Vytěžování dat

Filip Železný

Katedra počítačů
oddělení Inteligentní Datové Analýzy (IDA)

22. září 2014

Odhad rozdělení

Úloha:

- Vstup: data

$$D = \{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_m\},\ \vec{x}_i \in X\ (1 \leq i \leq m),\ m \in N$$

  náhodně, navz. nezávisle vybraná z rozdělení P_X na X

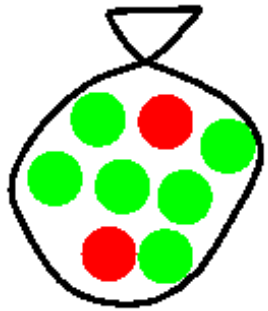
- Výstup: vzor

$$h \in \mathcal{L}$$

  reprezentující odhad P_X, tj. generativní model dat

Odhad rozdělení: příklad úlohy

2 druhy bonbónů v balíčku



Vybíráme náhodně (poslepu), dostaneme

Jaký je poměr (pravděpodobnost) zelených bonbónů v balíčku?

Odhad rozdělení: příklad úlohy (pokr.)

$$X = \{ \text{●}, \text{●} \}$$

P_X lze reprezentovat jedním číslem θ

$$P(\text{●}) = \theta,\ P(\text{●}) = 1 - \theta$$

Tedy prostor vzorů je reálný interval

$$\mathcal{L} = [0; 1]$$

Pozn.: ve skutečnosti konečná podmnožina $[0; 1]$, neboť reálná čísla se reprezentují konečným počtem číslic.

Odhad dle četnosti

Data

$$D = \{ \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \} \ (m = 10)$$

Odhad dle relativní četnosti:

$$P(\bullet) = \theta \approx \frac{9}{10}$$

Odůvodnění: četnost konverguje k pravděpodobnosti pro $m \to \infty$.

Odhad dle maximální věrohodnosti

- Obecnější metoda odhadu. Používá podmíněnou pravděpodobnost

$$P(D|\theta)$$

tj.: má-li parametr hodnotu θ, budeme data D pozorovat s touto pravděpodobností. Nazývá se věrohodnost (likelihood).

- Parametr odhadneme tak, že věrohodnost maximalizujeme

$$\theta^* = \arg\max_{\theta} P(D|\theta)$$

- Data $x_i \in D$ jsou vybírána navzájem nezávisle, tedy:

$$P(D|\theta) = \Pi_{i=1}^{m} P(x_i|\theta)$$

Věrohodnost: příklad

$$D = \{ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet\ \bullet \}\ (m = 10)$$

- Pro $\theta = P(\bullet) = 0.6$

  $$P(D|\theta) = P(\bullet|0.6)^9 \cdot P(\bullet|0.6)^1 = 0.6^9 \cdot 0.4 \approx 0.004$$

- Pro $\theta = P(\bullet) = 0.8$

  $$P(D|\theta) = P(\bullet|0.8)^9 \cdot P(\bullet|0.8)^1 = 0.8^9 \cdot 0.2 \approx 0.027$$

- Tedy $\theta = 0.8$ je věrohodnější než $\theta = 0.6$.

Obecně: jak najít θ, které věrohodnost maximalizuje?

Logaritmus věrohodnosti

Pro snazší výpočet používáme *logaritmus* věrohodnosti

$$L(D|\theta) = \log P(D|\theta)$$

tedy

$$L(D|\theta) = \log \Pi_{i=1}^{m} P(x_i|\theta) = \sum_{i=1}^{m} \log P(x_i|\theta)$$

V příkladě s bonbóny:

$$L(D|\theta) = \log \theta^z + \log(1-\theta)^c = c \log \theta + z \log(1-\theta)$$

kde c a z je počet červených resp. zelených bonbónů v datech

Hledání maxima věrohodnosti

θ maximalizuje věrohodnost právě tehdy, když maximalizuje její logaritmus. Hledáme maximum $L(D|\theta)$, tedy položíme

$$\frac{\mathbf{d}}{\mathbf{d}\theta} L(D|\theta) = 0$$

V příkladě s bonbóny:

$$\frac{\mathbf{d}}{\mathbf{d}\theta} \left(c \log \theta + z \log(1-\theta)\right) = \frac{c}{\theta} - \frac{z}{1-\theta} = 0$$

