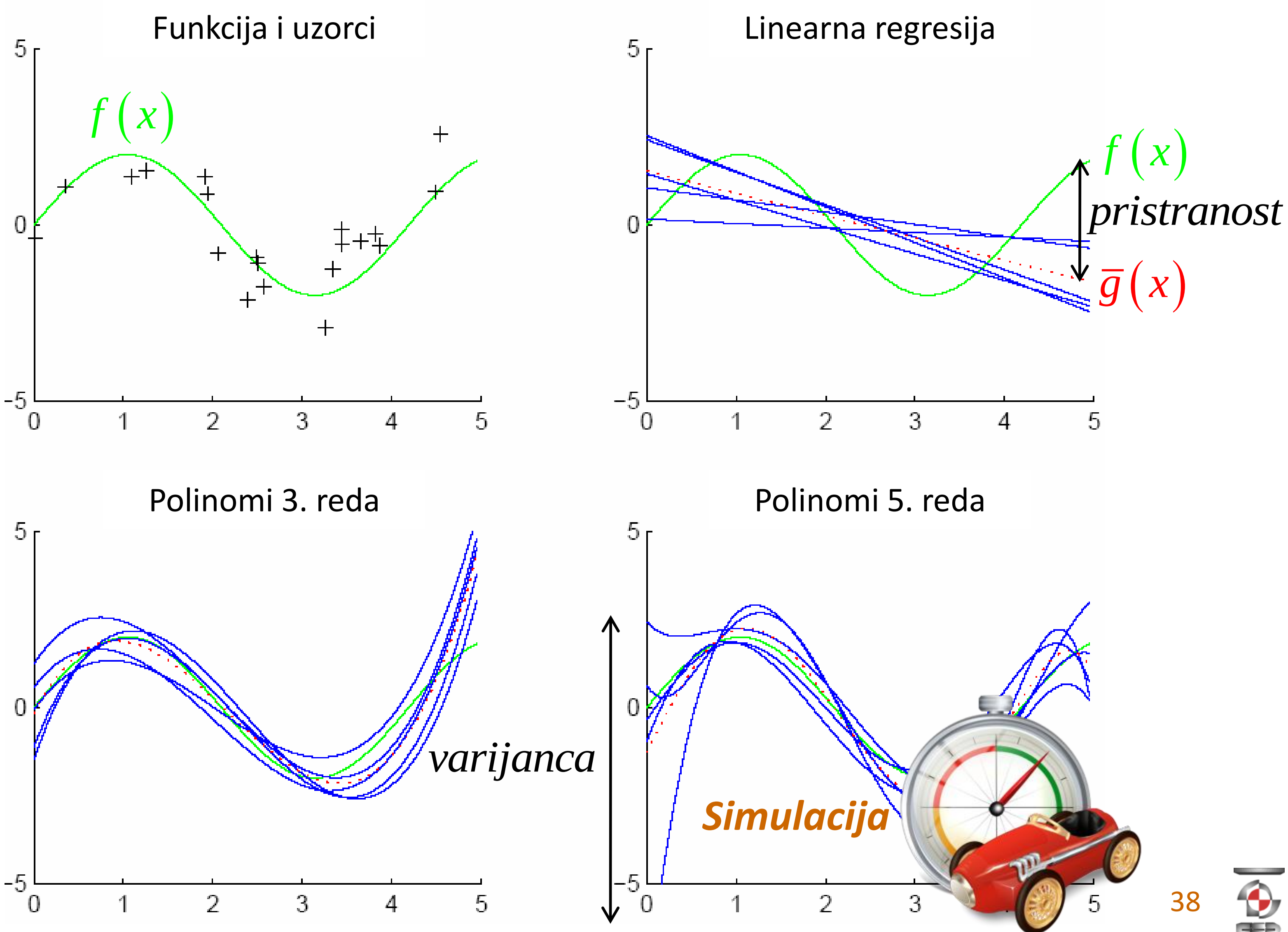
Dvojba pristranost/varijanca

- Dvojba pristranost/varijanca, (engl. *Bias/Variance dilemma):* (Geman et al., 1992)
- Kako povećavamo složenost modela:
  - pristranost se smanjuje (bolje poklapanje s podatcima)
  - varijanca se povećava (poklapanje više varira)

