NMAI059 Pravděpodobnost a statistika 1
12. přednáška

Robert Šámal

Přehled

Bayesovská statistika

Srovnání dvou přístupů ke statistice

Frekventistický/klasický přístup

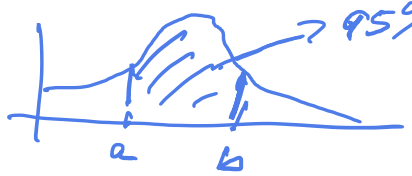
- Pravděpodobnost je dlouhodobá frekvence (z 6000 hodů kostkou padla šestka 1026-krát). Je to objektivní vlastnost reálného světa.
- Parametry jsou pevné, neznámé konstanty. Nelze o nich říkat smysluplné pravděpodobnostní výroky.
- Navrhujeme statistické procedury tak, aby měly žádané dlouhodobé vlastnosti. Např. 95 % z našich intervalových odhadů pokryje neznámý parametr.

Bayesovský přístup

- Pravděpodobnost popisuje, jak moc věříme nějakému jevu, jak moc jsme ochotní se vsadit. (Pravděpodobnost, že Thomas Bayes měl 18. prosince 1760 šálek čaje, je 90 %.)
- Můžeme vyslovovat pravděpodobnostní výroky i o parametrech (třebaže jsou to pevné konstanty).
- Spočítáme distribuci ϑ a z ní tvoříme bodové a intervalové odhady, atd.

Bayesovská metoda – základní popis

- neznámý parametr považujeme za náhodnou veličinu Θ
- zvolíme *apriorní distribuci (prior distribution)*, neboli hustotu pravděpodobnosti $f_\Theta(\vartheta)$ nezávislou na datech.
- zvolíme statistický model $f_{X|\Theta}(x|\vartheta)$, který popisuje, co naměříme (s jakou pravděpodobností), v závislosti na hodnotě parametru
- poté, co pozorujeme hodnotu $X = x$, spočítáme *posteriorní distribuci (posterior distribution)* $f_{\Theta|X}(\vartheta|x)$
- z té pak odvodíme, co potřebujeme např. najdeme a, b, aby $\int_a^b f_{\Theta|X}(\vartheta|x)d\vartheta \geq 1 - \alpha$



- $\vartheta = \theta$ malá théta, Θ je velká théta

Bayesova věta

$P(X=x \mid \Theta=\vartheta)$ ... podmíněná pravd. dle realizace v klas. statistd. $P(X=x;\vartheta)$ .... normální pravd., ϑ je parametr

Věta (Bayesova pro diskrétní náhodné veličiny)

X, Θ jsou diskrétní n.v.

$P(X=x \mid \Theta=\vartheta)\ P(\Theta=\vartheta)$

$P(\Theta=\vartheta \mid X=x)$

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

(sčítance s $p_\Theta(\vartheta') = 0$ považujeme za 0).

Věta (Bayesova pro spojité náhodné veličiny)

X, Θ jsou spojité n.v., které mají hustotu f_X, f_Θ i sdruženou hustotu $f_{X,\Theta}$

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')d\vartheta'}.$$

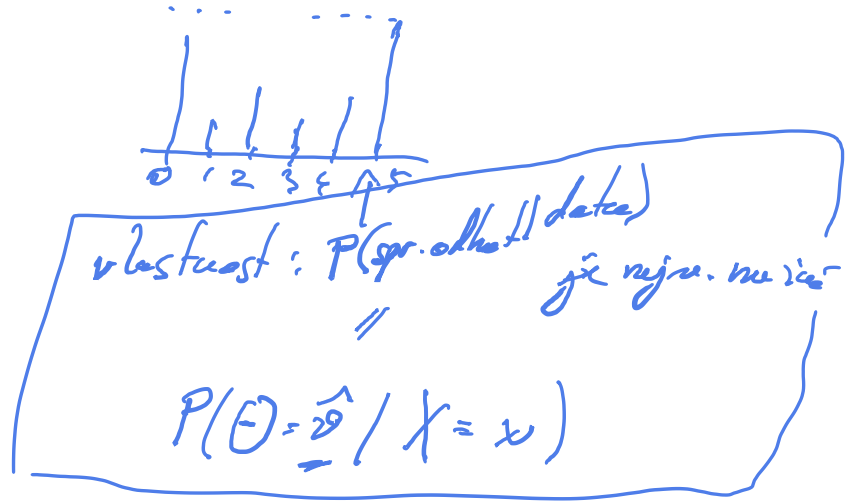
(sčítance s $f_\Theta(\vartheta') = 0$ považujeme za 0).

