14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.5

1 Funkcija izglednosti

- Skup podataka $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ koji su **iid**; pretpostavka: $\mathbf{x}^{(i)} \sim p(\mathbf{x}|\boldsymbol{\theta})$
- Vjerojatnost uzorka:

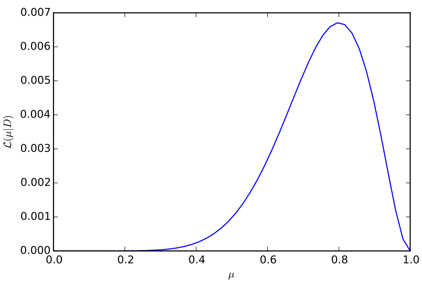
$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^N p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \ \equiv \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$$

gdje je $\mathcal{L} : \boldsymbol{\theta} \mapsto p(\mathcal{D}|\boldsymbol{\theta})$ **funkcija izglednosti**

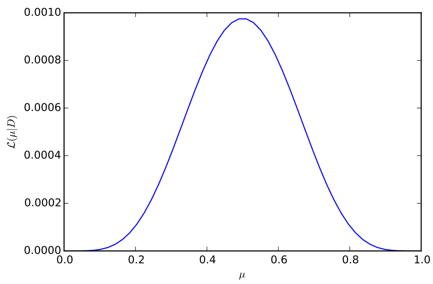
- $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ – vjerojatnost realizacije uzorka $\mathcal{D}$, ako je parametar populacije jednak $\boldsymbol{\theta}$
- Npr. funkcija izglednosti Bernoullijeve varijable – m pozitivnih ishoda u N pokusa:

$$\mathcal{L}(\mu|\mathcal{D}) = P(\mathcal{D}|\mu) = P(x^{(1)}, \ldots, x^{(N)}|\mu) = \prod_{i=1}^N P(x^{(i)}|\mu) = \mu^m(1-\mu)^{(N-m)}$$

gdje $m = \sum_i x^{(i)}$



$m = 8$, $N = 10$



$m = 5$, $N = 10$

2 Procjenitelj MLE

- Pretpostavka: uzorak $\mathcal{D}$ je **najvjerojatniji mogući**, inače ne bi bio izvučen
- Najbolja procjena za $\boldsymbol{\theta}$ je ona koja maksimizira izglednost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$

- **Procjenitelj najveće izglednosti** (*maximum likelihood estimator*) – **MLE**:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$$

- Radi matematičke jednostavnosti, maksimizirat ćemo **log-izglednost**:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \big(\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\big)$$

- MLE za parametar Bernoullijeve razdiobe:

$$\begin{aligned}
\ln \mathcal{L}(\mu|\mathcal{D}) &= \ln \prod_{i=1}^{N} \mu^{x^{(i)}}(1-\mu)^{1-x^{(i)}} = \sum_{i=1}^{N} x^{(i)} \ln \mu + \Big(N - \sum_{i=1}^{N} x^{(i)}\Big) \ln(1-\mu) \\
\frac{\partial \ln \mathcal{L}}{\partial \mu} &= \frac{1}{\mu} \sum_{i=1}^{N} x^{(i)} - \frac{1}{1-\mu}\Big(N - \sum_{i=1}^{N} x^{(i)}\Big) = 0 \\
\Rightarrow\ \hat{\mu}_{\mathrm{MLE}} &= \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \frac{m}{N}
\end{aligned}$$

$\Rightarrow$ **relativna frekvencija** (udio realizacije $x = 1$)

- MLE za parametre kategorijske razdiobe:

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

$\Rightarrow$ maksimizacijom po μ_k uz $\sum_{k=1}^{K} \mu_k = 1$ metodom **Lagrangeovih multiplikatora**:

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

$\Rightarrow$ relativna frekvencija k-te vrijednosti kategorijske varijable

- MLE za parametre normalne razdiobe:

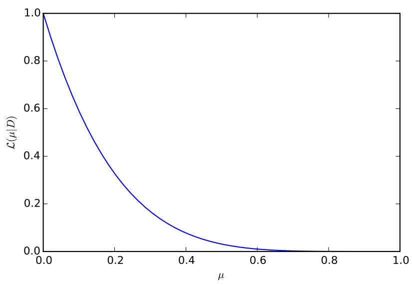
$$\begin{aligned}
\ln \mathcal{L}(\mu, \sigma^2|\mathcal{D}) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\Big) \\
&= -\frac{N}{2}\ln(2\pi) - N \ln \sigma - \frac{\sum_i (x^{(i)}-\mu)^2}{2\sigma^2} \\
\frac{\partial \ln \mathcal{L}}{\partial \mu} &= 0 \ \Rightarrow\ \hat{\mu}_{\mathrm{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \\
\frac{\partial \ln \mathcal{L}}{\partial \sigma^2} &= 0 \ \Rightarrow\ \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \hat{\mu}_{\mathrm{MLE}})^2
\end{aligned}$$

