

PARAMETRICKÉ ODHADY PRAVDĚPODOBNOSTNÍCH ROZDĚLENÍ

Václav Hlaváč

Fakulta elektrotechnická ČVUT v Praze
katedra kybernetiky, **Centrum strojového vnímání**
hlavac@fel.cvut.cz, <http://cmp.felk.cvut.cz/~hlavac>

OBSAH PŘEDNÁŠKY

- ◆ Taxonomie metod odhadů rozdělení pravděpodobností.
- ◆ Parametrické odhady podle maximální věrohodnosti.
- ◆ Parametrické bayesovské odhady.

Poděkování Ing. Vojtěchu Francovi, PhD. za první verzi z podzimu 2005.

TAXONOMIE ODHADŮ PRAVDĚPODOBNOSTNÍCH MODELŮ

1. Podle modelu

- ◆ Parametrické $p(x, \theta); \{x_1, \dots, x_n\} \rightarrow \hat{\theta}$.
- ◆ Neparametrické $p(x); \{x_1, \dots, x_n\} \rightarrow \hat{p}(x)$.

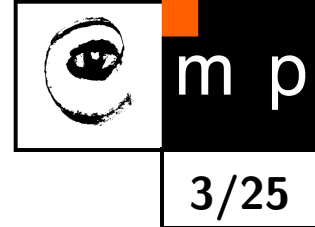
2. Podle vstupní informace

- ◆ Učení s učitelem $p(x, k; \theta); \{(x_1, k_1) \dots, (x_n, k_n)\} \rightarrow \hat{\theta}$.
- ◆ Učení bez učitele $p(x, k; \theta); \{x_1, \dots, x_n\} \rightarrow \hat{\theta}$.

3. Podle principu odhadu

- ◆ Maximální věrohodnost $p(x, k; \theta); \hat{\theta} = \underset{\theta}{\operatorname{argmax}} p(\text{data}, \theta)$.
- ◆ Bayesovský odhad $p(x, k; \theta); \hat{\theta} = \underset{\theta}{\operatorname{argmin}} R(\text{data}, \theta)$.
- ◆ Max-min učení $p(x, k; \theta); \hat{\theta} = \underset{\theta}{\operatorname{argmax}} \min_{i=1, \dots, m} p(x_i, \theta)$.

METODA MAXIMÁLNÍ VĚROHODNOSTI (angl. maximum likelihood method)



Dáno:

- ◆ $T = \{x_1, x_2, \dots, x_n\}$ jsou naměřená data. Předpokládá se i.i.d. = independently and identically distributed, tj. data naměřena nezávisle ze stejného rozdělení.
- ◆ Rozdělení $p(T; \theta)$ až na neznámé parametry $\theta \in \Theta$.

Pravděpodobnost naměření dat T při daném θ

$$p(T; \theta) = p(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$$

se nazývá **věrohodnostní funkce**.

MAXIMÁLNĚ VĚROHODNÝ ODHAD

- ◆ Maximálně věrohodný odhad $\hat{\theta}_{ML}$,

$$\begin{aligned}\hat{\theta}_{ML} &= \operatorname{argmax}_{\theta \in \Theta} p(T; \theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \log p(T; \theta) .\end{aligned}$$

- ◆ Logaritmická věrohodnostní funkce (angl. Log-likelihood) $L(\theta)$.

$$L(\theta) = \log p(T; \theta) .$$

HLEDÁNÍ MAXIMA VĚROHODNOSTNÍ FUNKCE

$$\begin{aligned}\hat{\theta}_{ML} &= \operatorname{argmax}_{\theta \in \Theta} \prod_{n=1}^n p(x_i; \theta) \\ &= \operatorname{argmax}_{\theta \in \Theta} \log \left(\prod_{n=1}^n p(x_i; \theta) \right) \\ &= \operatorname{argmax}_{\theta \in \Theta} \underbrace{\sum_{i=1}^n \log p(x_i; \theta)}_{L(\theta)}\end{aligned}$$

Sloupcový vektor parametrů $\theta = [\theta_1, \dots, \theta_d]^\top \in \mathbb{R}^d$.