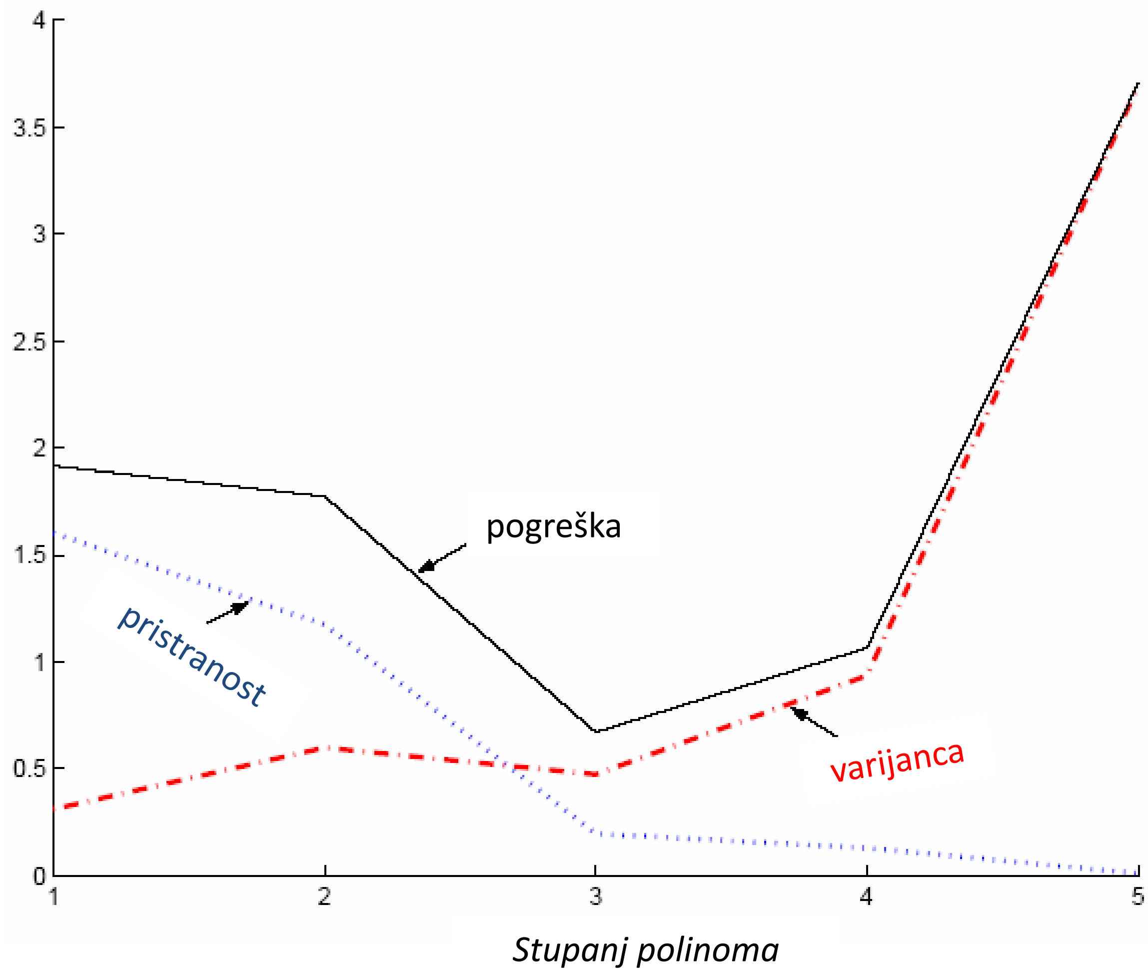


- PRIMJERI:
  - $g_i(x) = 2 = const.$ nema varijancu, ali ima veliku pristranost.
  - $g_i(x) = \sum_t r_i^t / N$ ima manju pristranost, ali ima i varijancu