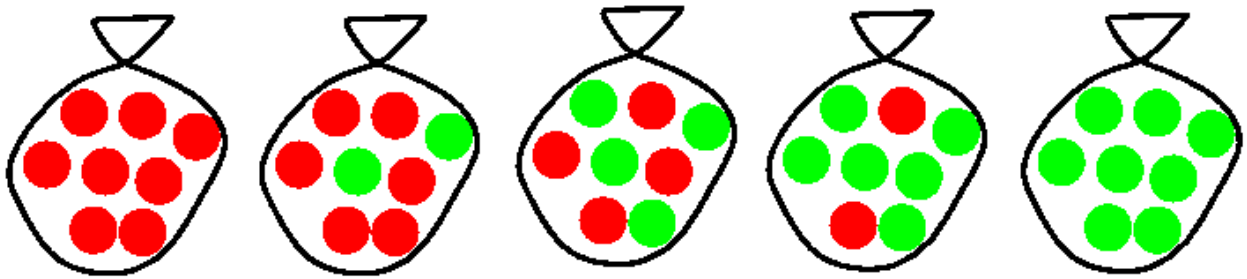
Řešení:

$$\theta = \frac{c}{c+z} = \frac{c}{m}$$

Tedy stejný výsledek jako u odhadu dle relativní četnosti. Metoda maximální věrohodnosti je ale obecnější - uvidíme dále.

Omezená množina vzorů

Tentokrát víme, že se vyrábí jen 5 typů balíčků:



1. 100% zelených
2. 75% zelených, 25% červených
3. 50% zelených, 50% červených
4. 25% zelených, 75% červených
5. 100% červených

Každý typ představuje jeden vzor pro generování dat (losování bonbónů), označme je po řadě $\mathcal{L} = \{h_1, h_2, h_3, h_4, h_5\}$.

Vzor s maximální věrohodnostní

Odhad dle četností již není použitelný, metoda maximální věrohodnosti je.

$$P(D|h_1) = 1^z \cdot 0^c$$
$$P(D|h_1) = 0.75^z \cdot 0.25^c$$
$$P(D|h_3) = 0.5^z \cdot 0.5^c$$
$$P(D|h_4) = 0.25^z \cdot 0.75^c$$
$$P(D|h_5) = 0^z \cdot 1^c$$

(z, c ... počet zelených resp. červených bonbónů v datech)

Dostáváme samé zelené: $D = \{$ ● ● ● $\ldots\}$, který vzor je nejvěrohodnější?

Apriorní pravděpodobnosti

Marginální rozdělení pravděpodobnosti $P_{\mathcal{L}}(h_i)$ vzorů může být známo před obdržením dat.
Např:

1. 100% zelených
2. 75% zelených, 25% červených
3. 50% zelených, 50% červených
4. 25% zelených, 75% červených
5. 100% červených

Apriorní pravděpodobnosti

Marginální rozdělení pravděpodobnosti $P_{\mathcal{L}}(h_i)$ vzorů může být známo před obdržením dat.
Např:

1. 100% zelených – 10% výroby
2. 75% zelených, 25% červených – 20% výroby
3. 50% zelených, 50% červených – 40% výroby
4. 25% zelených, 75% červených – 20% výroby
5. 100% červených – 10% výroby

Tedy

$$P_{\mathcal{L}}(h_1) = 0.1, P_{\mathcal{L}}(h_2) = 0.2, P_{\mathcal{L}}(h_3) = 0.4, P_{\mathcal{L}}(h_4) = 0.2, P_{\mathcal{L}}(h_5) = 0.1$$

Tyto pravděpodobnosti se nazývají *apriorní*.

Aposteriorní pravděpodobnost

Známe-li rozdělení

- $P_{\mathcal{L}}$
- a (po obdržení dat) $P(D|h_i)$ pro každý vzor h_i

můžeme podle Bayesova pravidla spočítat

$$P(h_i|D) = \frac{P(D|h_i)P_{\mathcal{L}}(h_i)}{P(D)}$$

$P(h_i|D)$ je *aposteriorní pravděpodobnost* vzoru h_i po obdržení dat D.