Plán: $\frac{\partial L(\theta)}{\partial \theta_j} = 0$; zkráceně $\nabla_{\theta} L = 0$

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \sum_{i=1}^n \log p(x_i; \theta) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \theta_j} \log p(x_i; \theta) \\ &= \sum_{i=1}^n \frac{1}{p(x_i; \theta)} \frac{\partial p(x_i; \theta)}{\partial \theta_j} = 0\end{aligned}$$

VLASTNOSTI ODHADŮ

- ◆ $T = \{x_1, x_2, \dots, x_n\}$ je náhodný výběr z rozdělení $p(x; \theta)$.
- ◆ θ je vektor neznámých parametrů.
- ◆ $\hat{\theta}$ je náhodná veličina. Lze u ní mluvit o střední hodnotě a rozptylu.

VLASTNOSTI ODHADŮ (2)

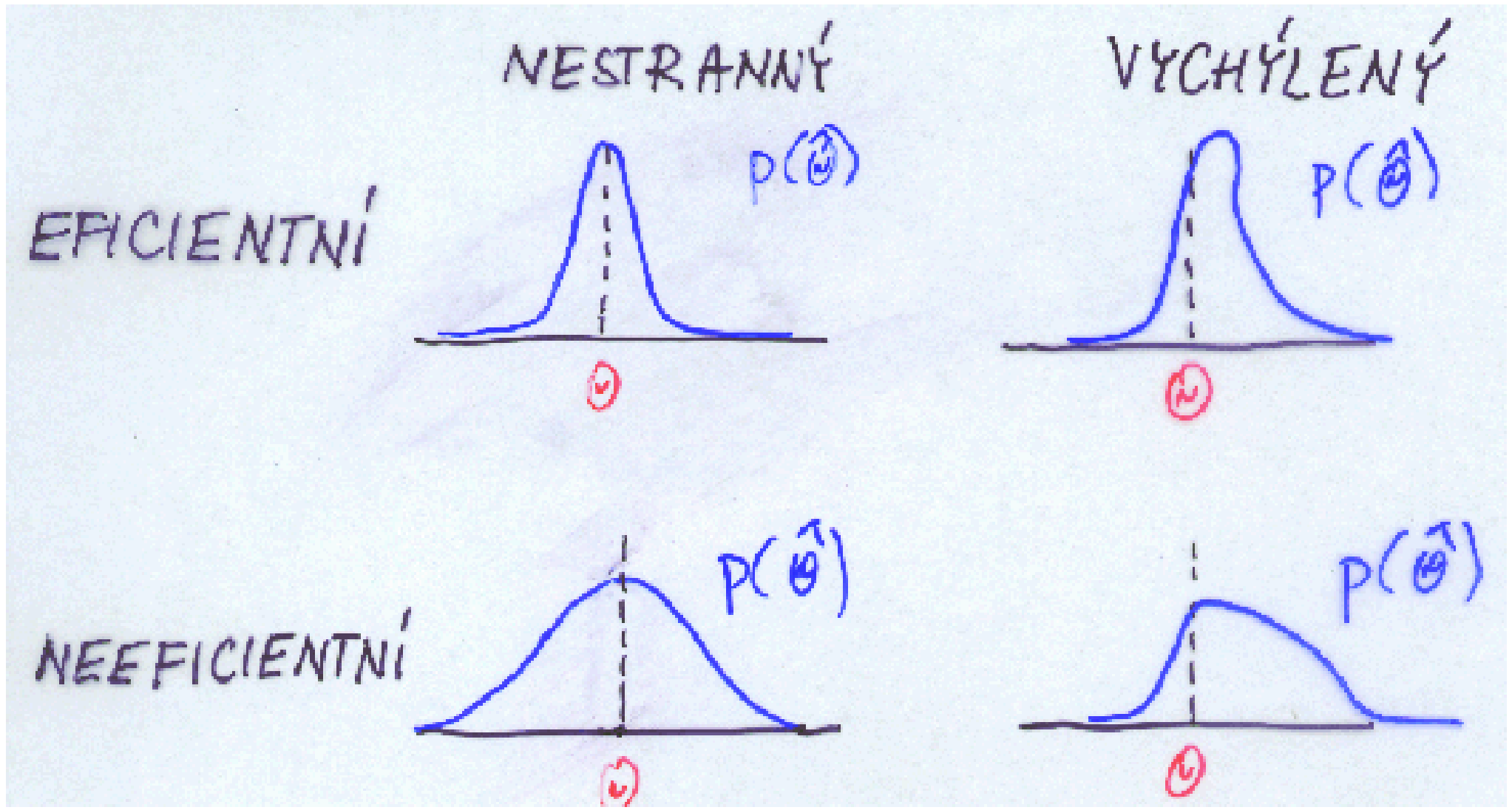
- ◆ Nestranný odhad $E(\hat{\theta}) = \theta$.
- ◆ Eficience odhadu $\text{var}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$.

