

4. Teorija vjerojatnosti

prof. dr. sc. Bojana Dalbelo Bašić
doc. dr. sc. Jan Šnajder

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

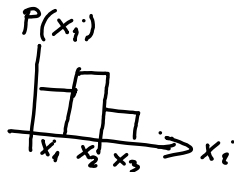
Ak. god. 2012/13.

- 1 Osnovni pojmovi
- 2 Teorijske razdiobe
- 3 Procjena parametara
- 4 Bayesovski procjenitelj

- 1 Osnovni pojmovi
- 2 Teorijske razdiobe
- 3 Procjena parametara
- 4 Bayesovski procjenitelj

Vjerojatnost

- X je s.v., $\{x_i\}$ su njezine vrijednosti
- $P(X = x) = P(X)$
- $P(x_i) \geq 0, \sum_i P(x_i) = 1$
- **distribucija (razdioba) vjerojatnosti**
- zajednička distribucija nad $\{X, Y\}$: $P(X = x, Y = y) = P(x, y)$
- kontinuirana s.v.: **funkcija gustoće vjerojatnosti (PDF)**:



$$\begin{aligned} p(x) &\geq 0 \\ \int_{-\infty}^{\infty} p(x) dx &= 1 \\ P(a \leq X \leq b) &= \int_a^b p(x) dx \end{aligned}$$



Dva pravila teorije vjerojatnosti

(1) Pravilo zbroja

$$P(x) = \sum_y P(x, y)$$

(Marginalna vjerojatnost varijable X)

	y_1	y_2	
x_1	0,5	0,1	$P(x_1) = 0,6$
x_2	0,2	0,2	0,4
	0,7	0,3	①

Uvjetna vjerojatnost:

$$P(y|x) = \frac{P(x, y)}{P(x)}$$

(2) Pravilo umnoška

$$P(x, y) = P(y|x)P(x)$$

$$= P(x|y) \cdot P(y)$$

Bayesovo pravilo

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_y P(x,y)} = \frac{P(x|y)P(y)}{\sum_y P(x,y)P(y)}$$

$= \sum_y P(y|x) \cdot P(y)$

Pravilo lanca (engl. *chain rule*)

$$P(x, y, z) = P(x) \underbrace{P(y|x)}_{P(x,y)} P(z|x, y)$$

Općenito: (faktORIZACIJA)

$$P(x_1, \dots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \cdots P(x_n|x_1, \dots, x_{n-1})$$
$$= \prod_{k=1}^n P(x_k|x_1, \dots, x_{k-1})$$


\rightarrow "faktor"

$$\mathbb{E}[X] = \sum_x xP(x) \quad \mathbb{E}(X) = \sum_x f(x) \cdot P(x)$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx$$

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \quad (a, b \in \mathbb{R})$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$


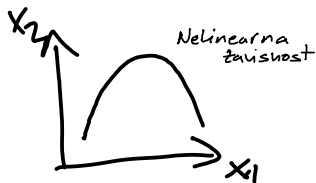
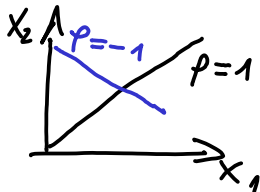
$$\text{Var}(\underline{aX}) = \mathbb{E}[(aX)^2] - \mathbb{E}[aX]^2 = a^2\mathbb{E}[X^2] - a^2\mathbb{E}[X]^2 = \underline{a^2\text{Var}(X)}$$

$$\begin{aligned}\text{Cov}(X, Y) &= \sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ \text{Cov}(X, Y) &= \text{Cov}(Y, X) \\ \text{Cov}(X, X) &= \text{Var}(X) = \sigma_X^2\end{aligned}$$

Pearsonov koeficijent korelacije (linearna zavisnost):

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\rho_{X,Y} \in [-1, +1]$$



Nezavisnost

Varijable X i Y su **nezavisne** akko:

ili

$$P(X, Y) = P(X)P(Y)$$

\swarrow \searrow

$$P(X|Y) = P(X) \quad \text{i} \quad P(Y|X) = P(Y)$$

Znanje o ishodu varijable Y ne utječe na vjerojatnost ishoda varijable X (i obrnuto).