Bayesovské bodové odhady – MAP a LMS

MAP – Maximum A-Posteriori

Volíme $\hat{\vartheta}$ tak, aby maximalizovalo

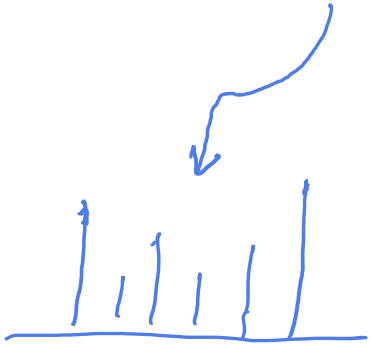
- $p_{\Theta|X}(\vartheta|x)$ v diskrétním případě
- $f_{\Theta|X}(\vartheta|x)$ ve spojitém případě

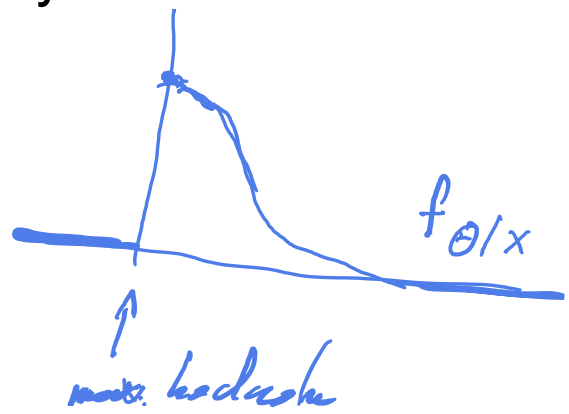


LMS – Least Mean Square

Též metoda podmíněné střední hodnoty.

- Volíme $\hat{\vartheta} = \mathbb{E}(\Theta \mid X = x)$





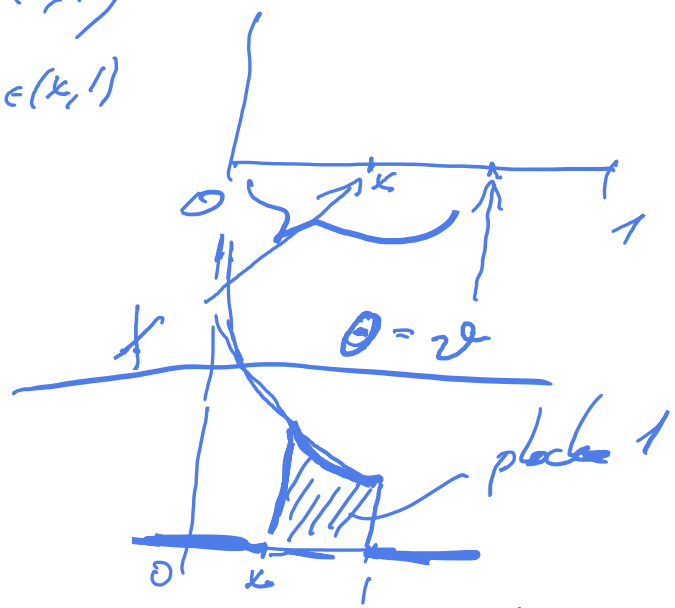
Příklad 1

Bayesovský klasifikátor spamů:

- vytvoříme seznam podezřelých slov (money, win, pharmacy, . . .)
- N.v. X_i popisuje, zda email obsahuje podezřelé slovo w_i.
- N.v. Θ popisuje, zda email je spam $\Theta = 1$ nebo ne $\Theta = 0$.
- Z předchozích emailů získáme odhady $p_{X|\Theta}$ a p_Θ
- Použijeme Bayesovu větu na výpočet $p_{\Theta|X}$

Příklad 2

Romeo a Julie se mají sejít přesně v poledne. Julie ale přijde pozdě o dobu popsanou náhodnou veličinou $X \sim U(0, \vartheta)$. Parametr ϑ modelujeme náhodnou veličinou $\Theta \sim U(0, 1)$. Co z naměřené hodnoty $X = x$ usoudíme o ϑ?



podana stó. kwadr.

$$\hat{\vartheta} := \mathbb{E}(\Theta \mid X = x) = \int_{-\infty}^{\infty} \vartheta f_{\Theta|X}(\vartheta|x)\, d\vartheta$$

$$= \int_{x}^{1} \vartheta \frac{1}{\vartheta |\ln x|}\, d\vartheta = \int_{x}^{1} \frac{1}{|\ln x|} = \frac{1-x}{|\ln x|}$$

wariant: mediana warunkowa w próbce $X_1, \ldots, X_4$

Příklad 3

Pozorujeme nezávislé náhodné veličiny $X = (X_1, \ldots, X_n)$, předpokládáme $X_i \sim N(\vartheta, \sigma_i^2)$ a ϑ je hodnota náhodné veličiny $\Theta \sim N(x_0, \sigma_0)$. Co z naměřených hodnot $X = x = (x_1, \ldots, x_n)$ usoudíme o ϑ?

$f_\Theta(\vartheta) = c_1 \exp\left(-\frac{(\vartheta - x_0)^2}{2\sigma_0^2}\right)$

apriorní distr. bude opět stejného typu (normální distr.)