Dvojba pristranost/varijanca



Polinomska regresija

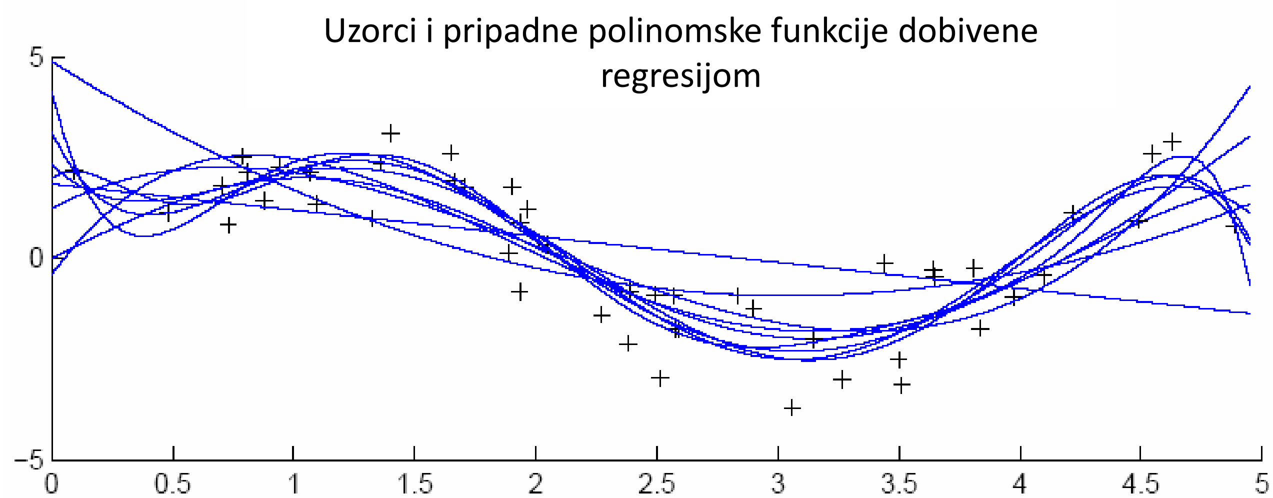


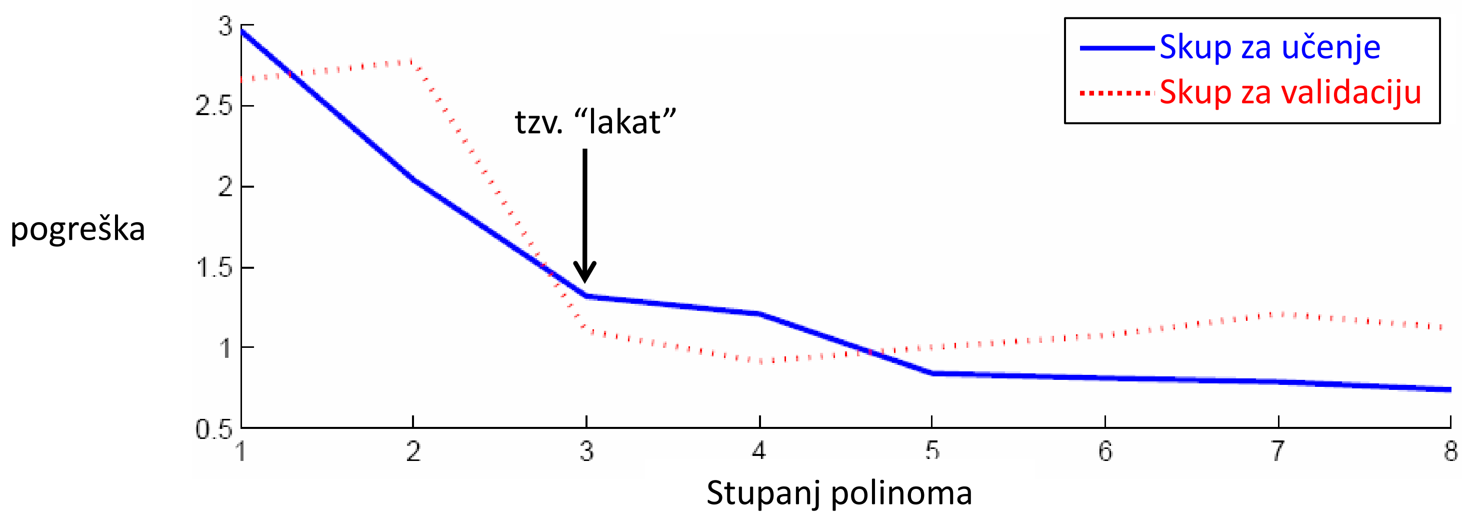
Postupci odabira modela

- **Unakrsna validacija (engl. *Cross-validation*):** ne možemo izračunati pristranost i varijancu za model, ali možemo izračunati ukupnu pogrešku. Uzorke podijelimo na skup za učenje i skup za validaciju. Uvježbamo modele različitih složenosti i ispitujemo njihove pogreške na skupu za validaciju.