Nejlepší odhad má minimální rozptyl.

- ◆ Konzistentní odhad
 - $E(\hat{\theta}) = \theta$ pro $n \rightarrow \infty$,
 - $\text{var}(\hat{\theta}) = 0$ pro $n \rightarrow \infty$.

S rostoucím rozsahem výběru se hodnoty statistiky blíží skutečné hodnotě.

ILUSTRACE VLASTNOSTÍ ODHADŮ



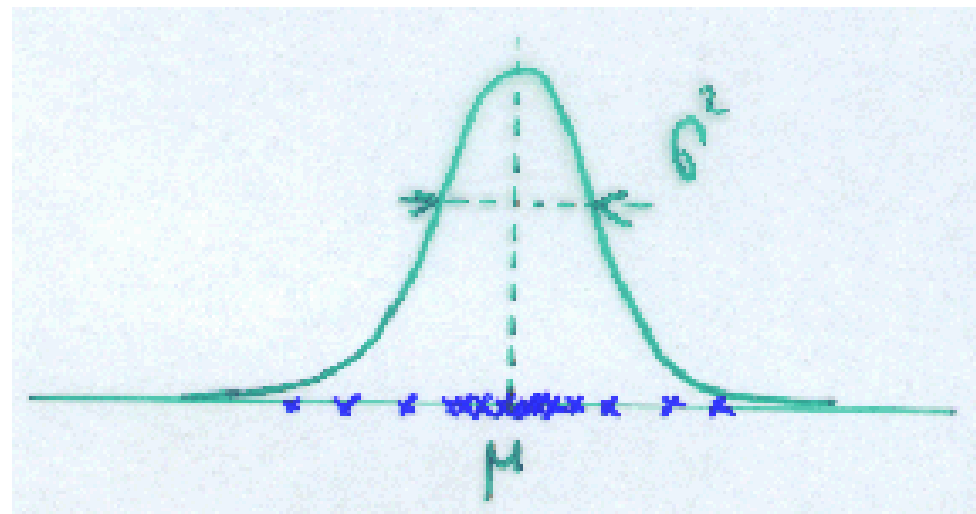
ML odhad je asymptoticky (tj. pro $n \rightarrow \infty$)

1. Nevychýlený.
2. Konzistentní.
3. Nejlepší ve smyslu eficeience.
4. Pravděpodobnostní rozdělení parametru získaného maximálně věrohodným odhadem se blíží Gaussovu rozdělení.

Poznámka: Odhad často mívá uvedené vlastnosti. V obecnosti záruku ovšem poskytnout nelze, protože např. hodnoty odhadu nemusí být reálná čísla, a tak nelze tvrdit, jaké má rozdělení.