Jmenovatel

$$P(D) = \sum_{j=1}^{|\mathcal{L}|} P(D|h_j)P(h_j)$$

nezávisí na h_i. Z tohoto důvodu

$$\arg\max_{h_i} P(h_i|D) = \arg\max_{h_i} P(D|h_i)P_{\mathcal{L}}(h_i)$$

Odhad dle MAP

Metoda *maximální aposteriorní pravděpodobnosti* (MAP) vybírá vzor h

$$h = \arg\max_{h_i} P(h_i|D) = \arg\max_{h_i} P(D|h_i)P_{\mathcal{L}}(h_i)$$

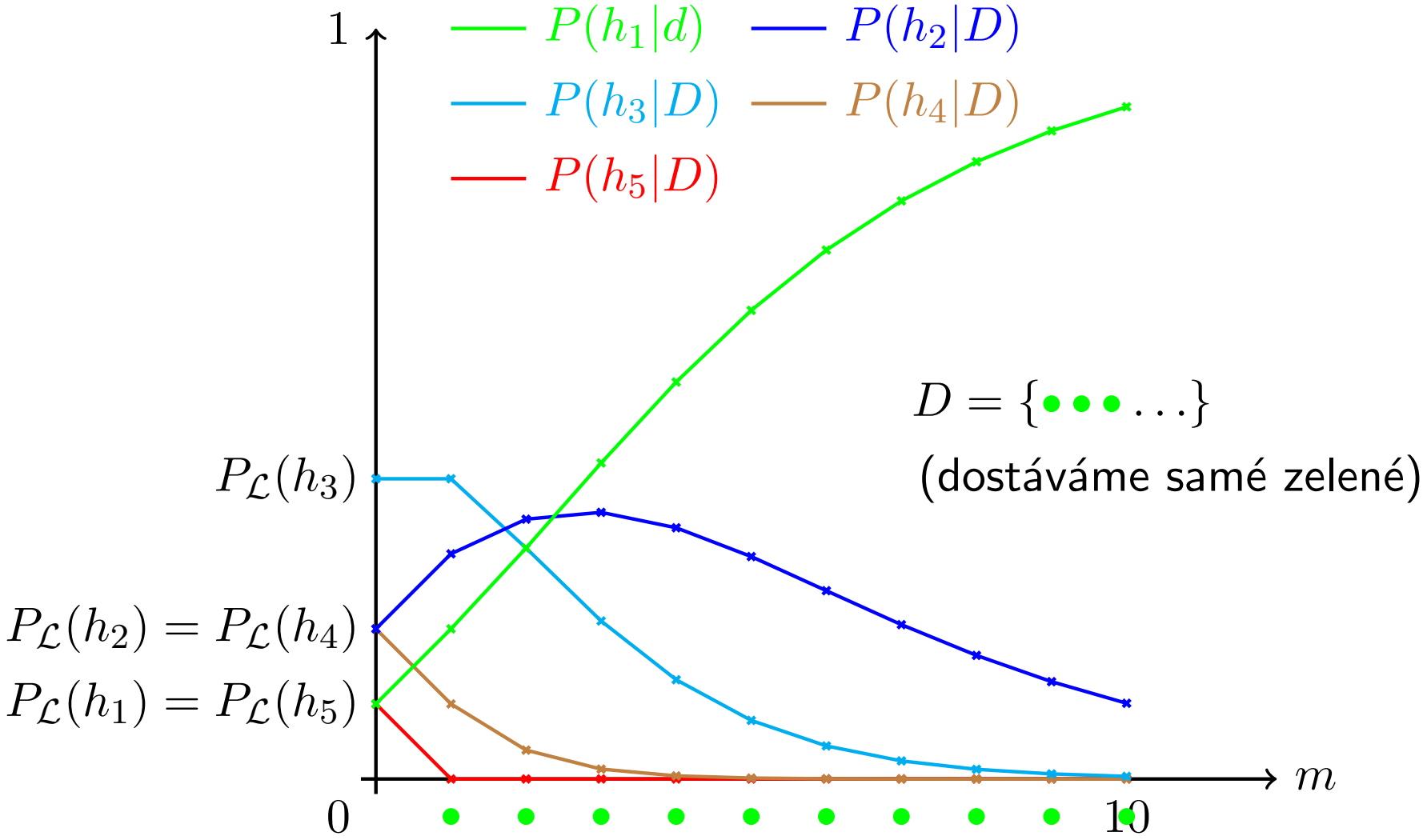
Srov. s metodou maximální věrohodnosti, kde

$$h = \arg\max_{h_i} P(D|h_i)$$

MAP tedy bere navíc úvahu informaci nesenou apriorním rozdělením $P_{\mathcal{L}}(h_i)$. Ta je významná pro malé množství dat, ale s rostoucím množstvím dat její význam klesá:

$$\arg\max_{h_i} P(D|h_i)P_{\mathcal{L}}(h_i) \rightarrow_{m\rightarrow\infty} \arg\max_{h_i} P(D|h_i)$$