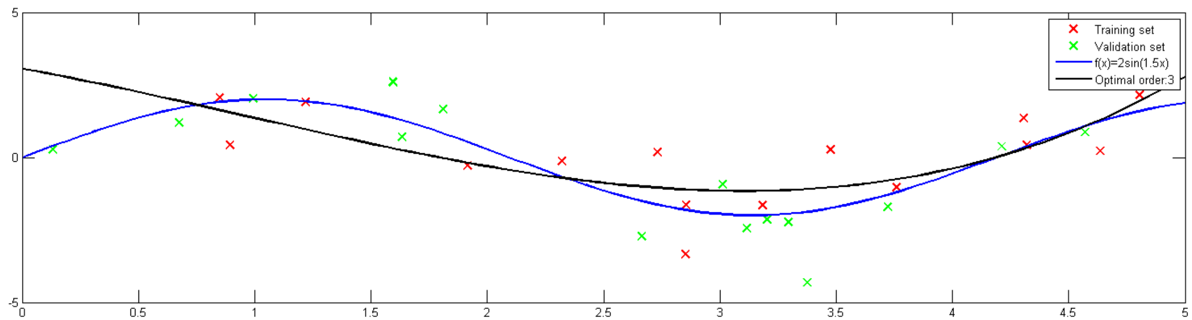
- Dok povećavamo složenost modela, pogreška na skupu za učenje se smanjuje. Pogreška na skupu za validaciju se smanjuje do određene razine složenosti. Pri toj razini pogreška se zaustavlja smanjivati, ili se beznačajno smanjuje. Ako je u podatcima šum jako izražen, pogreška se može i povećavati. Optimalnu složenost modela na grafu prepoznajemo prema karakterističnom obliku lakta (engl. *elbow*).