Příklad: Odhad parametrů Gaussova rozdělení

- ◆ $T = \{x_1, x_2, \dots, x_n\}$
- ◆ $p(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- ◆ $\theta = [\mu, \sigma]^T$



Příklad: Odhad parametrů Gaussova rozdělení

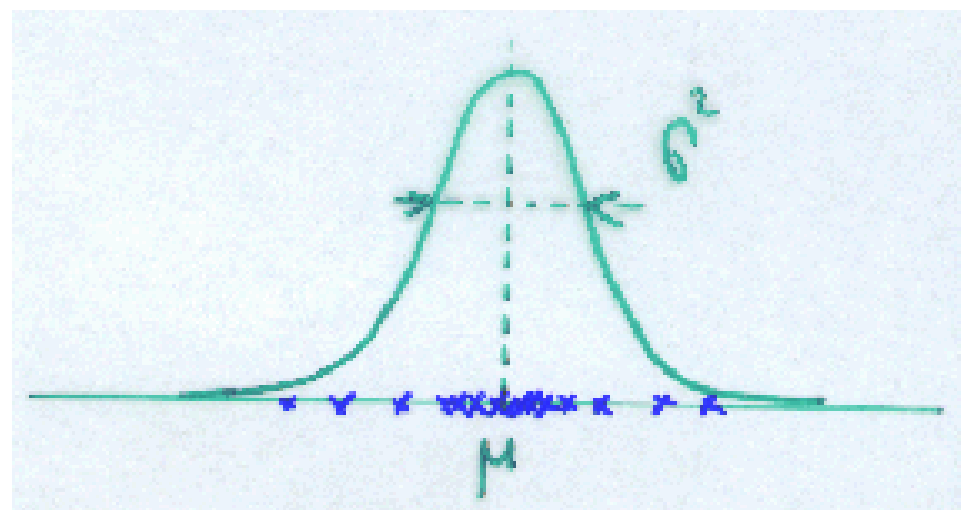
◆ $T = \{x_1, x_2, \dots, x_n\}$

◆ $p(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

◆ $\theta = [\mu, \sigma]^T$

◆ $p(T; \theta) = \prod_{i=1}^n p(x_i; \theta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$

◆ $L(\theta) = \log p(T; \theta) = \sum_{i=1}^n \left(-\log \sigma - \log \sqrt{2\pi} + \frac{-(x_i-\mu)^2}{2\sigma^2} \right)$



Dále budeme hledat maximum věrohodnostní funkce, tj. najdeme její stacionární body $\nabla_{\theta} L = 0$.

Příklad: Odhad parametrů Gaussova rozdělení (2)

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^n \left(\frac{-2(x_i - \mu)}{2\sigma^2} \right) = 0 \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^n \left(\frac{-1}{\sigma} - \frac{(x_i - \mu)^2}{2} \left(\frac{-2}{\sigma^3} \right) \right) = 0 \quad | \cdot \sigma^3$$

$$\sum_{i=1}^n ((x_i - \mu)^2) = 0 \Rightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Příklad: ML odhad diskrétního rozdělení

- ◆ $T = \{x_1, x_2, \dots, x_n\}$
- ◆ $x \in X = \{1, 2, \dots, m\}$

$$p(x; \theta) = \begin{bmatrix} p(x = 1) = \theta_1 \\ p(x = 2) = \theta_2 \\ \vdots \\ p(x = m) = \theta_m \end{bmatrix}$$

$$\begin{aligned} p(x_1, \dots, x_n; \theta) &= \prod_{i=1}^n p(x_i) = \prod_{i=1}^n \sum_{x \in X} p(x) \cdot \delta(x_i = x) \\ &= \prod_{x \in X} p(x)^{n(x)} \end{aligned}$$

$$L(\theta) = \log p(x_1, \dots, x_n; \theta) = \sum_{x \in X} n(x) \log p(x)$$

Příklad: ML odhad diskrétního rozdělení (2)

Minimalizujeme $L(\theta) = \sum_{x \in X} n(x) \log p(x)$

za omezení $\sum_{x \in X} p(x) = 1$, $p(x) \geq 0$

Použijeme metodu Lagrangeových multiplikátorů

$$F(\theta) = \sum_{x \in X} n(x) \log p(x) + \lambda \left(\sum_{x \in X} p(x) - 1 \right)$$

$$\frac{\partial F(\theta)}{\partial p(x)} = \frac{n(x)}{p(x)} + \lambda = 0, \quad \text{pro } \forall x$$

Příklad: ML odhad diskrétního rozdělení (3)