$\Rightarrow$ procjenitelj varijance je pristran (za malen N preporuča ga se korigirati)

- MLE za parametre multivarijatne normalne razdiobe:

$$
\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) &= \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= -\frac{nN}{2}\ln(2\pi) - \frac{N}{2}|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}) \\
\nabla_{\boldsymbol{\mu}} \ln \mathcal{L} = 0 \;\Rightarrow\; \hat{\boldsymbol{\mu}}_{\mathrm{MLE}} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)} \\
\nabla_{\boldsymbol{\Sigma}} \ln \mathcal{L} = 0 \;\Rightarrow\; \hat{\boldsymbol{\Sigma}}_{\mathrm{MLE}} &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{MLE}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{MLE}})^{\mathrm{T}}
\end{aligned}
$$

- MLE smo već koristili kod izvoda funkcije gubitka za poopćene linearne modele
- Minimizacija empirijske pogreške $\Leftrightarrow$ MLE procjena za $\mathbf{w}$ uz odgovarajuću $p(y|\mathbf{x})$
- MLE je sklon **prenaučenosti** – osobito problematično kada je N malen
- Npr., bacanje novčića (Bernoullijeva varijabla): $m = 0$, $N = 5 \Rightarrow \hat{\mu}_{\mathrm{MLE}} = 0$:



3 Procjenitelj MAP

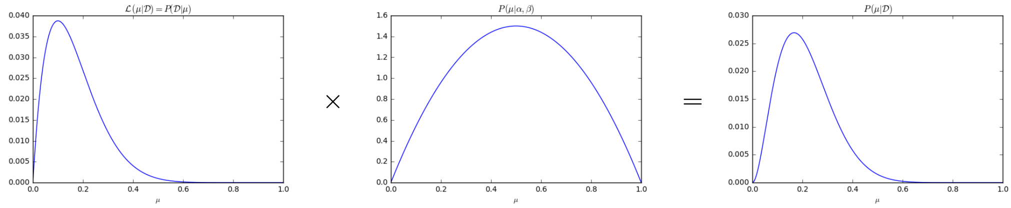
- Želimo kombinirati informacije iz podataka (izglednost $\boldsymbol{\theta}$) s **apriornim znanjem** o $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ – **apriorna razdioba parametra** $\boldsymbol{\theta}$ (*parameter prior*)
- **Aposteriorna vjerojatnost parametra** $\boldsymbol{\theta}$ (Bayesov teorem):

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

- Procjenitelj **maksimum aposteriori (MAP)**:

$$
\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; p(\boldsymbol{\theta}|\mathcal{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\, p(\boldsymbol{\theta})
$$

- Izglednost $\times$ Prior $\propto$ Posterior:



- Rješivo **analitički**, ako $p(\mathcal{D}|\boldsymbol{\theta})$ i $p(\boldsymbol{\theta})$ odaberemo tako da daju neku standardnu $p(\boldsymbol{\theta}|\mathcal{D})$
- **Konjugatne distribucije** $\Leftrightarrow$ $p(\boldsymbol{\theta}|\mathcal{D})$ i $p(\boldsymbol{\theta})$ su iste vrste distribucija
- $p(\boldsymbol{\theta})$ je **konjugatna apriorna distribucija** za $p(\mathcal{D}|\boldsymbol{\theta})$ $\Rightarrow$ $p(\boldsymbol{\theta}|\mathcal{D})$ i $p(\boldsymbol{\theta})$ su konjugatne
- Svaka $p(\mathcal{D}|\boldsymbol{\theta})$ iz **eksponencijalne familije** ima svoju konjugatnu apriornu distribuciju:
  - $p(\mathcal{D}|\boldsymbol{\theta})$ Bernoullijeva $\Rightarrow$ $p(\boldsymbol{\theta})$ beta
  - $p(\mathcal{D}|\boldsymbol{\theta})$ kategorijska $\Rightarrow$ $p(\boldsymbol{\theta})$ Dirichletova
  - $p(\mathcal{D}|\boldsymbol{\theta})$ normalna $\Rightarrow$ $p(\boldsymbol{\theta})$ normalna
  - $p(\mathcal{D}|\boldsymbol{\theta})$ multiv. normalna $\Rightarrow$ $p(\boldsymbol{\theta})$ multiv. normalna