Aposteriorní pravděpodobnost jako funkce množství dat



Odhad parametrů normálního rozdělení

Data $D = \{x_1, x_2, \ldots x_m\}$ vybrána navz. nezávisle z rozdělení

$$P_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Z dat odhadujeme parametry μ, σ. Aplikace metody max. věrohodnosti:

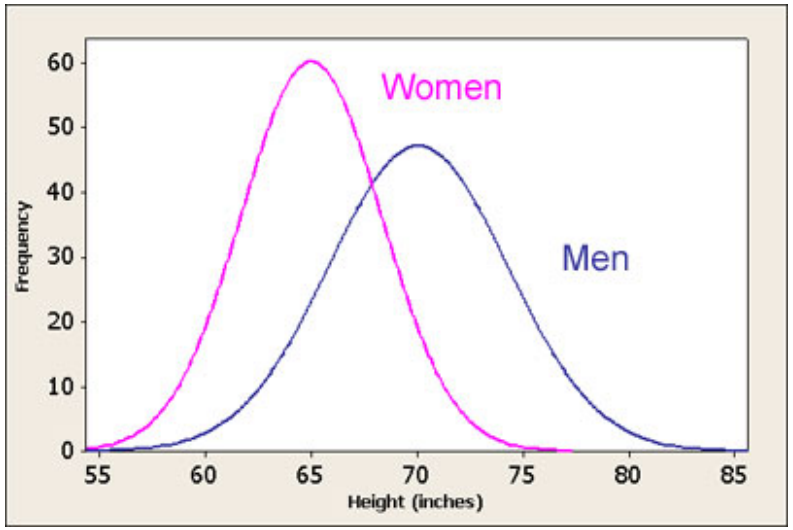
$$L(D|\mu,\sigma) = \sum_{i=1}^{m} \log P(x_i|\mu,\sigma) = m(-\log\sqrt{2\pi} - \log\sigma) - \sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\mathbf{d}}{\mathbf{d}\mu} L(D|\theta) = -\frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu) = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{m} x_i}{m}$$

$$\frac{\mathbf{d}}{\mathbf{d}\sigma} L(D|\theta) = -\frac{m}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{m}(x_i-\mu)^2 = 0 \Rightarrow \sigma = \sqrt{\frac{\sum_{i=1}^{m}(x_i-\mu)^2}{m}}$$

Směs normálních rozdělení (pokr.)

pohlaví	výška
žena	171
žena	164
muž	182
žena	169
muž	178
muž	184
...	



$$X = P \times V = \{\text{muž, žena}\} \times R^+$$

$$D = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots\} = \{(p_1, v_1), (p_2, v_2), (p_3, v_3), \ldots\}$$
$$= \{(\text{žena}, 171), (\text{žena}, 164), (\text{muž}, 182), \ldots\}$$

Směs normálních rozdělení (pokr.)

- Rozdělení výšek je součtem dvou normálních rozdělení (muži, ženy)
- Každé má svoji střední hodnotu a rozptyl

Rozdělení P_X na X lze vyjádřit jako

$$P_X(\vec{x}) = P_X([p, v]) = P_P(\text{muž})P_{V|P}(v|\text{muž}) + P_P(\text{žena})P_{V|P}(v|\text{žena})$$

$$P_{V|P}(v|\text{muž}) = \frac{1}{\sqrt{2\pi\sigma_{\text{muž}}}} \exp\left(-\frac{(x-\mu_{\text{muž}})^2}{2\sigma^2_{\text{muž}}}\right)$$

$$P_{V|P}(v|\text{žena}) = \frac{1}{\sqrt{2\pi\sigma_{\text{žena}}}} \exp\left(-\frac{(x-\mu_{\text{žena}})^2}{2\sigma^2_{\text{žena}}}\right)$$

Směs normálních rozdělení (pokr.)

