Rekonstrukce diskrétního rozdělení psti metodou maximální entropie

Příklad Lze nalézt ‚četnosti' nepozorovaných stavů tak, abychom si „vymýšleli co nejméně"?

Nechť n_i, $i = 1, 2, \ldots, N$ jsou známé (absolutní) četnosti z neznámého diskrétního rozdělení pravděpodobnosti a nechť m_j, $j = 1, 2, \ldots, M$ jsou neznámé četnosti z tohoto rozdělení. Hledáme hodnoty m_j tak, aby entropie celého rozdělení byla maximální.

$$J = -\sum_{i=1}^{N} \frac{n_i}{n} \log \frac{n_i}{n} - \sum_{j=1}^{M} \frac{m_j}{n} \log \frac{m_j}{n} + \lambda (n - \sum_{i=1}^{N} n_i - \sum_{j=1}^{M} m_j),$$

neznámé jsou m_j, $j = 1, 2, \ldots, M$, n a λ.

Výsledek:

$$m_j = m = n_0 \, e^{-H_0}, \quad \text{kde} \quad n_0 = \sum_{i=1}^{N} n_i \quad \text{a} \quad H_0 = -\sum_{i=1}^{N} \frac{n_i}{n_0} \log \frac{n_i}{n_0}$$

$$H = H_0 + \log \frac{n}{n_0}, \quad \text{kde} \quad n = n_0 + M\, m$$

Rekonstrukce spojitého rozdělení ze známých momentů

Příklad

Nechť μ je střední hodnota a σ rozptyl neznámého rozdělení pravděpodobnosti s hustotou $f(x)$. Jak vypadá $f(x)\colon \mathbb{R} \mapsto [0,1]$, která má za těchto podmínek maximální diferenciální entropii?

$$Q = -\int_{-\infty}^{\infty} f(x)\log f(x)\mathrm{d}x + \lambda_0\left(1 - \int_{-\infty}^{\infty} f(x)\mathrm{d}x\right) + \lambda_1\left(\mu - \int_{-\infty}^{\infty} x f(x)\mathrm{d}x\right) + \\ + \lambda_2\left(\sigma^2 - \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x\right) \tag{1}$$

$g_1(x) = 1$ $g_2(x) = x$ $g_3(x) = (x-\mu)^2$

$\frac{\partial Q}{\partial f} = 0$ $\frac{\partial Q}{\partial \lambda_i} = 0$ $i = 0,1,2$

Přehled užité matematiky

Výsledek

$$f(x) = \frac{1}{T}\, e^{a_1 x + a_2 x^2} = \cdots = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

Důkaz

1. pomocí Gibbsovy nerovnosti (viz dále)
2. pomocí variačního počtu z rovnice (1) ⊛1

Zobecnění předchozího výsledku

Nechť $g_i(x)$, $i = 1, 2, \ldots, K$ jsou známé funkce a $c_i = \int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x$ jsou známé hodnoty. Rekonstruujeme z nich neznámou hustotu $f(x)$ tak, aby maximalizovala entropii $H(x)$.

odjinud

Výsledek

$$f(x) = \frac{1}{T} e^{-\sum_{j=1}^{K} a_j g_j(x)}$$

a

$$H(x) = -\int_{-\infty}^{\infty} f(x) \log f(x) \mathrm{d}x = \log T + \sum_{j=1}^{K} a_j c_j$$

Důkaz Nechť $\phi(x)$ je hustota rozdělení, které má větší entropii než $f(x)$ a vyhovuje podmínkám (má stejné hodnoty c_i). Platí Gibbsova nerovnost

$$c_i = \int_{-\infty}^{\infty} g_i(x) \phi(x) \mathrm{d}x$$

$$H_\phi(x) = -\int_{-\infty}^{\infty} \phi(x) \log \phi(x) \leq -\int_{-\infty}^{\infty} \phi(x) \log f(x) \mathrm{d}x =$$

$H_\phi(x) > H_f(x)$

$$= \int_{-\infty}^{\infty} \phi(x) \left(\log T + \sum_{j=1}^{K} a_j g_j(x) \right) \mathrm{d}x = \log T + \sum_{j=1}^{K} a_j c_j = H_f(x)$$

$$\sum_j a_j \int_{-\infty}^{\infty} \phi(x) g_j(x) \mathrm{d}x = \sum_j a_j c_j$$

což je spor.