Přepsáno z minulé obrazovky $\frac{\partial F(\theta)}{\partial p(x)} = \frac{n(x)}{p(x)} + \lambda = 0$, pro $\forall x$

$$n(x) = -\lambda p(x) \quad \Rightarrow \quad p(x) = \frac{n(x)}{-\lambda}$$

Z omezení plyne $1 = \sum_{x \in X} p(x) = \sum_{x \in X} \frac{n(x)}{-\lambda} = \frac{1}{-\lambda} \sum_{x \in X} n(x) = \frac{n}{-\lambda}$

$$\lambda = -n$$

$$p(x) = \frac{-n(x)}{\lambda} = \frac{n(x)}{n}$$

POUŽITÍ ML V UČENÍ S UČITELEM

- ◆ $T_{XK} = \{(x_1, k_1), (x_2, k_2) \dots, (x_n, k_n)\}$
- ◆ $x \in X \dots$ diskrétní nebo spojité
- ◆ $k \in K \dots$ diskrétní
- ◆ $p(x, k; \theta) = p(x|k; \theta) p(k; \theta_A)$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{|K|} \\ \theta_A \end{bmatrix}$$

$$p(T_{XK}; \theta) = \prod_{j=1}^n p(x_j, k_j; \theta) = \prod_{j=1}^n p(x_j|k_j; \theta_{k_j}) p(k_j|\theta_A)$$

Vytvoříme logaritmickou věrohodnostní funkci $L(\theta)$

POUŽITÍ ML V UČENÍ S UČITELEM (2)

$$\begin{aligned} L(\theta) &= \log p(T_{XK}; \theta) = \sum_{j=1}^n \log p(x_j | k_j; \theta_{k_j}) + \sum_{j=1}^n p(k_j | \theta_A) \\ &= \sum_{k \in K} \sum_{j \in T_k} \log p(x_j | k; \theta_k) + \sum_{j=1}^n p(k_j | \theta_A) \end{aligned}$$

Původní úloha se rozloží na $|K| + 1$ nezávislých úloh.

$$\hat{\theta}_k = \operatorname{argmax}_{\theta_k} L(\theta_k) = \sum_{j \in T_k} \log p(x_j | k; \theta_k), \text{ pro } \forall k \in K$$

$$\hat{\theta}_A = \operatorname{argmax}_{\theta_A} L(\theta_A) = \sum_{j=1}^n p(k_j | \theta_A)$$

POUŽITÍ ML V UČENÍ BEZ UČITELE

- ◆ $T_X = \{x_1, x_2, \dots, x_n\}$... pouze pozorovatelné stavy.
- ◆ $p(x, k; \theta) = p(x|k; \theta_k) p(k, \theta_A)$... statistický model uvažuje skrytý stav, který je neznámý.

$$\begin{aligned}
 p(T_x; \theta) &= \prod_{j=1}^n p(x_j; \theta) = \prod_{j=1}^n \sum_{k \in K} p(x_j, k; \theta) \\
 &= \prod_{j=1}^n \sum_{k \in K} p(x_j|k; \theta) p(k; \theta_A)
 \end{aligned}$$

$$\begin{aligned}
 L(\theta) &= \log p(T_x; \theta) = \sum_{j=1}^n \log \sum_{k \in K} p(x_j|k; \theta) p(k; \theta_A) \\
 &= \sum_{j=1}^n \log \left(\sum_{k \in K} p(x_j|k; \theta) p(k; \theta_A) \right) \quad (\text{EM})
 \end{aligned}$$

$\nabla_{\theta}L(\theta) = 0$ nemá analytické řešení!

Používá se Expectation-Maximization (EM) algoritmus

- ◆ Iterační metoda pro odhad ML podle tvaru (EM) na předchozí obrazovce.
- ◆ Lze použít, když umíme udělat ML odhad z T_{XK} .
- ◆ EM algoritmu bude věnována samostatná přednáška.

VLASTNOSTI ML ODHADU

Klady

- ◆ ML má příznivé statistické vlastnosti. Je asymptoticky ($n \rightarrow \infty$) nevychýlený, má nejmenší rozptyl, je konzistentní.
- ◆ Pro mnoho “jednoduchých rozdělání” vede ML na snadné analytické řešení, viz příklad s gaussiánem.