Odhady dle maximální věrohodnosti, zvlášť pro každé pohlaví:

pohlaví	výška
žena	171
žena	164
žena	169

pohlaví	výška
muž	182
muž	173
muž	188

$$\mu_{\text{žena}} \approx \frac{\sum_{i=1}^{m} x_i}{m} = \frac{504}{3} = 168$$

$$\sigma_{\text{žena}} \approx \sqrt{\frac{3^2 + 4^2 + 1^2}{3}} \approx 2.94$$

$$\mu_{\text{muž}} \approx \frac{\sum_{i=1}^{m} x_i}{m} = \frac{543}{3} = 181$$

$$\sigma_{\text{muž}} \approx \sqrt{\frac{1^2 + 8^2 + 7^2}{3}} \approx 6.16$$

Skrytá proměnná

Víme, že v populaci jsou muži a ženy, ale proměnná (příznak) pohlaví v datech není.

pohlaví	výška
žena	171
žena	164
muž	182
žena	169
muž	178
muž	184

Jak nyní odhadnout P_X, tedy parametry $\mu_{\text{muž}}, \sigma_{\text{muž}}, \mu_{\text{žena}}, \sigma_{\text{žena}}$ a $P(\text{žena})$?

Algoritmus EM

1. ‘Nastřel’ počáteční hodnoty parametrů, např.

$$\mu_{\text{žena}} = 150,\ \sigma_{\text{žena}} = 10$$
$$\mu_{\text{muž}} = 200,\ \sigma_{\text{muž}} = 10$$
$$P(\text{žena}) = 0.5,\ P(\text{muž}) = 0.5$$

2. **Krok E (expectation)**: Se stanovenými parametry spočti pravděpodobnosti hodnot skryté proměnné pro každou instanci, např.

$$P(\text{žena}|171) = P(171|\text{žena})P(\text{žena})/P(171) =$$
$$\frac{1}{\sqrt{2\pi\sigma_{\text{žena}}}} \exp\left(-\frac{(171-\mu_{\text{žena}})^2}{2\sigma_{\text{žena}}^2}\right) \cdot 0.5/P(171)$$
$$= 0.01391 \cdot 0.5/P(171)$$

Algoritmus EM (pokr.)

2. **Krok E (pokr.)**

$$P(\text{muž}|171) = P(171|\text{muž})P(\text{muž})/P(171) =$$
$$\frac{1}{\sqrt{2\pi\sigma_{\text{muž}}}} \exp\left(-\frac{(171-\mu_{\text{muž}})^2}{2\sigma^2_{\text{muž}}}\right) \cdot 0.5/P(171)$$
$$= 0.00188 \cdot 0.5/P(171)$$

$$P(\text{žena}|171) + P(\text{muž}|171) = 1$$
$$P(\text{žena}|171) = \frac{0.01391 \cdot 0.5}{0.01391 \cdot 0.5 + 0.00188 \cdot 0.5} = 0.88$$
$$P(\text{muž}|171) = 1 - 0.88 = 0.12$$

Algoritmus EM (pokr.)

3. **Krok M (maximization)**: Se spočtenými pravděpodobnostmi pro hodnoty skrytých proměnných znovu odhadni parametry rozdělení

$$\mu_{\text{žena}} \leftarrow \frac{1}{N_{\text{žena}}} \sum_{i=1}^{m} P(\text{žena}|v_i) v_i$$

$$\sigma_{\text{žena}} \leftarrow \sqrt{\frac{1}{N_{\text{žena}}} \sum_{i=1}^{m} P(\text{žena}|v_i)(v_i - \mu_{\text{žena}})^2}$$

$$P(\text{žena}) \leftarrow \frac{1}{m} \sum_{i=1}^{m} P(\text{žena}|v_i)$$

   - $N_{\text{žena}} = \sum_{i=1}^{m} P(\text{žena}|v_i)$ ... normalizační konstanta, zaručuje, že součet $P(\text{žena}|v_i)$ přes všechny instance je 1.

   Analogicky spočteme pro muže.

4. Opakuj krokem 2 (dokud změny nejsou dostatečně malé)

Poznámka ke kroku M

- Pozorujte:

$$\frac{1}{N}\sum_{i=1}^{m} v_i$$

je průměrná výška v celém vzorku dat.

- Odhad v kroku M

$$\frac{1}{N_{\text{žena}}}\sum_{i=1}^{m} P(\text{žena}|v_i)v_i$$

je vlastně "vážený průměr". Vahou je pravděpodobnost $P(\text{žena}|v_i)$, že osoba s výškou v_i je žena. Přitom $\sum_{i=1}^{m} P(\text{žena}|v_i) = N_{\text{žena}}$.

- Analogicky pro odhad $\sigma_{\text{žena}}$
- Analogicky pro muže

Algoritmus EM (pokr.)

Konvergence algoritmu EM

iterace	$\mu_{\text{žena}}$	$\mu_{\text{muž}}$
0	150.000	200.000
1	167.007	181.972
2	167.951	181.956
(správné hodnoty)	168	181