4 Beta-Bernoullijev model

- Konjugatna apriorna distr. za izglednost Bernoullijeve varijable je **beta-distribucija**:

$$p(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

  gdje beta-funkcija B služi za normalizaciju, te $\alpha > 0$ i $\beta > 0$

  - $\alpha = \beta = 1 \Leftrightarrow$ uniformna distribucija $\Rightarrow$ **neinformativna apriorna distribucija**
  - $\alpha > 1,\ \beta > 1 \Rightarrow$ veća gustoća vjerojatnosti za $\mu = 0.5$
  - $\alpha > \beta \Rightarrow$ veća gustoća vjerojatnosti za $\mu \in (0.5, 1)$
  - $\alpha < \beta \Rightarrow$ veća gustoća vjerojatnosti za $\mu \in (0, 0.5)$

- Maksimizator (mod) beta-distribucije: $\frac{\alpha-1}{\alpha+\beta-2}$ (za $\alpha,\beta > 1$)
- Aposteriorna beta-distribucija:

$$\begin{aligned}
p(\mu|\mathcal{D},\alpha,\beta) &= \mu^m(1-\mu)^{N-m}\frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}\frac{1}{p(\mathcal{D})} \\
&= \mu^{m+\alpha-1}(1-\mu)^{N-m+\beta-1}\frac{1}{B(\alpha,\beta)p(\mathcal{D})} \\
&= \mu^{\alpha'-1}(1-\mu)^{\beta'-1}\frac{1}{B(\alpha',\beta')}
\end{aligned}$$

  gdje $\alpha' = m + \alpha$ i $\beta' = N - m + \beta$

- MAP-procjenitelj odgovara modu aposteriorne beta-distribucije:

$$\hat{\mu}_{\mathrm{MAP}} = \frac{\alpha'-1}{\alpha'+\beta'-2} = \frac{m+\alpha-1}{\alpha+N+\beta-2}$$

- Za $N \to \infty$ procjenom dominiraju podatci; za $\alpha = \beta = 1$ MAP degenerira u MLE
- MAP provodi **zaglađivanje** (*smoothing*) – preraspoređivanje mase vjerojatnosti
- **Laplaceovo zaglađivanje (Laplace smoothing)** – MAP sa $\alpha = \beta = 2$:

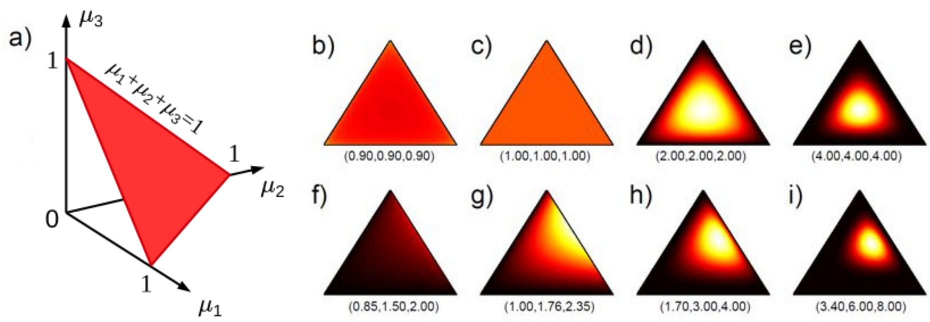
$$\hat{\mu}_{\mathrm{MAP}} = \frac{m+1}{N+2}$$

5 Dirichlet-kategorijski model

- Konjugatna apriorna distr. za multinomijalnu izglednost je **Dirichletova distribucija**

$$P(\boldsymbol{\mu}|\boldsymbol{\alpha}) = P(\mu_1, \dots, \mu_K|\alpha_1, \dots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

  gdje beta-funkcija B služi za normalizaciju, te $\alpha_k > 0$
- Dirichletova distribucija je poopćenje beta-distribucije na K varijabli
- μ_k leže na $(K-1)$-dimenzijskom **standardnom simpleksu**, tj. $\sum_{k=1}^{K} \mu_k = 1$ i $\mu_k \geqslant 0$
- Npr., za $K = 3$, to je trokut u trodimenzijskome prostoru:



- MAP-procjenitelj odgovara modu Dirichletove distribucije:

$$\hat{\mu}_{k,\mathrm{MAP}} = \frac{\alpha'_k - 1}{\sum_{k=1}^{K} \alpha'_k - K}$$

  gdje $\alpha'_k = N_k + \alpha_k$ i $N_k = \sum_i x_k^{(i)}$ (broj nastupanja k-te vrijednosti)
- Uz $\alpha_k = 2$, najvjerojatnija je uniformna distribucija po $\boldsymbol{\mu}$, a procjenitelj je:

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