Za nezavisne varijable X i Y vrijedi:

$$\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}[X] \mathbb{E}[Y] \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) \\ \text{Cov}(X, Y) &= \rho_{X,Y} = 0 \end{aligned} \quad \left. \vphantom{\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}[X] \mathbb{E}[Y] \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) \\ \text{Cov}(X, Y) &= \rho_{X,Y} = 0 \end{aligned}} \right\} \text{DZ}$$

Nezavisne varijable su nekorelirane, ali obrat općenito ne vrijedi.

Uvjetna nezavisnost

Varijable X i Y su **uvjetno nezavisne** uz danu varijablu Z , što označavamo kao $X \perp Y | Z$, akko

$$P(X|Y, Z) = P(X|Z)$$

ili

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Jednom kada nam je poznat ishod varijable Z , znanje o ishodu varijable Y ne utječe na ishod varijable X (i obrnuto).

X = "student je priključen na FER"

Y = "student je priključen na PMF-MO"

$P(Y|X) \neq P(Y)$ (nisu marginalno nezavisne!)

Z = "student je sudjelovao na autem. karticirajućima"

$X \perp Y | Z$ $P(Y|X, Z) = P(Y|Z)$

Matrica kovarijacije

(X_1, \dots, X_n) je n -dimenzijski slučajni vektor

Matrica kovarijacije Σ :

$$\Sigma_{ij} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{pmatrix}_{n \times n}$$

$$\Sigma = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]$$

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} \times \begin{bmatrix} \text{---} \end{bmatrix}_{1 \times n} = \begin{bmatrix} \text{---} \end{bmatrix}_{n \times n}$$

1 Osnovni pojmovi

2 Teorijske razdiobe

3 Procjena parametara

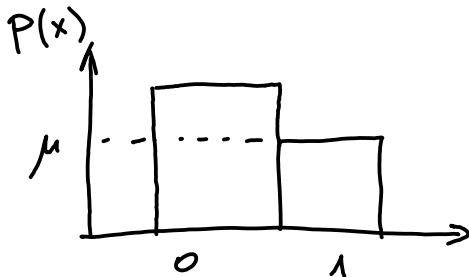
4 Bayesovski procjenitelj

Bernoullijeva razdioba

↓
Binarna
varijabla

$$P(X = x|\mu) = \mu^x(1 - \mu)^{1-x} = \begin{cases} \mu & \text{ako } X = 1 \\ 1 - \mu & \text{inače} \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \text{Var}(X) &= \mu(1 - \mu) \end{aligned} \quad \left. \vphantom{\begin{aligned} \mathbb{E}[X] \\ \text{Var}(X) \end{aligned}} \right\} \text{Dz}$$



Multinomijalna razdioba

Varijabla koja poprima jednu od K vrijednosti

$\mathbf{x} = (x_1, x_2, \dots, x_K)^T$ je binarni vektor indikatorskih varijabli


vektor 1-od-K

$$P(\mathbf{X} = \mathbf{x} | \boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T, \sum_k \mu_k = 1, \mu_k \geq 0$$


$$X = x_3$$

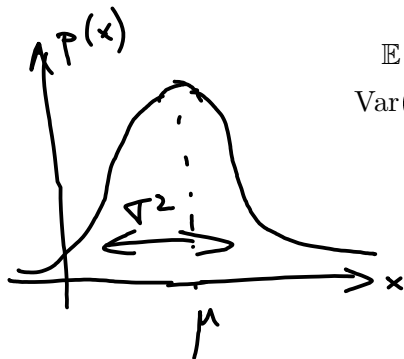
$$X = (0, 0, 1, 0)$$

$\mu_1 \quad \mu_2 \quad \mu_3 \quad \mu_4$

$$P(X = (0, 0, 1, 0)) = \prod_{k=1}^4 \mu_k^{x_k} = 1 \cdot 1 \cdot \mu_3 \cdot 1 = \mu_3$$

Gaussova razdioba

$$p(X = x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$



$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \text{Var}(X) &= \sigma^2 \end{aligned}$$

parametri populacije
(teorijske razdiobe)

Multivarijatna Gaussova razdioba

$$\mathbf{X} = (x_1, \dots, x_n)^T$$

$$p(\mathbf{X} = \mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ - \underbrace{\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{\Delta^2} \right\}$$

inverz!

vektor

kovarijacijska matrica

kvadratna forma

$\boldsymbol{\Sigma}$ mora biti pozitivno definitna (tada je nesingularna i ima inverz).