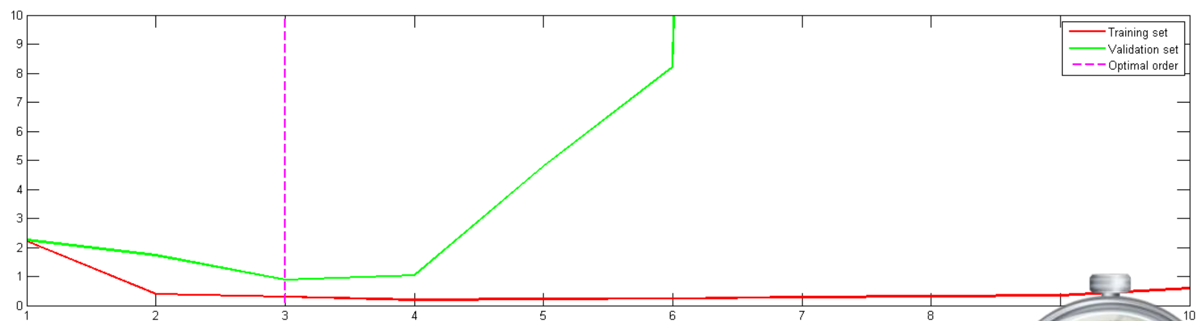
Postupci odabira modela



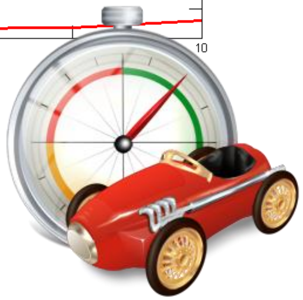


Postupci odabira modela





Simulacija



Postupci odabira modela

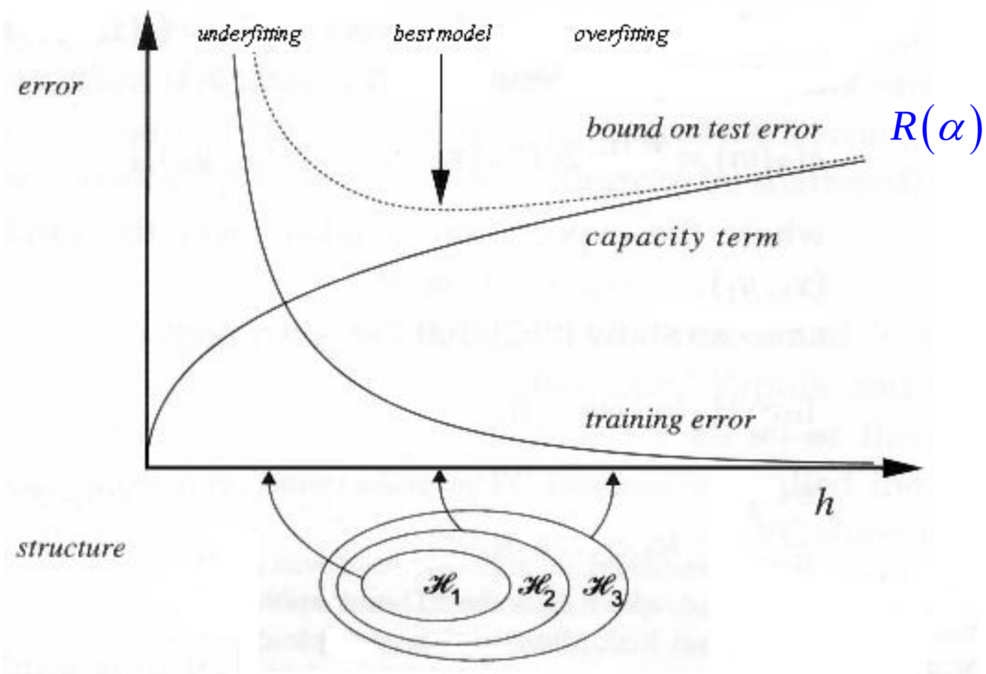
- **Regularizacija**, (Breiman, 1998): Koristimo uvećanu (engl. *augmented*) funkciju pogreške:

$$E' = \text{pogreška nad podatcima} + \lambda \cdot \text{složenost modela}$$

Drugi član izraza "kažnjava" složene modele koji imaju veliku varijancu. Parametar λ predstavlja težinu kazne.

Postupci odabira modela

- **Minimizacija strukturnog rizika** (Vapnik, 1995):
  Predstavlja opći model kontrole složenosti modela i omogućuje nam balansiranje između složenosti samog modela (u nekim slučajevima se može predstaviti VC dimenzijom) i kvalitete “poklapanja” modela s uzorcima iz skupa za učenje. Vrsta regularizacije.



$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h\log\left(\frac{2m}{h}+1\right)-\log\left(\frac{\eta}{4}\right)}{m}}$$

Empirijski rizik (skup za učenje)

Složenost skupa model

Postupci odabira modela

Minimizacija strukturnog rizika

- PRIMJER:

1. Koristeći apriorno znanje iz domene odabiremo klasu funkcija kao što su: polinomi stupnja ***n***, neuronske mreže sa ***n*** skrivenih slojeva i slično.
2. Podijelimo klasu funkcija u hijerarhijski ugnježđene podskupove poredane po rastućoj složenosti. Npr. polinomi rastućeg stupnja.
3. Minimiziramo empirijski rizik na svakom podskupu (odabir optimalnih parametara u pogledu empirijskog rizika).
4. Odabiremo model iz niza sortiranih modela za kojeg je zbroj empirijskog rizika i funkcije ovisne o VC dimenziji minimalan.

Postupci odabira modela

- **Dužina minimalnog opisa (engl. *Minimum description length - MDL*)**, (Rissanen, 1978): promotrimo maksimum a posteriori hipotezu u svijetlu teorije informacije:

$$\theta_{MAP} = \arg\max_{\theta} p(X \mid \theta)\, p(\theta)$$

$$\theta_{MAP} = \arg\max_{\theta} \left(\log_2 P(X \mid \theta) + \log_2 P(\theta)\right)$$

$$\theta_{MAP} = \arg\min_{\theta} \left(\underbrace{-\log_2 P(X \mid \theta) - \log_2 P(\theta)}_{entropija} \right)$$

- MDL nam pruža način odabira složenosti modela. Možemo odabrati između jednostavnijeg modela koji radi određenu pogrešku i složenijeg modela koji savršeno klasificira primjere iz skupa za učenje.

Postupci odabira modela

- **Bayesov odabir modela:** Koristi se kada unaprijed postoji znanje o prikladnom razredu aproksimacijskih funkcija.

$$p(model \mid podatci) = \frac{p(podatci \mid model)\, p(model)}{p(podatci)}$$

- Unakrsna validacija se razlikuje od svih ostalih postupaka odabira modela jer ne pretpostavlja nikakve *a priori* pretpostavke o modelu. Unakrsna validacija je najbolji pristup odabiru modela ako je skup za validaciju dovoljno velik. Ostali modeli postaju korisni kada je skup uzoraka malen.