$f_{X/\Theta}(x/\vartheta) = c_2 \exp\left(-\frac{(x_1-\vartheta)^2}{2\sigma_1^2}\right) \cdots \exp\left(-\frac{(x_n-\vartheta)^2}{2\sigma_n^2}\right)$

$= \text{konst} \cdot \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdots \sim$

$= c_4 \cdot \exp\left(-\frac{(\vartheta - m)^2}{2v}\right)$ ---- $N(m, v)$

Bayes:

$f_{\Theta/X}(\vartheta/x) = \frac{f_\Theta(\vartheta) \cdot f_{X/\Theta}(x/\vartheta)}{\int \cdots d\vartheta'}$ = $c_3 \cdot c_1 \exp\left(-\frac{(\vartheta - x_0)^2}{2\sigma_0^2}\right) \cdot c_2 \exp \ldots$

nezávisí na ϑ

$= d \cdot \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \vartheta)^2}{2\sigma_i^2}\right)$

$\vartheta^2 - 2m\vartheta + m^2 \to \sigma^2 - 2\sigma x_1 + x_1^2$

$-\frac{\vartheta^2}{2} \cdot \sum_{i=0}^{n} \frac{1}{\sigma_i^2} + \vartheta \cdot \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + \sum_{i=1}^{n} \frac{x_i^2}{2\sigma_i^2} = -\frac{(\vartheta - m)^2}{2v} + \text{konst.}$

$v = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$ $\quad m = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$

Příklad 4

Házíme mincí, pravděpodobnost, že padne panna je ϑ. Z n hodů padla panna v $X = k$ případech. Pokud naše apriorní distribuce byla $U(0, 1)$, jaká bude distribuce posteriorní?

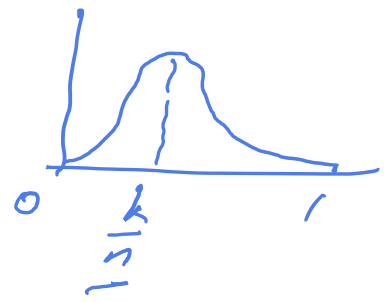
$f_\Theta(\vartheta) = 1 \quad (\vartheta \in (0,1))$

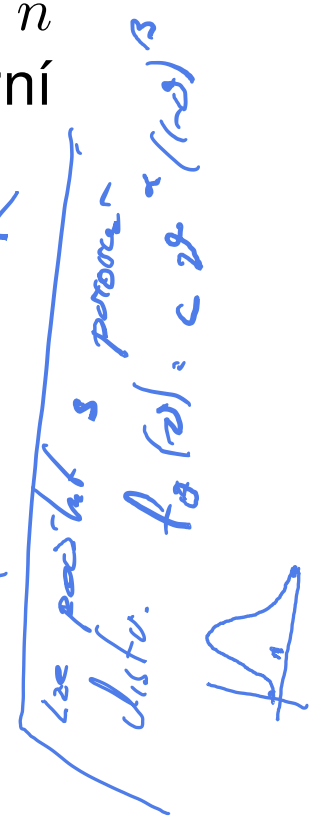
$P(X = k \mid \Theta = \vartheta) = \binom{n}{k} \cdot \vartheta^k (1-\vartheta)^{n-k}$

$P_{X|\Theta}(k, \vartheta)$

MAP $\Leftrightarrow$ ML

$\hat{\vartheta} = \frac{k}{n}$





$$f_{\Theta|X}(\vartheta \mid k) = \frac{f_\Theta(\vartheta) P_{X|\Theta}(k \mid \vartheta)}{\int \ldots} = c \cdot \vartheta^k (1-\vartheta)^{n-k}$$

c tak, aby $\int_0^1 c\,\vartheta^k (1-\vartheta)^{n-k} = 1$ — Beta distribuce

podm. stř. h.: $E(\Theta \mid X = k) = \int_0^1 c\,\vartheta \cdot \vartheta^k (1-\vartheta)^{n-k} = \frac{k+1}{n+2}$

$\frac{k}{n}$ a $\frac{1}{2}$

Střední hodnota a součet čtverců

Věta

Pro libovolnou n.v. Θ *je hodnota* $\mathbb{E}\big(\Theta - \hat{\vartheta}\big)^2$ *nejmenší pro* $\hat{\vartheta} = \mathbb{E}\big(\Theta\big)$.

$\hat{\vartheta}$ reálné číslo

Dk

$\operatorname{var} Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$

$\mathbb{E}Z^2 = \operatorname{var} Z + (\mathbb{E}Z)^2$

$Z := \Theta - \hat{\vartheta}$

$\mathbb{E}(\Theta-\hat{\vartheta})^2 = \operatorname{var}\Theta + (\mathbb{E}(\Theta)-\hat{\vartheta})^2$

$\operatorname{var} Z = \operatorname{var}\Theta$

$\mathbb{E}Z = \mathbb{E}\Theta - \hat{\vartheta}$

≥ 0

minimum pro $\mathbb{E}(\Theta) = \hat{\vartheta}$.

Podmíněná hodnota dává nejmenší součet čtverců

Věta

Bodový odhad $\hat{\vartheta} = \mathbb{E}(\Theta \mid X = x)$ *je nestranný a má nejmenší možnou hodnotu* $\mathbb{E}(\Theta - \hat{\vartheta})^2$.