$\Delta = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ je Mahalanobisova udaljenost između \mathbf{x} i $\boldsymbol{\mu}$.

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$$

$$\text{Cov}(X_i, X_j) = \boldsymbol{\Sigma}_{ij}$$

- 1 Osnovni pojmovi
- 2 Teorijske razdiobe
- 3 Procjena parametara**
- 4 Bayesovski procjenitelj

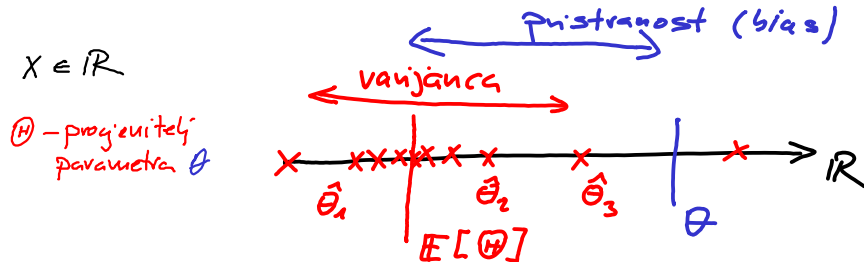
Procjenitelj (engl. *estimator*)

Ideja: na temelju slučajnog uzorka izračunati procjenu (estimaciju) parametra teorijske razdiobe.

Statistika, procjenitelj i procjena

Neka je (X_1, X_2, \dots, X_n) uzorak (n -torka slučajnih varijabli koje su iid). Slučajna varijabla $\Theta = g(X_1, X_2, \dots, X_n)$ naziva se **statistika**. Statistika Θ je **procjenitelj (estimator)** parametra populacije θ . Vrijednost procjenitelja $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ naziva se **procjena**.

Procjenitelj je s.v., dakle ima očekivanje i varijancu.



Pristranost procjenitelja

Nepristran procjenitelj (engl. *unbiased estimator*)

Procjenitelj Θ je **nepristran procjenitelj** parametra θ akko $\mathbb{E}[\Theta] = \theta$.

Pristranost procjenitelja (engl. *estimator bias*):

$$b_{\theta}(\Theta) = \mathbb{E}[\Theta] - \theta$$

Procjenitelj – primjer

X je s.v. sa $x \in \mathbb{R}$.

Označimo $\mathbb{E}[X] = \mu$ (srednja vrijednost) i $\text{Var}(X) = \sigma^2$ (varijanca).

Parametri populacije μ i σ^2 su nepoznati. Možemo ih procijeniti na temelju uzorka $\{x^{(i)}\}_{i=1}^N$ pomoću procjenitelja.

Za procjenitelje možemo upotrijebiti bilo koje statistike. Npr.

$$\hat{\mu} = \frac{1}{N} \sum_i x^{(i)} \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \hat{\mu})^2$$

Jesu li ovo dobri procjenitelji? (Jesu li nepristrani?)

$$\mathbb{E}[\hat{\mu}] = \mu ?$$

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 ?$$

Procjenitelj – primjer

$\mathbb{E}[\hat{\mu}] = \mu$, tj. $\hat{\mu}$ je nepristran procjenitelj srednje vrijednosti. $\leftarrow DZ$

$\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$, tj. $\hat{\sigma}^2$ **nije nepristran** procjenitelj varijance! $\leftarrow DZ$

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$$

Pristranost od $\hat{\sigma}^2$ je

$$b(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{N}$$

$N \rightarrow \infty$
 $b(\hat{\sigma}^2) \rightarrow 0$

Procjenitelj **podcjenjuje** (engl. *underestimates*) pravu varijancu!