Zápory

- ◆ Může mít více než jedno řešení stejné kvality.
- ◆ Existuje jediné globální řešení, ale je těžké ho najít.
- ◆ Pro směsi rozdělání neexistuje analytické řešení, viz EM algoritmus.

BAYESOVSKÝ ODHAD PARAMETRŮ PRAVDĚPODOBNOSTNÍHO ROZDĚLENÍ

- ◆ $T = \{x_1, x_2, \dots, x_n\}$... naměřená data.
- ◆ $p(T, \theta)$... sdružené rozdělení T a θ .
- ◆ $\theta \in \Theta$... chápe se jako realizace náhodné veličiny (na rozdíl od ML).
- ◆ $p(\theta)$... máme apriorní znalost o θ .

Odhad $\hat{\theta} = \Psi(T)$ formulujeme stejně jako v bayesovském rozhodování se ztrátovou funkcí $W: \Theta \times \Theta \rightarrow \mathbb{R}$.

BAYESOVSKÝ ODHAD PARAMETRŮ PRAVDĚPODOBNOSTNÍHO ROZDĚLENÍ (2)

Bayesovské riziko

$$R(\Psi) = \sum_T \sum_{\theta} p(T, \theta) W(\theta, \Psi(T))$$

Optimální rozhodovací funkce

$$\Psi^*(T) = \operatorname{argmin}_{\Psi} R(\Psi) = \operatorname{argmin}_{\psi} \sum_{\theta} p(T, \theta) W(\theta, \Psi(T))$$

BAYESOVSKÝ ODHAD PRO

$$W(\theta, \Psi(T)) = (\theta - \psi(T))^2$$

Po dosazení $\Psi^*(T) = \operatorname{argmin}_{\Psi} \sum_{\theta} p(T, \theta) (\theta - \Psi)^2$

Pro minimum ztrát musí platit

$$\frac{\partial}{\partial \Psi} \left(\sum_{\theta} p(T, \theta) (\theta - \Psi)^2 \right) = 0$$

$$\Psi^* \sum_{\theta} p(T, \theta) = \sum_{\theta} \theta p(T, \theta)$$

$$\Psi^*(T) = \frac{\sum_{\theta} \theta p(T, \theta)}{p(T)} = \frac{\sum_{\theta} \theta p(\theta|T) p(T)}{p(T)} = \frac{\cancel{p(T)} \sum_{\theta} \theta p(\theta|T)}{\cancel{p(T)}}$$

$$\Psi^*(T) = \sum_{\theta} \theta p(\theta|T)$$

PŘÍKLAD, HÁZENÍ MINCÍ

- ◆ Data (trénovací posloupnost): $T = \{\text{panna, panna, orel, panna, \dots, orel}\}$.
- ◆ Pravděpodobnostní model neznámé mince $p(\text{orel}) = 1 - p(\text{panna}) = \Theta$.
- ◆ Předpokládejme pravděpodobnostní Bernoulliho rozdělení $p(T|\Theta) = \Theta^k (1 - \Theta)^{n-k}$, kde n je počet vzorků v T , k je číslo udávající kolikrát padl orel.
- ◆ Necht' $p(\Theta) = \begin{cases} 1 & \text{pro } \Theta \in \langle 0, 1 \rangle, \\ 0 & \text{pro ostatní hodnoty.} \end{cases}$

$$\begin{aligned} \hat{\Theta}_B &= \Psi^*(T) = \frac{\int_0^1 \Theta p(T|\Theta) p(\Theta) d\Theta}{p(T)} \\ &= \frac{\int_0^1 \Theta^{k+1} (1 - \Theta)^{n-k} d\Theta}{\int_0^1 \Theta^k (1 - \Theta)^{n-k} d\Theta} = \frac{k + 1}{n + 2}. \end{aligned}$$

- ◆ Poznamenejme, že $\hat{\Theta}_{ML} = \frac{k}{n}$. Pro $n \rightarrow \infty$ také $\hat{\Theta}_B \rightarrow \hat{\Theta}_{ML}$.