Nepristran procjenitelj varijance:

$$\hat{\sigma}_{\text{nepr.}}^2 = \frac{1}{N-1} \sum_{i=1}^N (x^{(i)} - \hat{\mu})^2$$

*Za $N > 30$
nije važno!*

Rastav na pristranost i varijancu

Srednja kvadranta pogreška procjenitelja Θ :

$$r(\Theta, \theta) = \mathbb{E}[(\Theta - \theta)^2]$$

Općenito, za s.v. X i konstantu c vrijedi:

$$\mathbb{E}[(X - c)^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X] - c)^2$$

Dakle, srednju kvadratnu pogrešku možemo izraziti kao:

$$r(\Theta, \theta) = \underbrace{\mathbb{E}[(\Theta - \mathbb{E}[\Theta])^2]}_{\text{varijanca}} + \underbrace{(\mathbb{E}[\Theta] - \theta)^2}_{\text{pristranost}^2} = \text{Var}(\Theta) + b_{\theta}(\Theta)^2$$

Procjenitelji

- 1 Procjenitelj najveće izglednosti
(engl. *maximum likelihood estimator*, MLE) ← najjednostavniji
- 2 Procjenitelj *maximum a posteriori* (MAP)
- 3 Bayesovski procjenitelj ← suvremene ML-metode

Procjenitelj najveće izglednosti (MLE)

Skup neoznačenih primjera $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ koji su iid

$$\mathbf{x}^{(i)} \sim p(\mathbf{x}|\theta) \leftarrow \text{primjeri se pokoravaju PDF-u } p(\mathbf{x}|\theta)$$

MLE određuje najizglednije parametre θ : parametre koje izvlačenje uzorka \mathcal{D} čine najvjerojatnijim

$$p(\mathcal{D}|\theta) = p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}|\theta) = \prod_{i=1}^N p(\mathbf{x}^{(i)}|\theta) \equiv \mathcal{L}(\theta|\mathcal{D})$$

Handwritten notes:
- A red arrow points from the word "izglednosti" in the text above to the product symbol in the equation.
- A red arrow points from the word "izglednosti" in the text above to the term $\mathcal{L}(\theta|\mathcal{D})$.
- A red arrow points from the word "izglednosti" in the text above to the term $\mathcal{L}(\theta|\mathcal{D})$.
- A red arrow points from the word "izglednosti" in the text above to the term $\mathcal{L}(\theta|\mathcal{D})$.

Funkcija izglednosti $\mathcal{L} : \theta \mapsto p(\mathcal{D}|\theta)$ parametrima pridjeljuje vjerojatnost

\mathcal{L} nije PDF! Općenito ne vrijedi $\int_{\theta} \mathcal{L}(\theta|\mathcal{D}) d\theta = 1$.

~~$P(\mathcal{D}|\theta)$~~ !

Funkcija izglednosti – primjer

$\mathcal{D} \equiv 10$ bacanja novčića ($N = 10$)

Glava (H) 8 puta, pismo (T) 2 puta

} Bernoullijeva varijabla

$\mu \equiv$ vjerojatnost da dobijem H (očekivanje varijable)

$$\mathcal{L}(\mu | \mathcal{D}) = P(\mathcal{D} | \mu) = P(x_1, x_2, \dots, x_N | \mu) = \swarrow \text{iid}$$

$$\prod_{i=1}^{10} P(x^i | \mu) = \mu^8 \cdot (1-\mu)^2$$

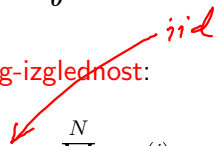
$$P(x | \mu) = \mu^x \cdot (1-\mu)^{1-x}$$

↑
fja parametra μ
Kada je maksimum?
↑
 $\hat{\mu}_{MLE} = ?$

Procjenitelj najveće izglednosti (MLE)

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta|\mathcal{D})$$

Jednostavnije je maksimizirati **log-izglednost**:


$$\ln \mathcal{L}(\theta|\mathcal{D}) \equiv \ln \cancel{\mathcal{L}(\theta|\mathcal{D})} = \ln \prod_{i=1}^N p(\mathbf{x}^{(i)}|\theta) = \sum_{i=1}^N \ln p(\mathbf{x}^{(i)}|\theta)$$

$\ln p(\mathcal{D}|\theta)$

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}} (\ln \mathcal{L}(\theta|\mathcal{D}))$$

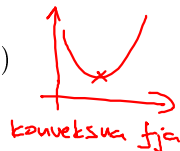
Maksimizaciju provodimo analitički (ako je moguće) ili iterativnim metodama.

MLE za Bernoullijevu razdiobu

log-izjednost

Bernoullijeva razdioba

$$\begin{aligned}\ln \mathcal{L}(\mu|\mathcal{D}) &= \ln \prod_{i=1}^N P(x|\mu) = \ln \prod_{i=1}^N \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}} \\ &= \sum_{i=1}^N x^{(i)} \ln \mu + \left(N - \sum_{i=1}^N x^{(i)}\right) \ln(1-\mu)\end{aligned}$$



$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{\mu} \sum_{i=1}^N x^{(i)} - \frac{1}{1-\mu} \left(N - \sum_{i=1}^N x^{(i)}\right) = 0$$

$$\Rightarrow \hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$$

relativna
frekvencija!

Vrijedi $\mathbb{E}[\hat{\mu}_{\text{ML}}] = \mathbb{E}[X] = \mu$, tj. ovo je nepristran procjenitelj.

MLE za multinomijalnu varijablu

multinomijalna varijabla

$$\ln \mathcal{L}(\boldsymbol{\mu}|\mathcal{D}) = \ln \prod_{i=1}^N P(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = \ln \prod_{i=1}^N \prod_{k=1}^K \mu_k^{x_k^{(i)}} = \sum_{k=1}^K \sum_{i=1}^N x_k^{(i)} \ln \mu_k$$

Izraz treba maksimizirati prema μ_k uz ograničenje $\sum_{k=1}^K \mu_k = 1$.

Primjenom metode Lagrangeovih multiplikatora dobivamo:

$$\hat{\mu}_{k,\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_k^{(i)} = \frac{N_k}{N}$$

relativna frekvencija

MLE za Gaussovu razdiobu

$$\begin{aligned}\ln \mathcal{L}(\mu, \sigma | \mathcal{D}) &= \ln \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right\} \\ &= -\frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{\sum_i (x^{(i)} - \mu)^2}{2\sigma^2}\end{aligned}$$

$$\nabla \ln \mathcal{L}(\mu, \sigma | \mathcal{D}) = 0$$

\vdots

$$\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$$

*srednja vrijednost
uzorka*

pristran!

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \hat{\mu}_{\text{ML}})^2$$

MLE za multivarijatnu Gaussovu razdiobu

$$\begin{aligned}\ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathcal{D}) &= \ln \prod_{i=1}^N p(\mathbf{x}^{(i)} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= -\frac{nN}{2} \ln(2\pi) - \frac{N}{2} |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\end{aligned}$$

$$\nabla \ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathcal{D}) = 0$$

\vdots

$$\hat{\boldsymbol{\mu}}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)}$$

centroid

$$\hat{\boldsymbol{\Sigma}}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{ML}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{ML}})^T$$

Procjenitelj MAP


MLE lako dovodi do **prenaučenosti** modela.

Npr. za uzorak za koji $x_k^{(i)} = 0$ procjena je $\hat{\mu}_{k\text{ML}} = 0$.

Možemo definirati **apriornu razdiobu parametara** $p(\theta)$ i maksimizirati aposteriornu vjerojatnost:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

↑
Bayes



A hand-drawn diagram in red ink showing a bell-shaped curve representing a probability distribution. The vertical axis is labeled $P(\theta)$ and the horizontal axis is labeled μ . A vertical line drops from the peak of the curve to the horizontal axis, where the value $0,5$ is written. An arrow points from the $p(\theta)$ term in the equation above to the curve.

MLE:

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta|\mathcal{D})$$

MAP:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta) p(\theta)$$

Problem: MLE i MAP su **točkaste procjene** (engl. *point estimates*).

- 1 Osnovni pojmovi
- 2 Teorijske razdiobe
- 3 Procjena parametara
- 4 Bayesovski procjenitelj**

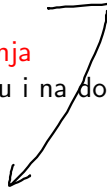
Frekventističko vs. bayesovsko tumačenje

Frekventističko shvaćanje vjerojatnost tumači kao vjerojatnost ishoda kod ponavljanja eksperimenta (relativna frekvencija).

Bayesovska statistika (engl. *bayesian statistics*) vjerojatnost tumači kao nesigurnosti znanja.

Bayesovski procjenitelj: nastavak na ideju MAP-procjenitelja. Kombinira apriorno znanje $p(\theta)$ sa znanjem koje proizlazi iz uzorka. Ne znamo točno kako izgleda $p(\theta)$ (nesigurnost znanja), pa ćemo raditi s očekivanjem.

Bayesovski procjenitelj povezan je s idejom **zaglađivanja** (engl. *smoothing*): raspoređujemo vjerojatnosnu masu i na događaje koji se nisu ostvarili u uzorku.



Ne računamo $P(D|\theta)$ u svakoj točki θ posebno, kao što to rade MLE i MAP

Bayesovski procjenitelj

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta')p(\theta')d\theta'}$$

Bayes (pointing to the numerator)

Aposteriora PDF (pointing to the left side)

Bayesovski procjenitelj je **očekivana vrijednost** od θ s obzirom na distribuciju $p(\theta|\mathcal{D})$:

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{D}] = \int \theta p(\theta|\mathcal{D})d\theta$$

$\mathbb{E}[x] = \int x \cdot p(x)dx$ (with an arrow pointing to the integral)

Integral je analitički izračunljiv uz prikladan udabir apriorne distribucije: onaj kada su $p(\theta)$ i $p(\theta|\mathcal{D})$ istog oblika, tzv. **konjugatne distribucije**.

Ako su $p(\theta)$ i $p(\theta|\mathcal{D})$ konjugatne distribucije, onda $p(\theta)$ zovemo **konjugatnom apriornom distribucijom** za izglednost $p(\mathcal{D}|\theta)$.

Konjugatne apriorne distribucije

Sve razdiobe koje pripadaju **ekponencijalnoj familiji distribucija** (Gaussova, Bernoullijeva, multinomijalna, ...) imaju svoju konjugatnu apriornu distribuciju.

- Gaussova distribuciju: Gaussovu distribucija
- Bernoullijeva distribucija: Beta-distribuciju
- Multinomijalna distribucija: **Dirichletovu distribuciju**

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \text{Dir}(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_k)$$

α_i su hiperparametri modela (parametri distribucije parametara)

Laplaceovo zaglađivanje

Bayesovski procjenitelj za multinomijalnu varijablu: $\mathbf{x} = (x_1, x_2, \dots, x_K)^T$

$$\begin{aligned}\hat{\mu}_{k,\text{Bayes}} &= \mathbb{E}[\mu_k | \mathcal{D}] = \int \boldsymbol{\theta} p(\mu_k | \mathcal{D}) d\mu_k = \frac{p(\mathcal{D} | \mu_k) \text{Dir}(\mu_k | \alpha_k)}{p(\mathcal{D})} \\ &\vdots \\ &= \frac{N_k + \alpha_k}{N + \sum_{i=1}^K \alpha_i}\end{aligned}$$

Uz pojednostavljenje $\alpha_i = \lambda$ dobivamo **Laplaceov procjenitelj**:

$$\hat{\mu}_{k,\text{Bayes}} = \frac{N_k + \lambda}{N + K\lambda}$$

Uz $\lambda = 1$ dobivamo **Laplaceovo pravilo** (*add-one smoothing*):

$$\hat{\mu}_{k,\text{Bayes}} = \frac{N_k + 1}{N + K}$$

relativna frekvencija

- Podsjetili smo se **Bernoullijeve**, **multinomijalne** i **Gaussove** razdiobe
- **Procjenitelj** je statistika (slučajna varijabla izračunata iz uzorka) kojom se procjenjuju parametri neke teorijske distribucije
- Dobri procjenitelji su **nepristrani**
- **Procjenitelj najveće izglednosti (MLE)** odabire parametre koji maksimiziraju vjerojatnost realizacije uzorka (tj. izglednost)
- **MAP-procjenitelj** dodatno koristi apriornu razdiobu parametara i maksimizira aposteriornu vjerojatnost parametara
- **Bayesovski procjenitelj** računa očekivanje aposteriorne vjerojatnosti parametra po svim vrijednostima parametara
- Izvođenjem bayesovskog procjenitelja za multinomijalnu varijablu dobivamo **Laplaceovo zaglađivanje**



Sljedeća tema: Bayesov klasifikator