Výpočet a_i, $i = 1, 2, \ldots, K$

Zavedeme funkci

partition function

$$Z(a_1, \ldots, a_K) = T = \int_{-\infty}^{\infty} e^{-\sum_{j=1}^{K} a_j g_j(x)} \mathrm{d}x$$

Platí

$$\frac{\partial Z}{\partial a_i} = \int_{-\infty}^{\infty} -g_i(x) e^{-\sum_{j=1}^{K} a_j g_j(x)} \mathrm{d}x = -Z \int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x = -Z\, c_i$$

Takže a_i jsou řešením soustavy nelineárních rovnic

$$-\frac{1}{Z}\frac{\partial Z}{\partial a_i} = c_i, \qquad i = 1, 2, \ldots K, \tag{3}$$

kde neznámé jsou a_i, kde c_i jsou dány a kde $g_i(x)$ jsou známé funkce.

Pozn

- soustava (3) je většinou obtížně řešitelná analyticky

⊛2 použijte pro důkaz výsledku (2) na str 28.

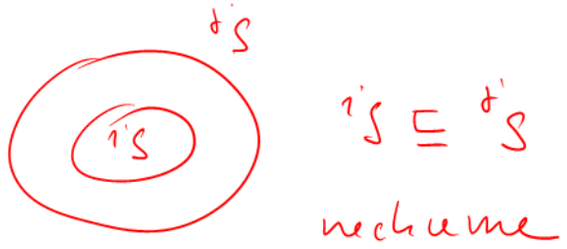
Identifikace struktury

Identifikace struktury

Nejlepší dekompozice systému $\mathbf{S}$ na množinu podsystémů $G = \{{}^1\mathbf{S}, {}^2\mathbf{S}, \ldots, {}^q\mathbf{S}\}$

Podproblémy

1. systematické generování rekonstrukčních hypotéz G
   - neredundance $\quad \forall i, j \in \{1, 2, \ldots, q\} : {}^i\mathbf{S} \not\sqsubseteq {}^j\mathbf{S}$
   - pokrytí $\quad \bigcup_i {}^i\mathbf{S} = \mathbf{S}$
2. Rekonstrukce celkového systému $\mathbf{S}^*$ z G.
3. Vyhodnocení rekonstrukční chyby $\Delta(G) = L(\mathbf{S}, \mathbf{S}^*)$



($\mathbf{S}$. . . množina proměnných systému)

Bez podmínky nerendundance a pokrytí by byl počet hypotéz

$$2^{2^n - 1} - 1.$$

(n – počet proměnných, $n = 2 \Rightarrow$ 7 hypotéz, $n = 4 \Rightarrow$ 32767 hypotéz)

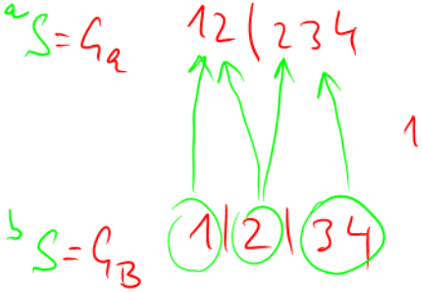
Zjemnění rekonstrukční hypotézy

Rekonstrukční hypotéza $G = \{{}^1\mathbf{S}, {}^2\mathbf{S}, \ldots, {}^q\mathbf{S}\}$

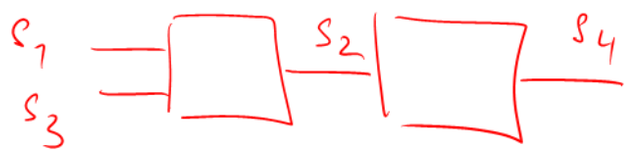
Zjemnění rekonstrukční hypotézy $\prec$

G_i je zjemněním G_j když pro každé ${}^x\mathbf{S} \in G_i$ existuje ${}^y\mathbf{S} \in G_j$ takové, že ${}^x\mathbf{S} \sqsubseteq {}^y\mathbf{S}$. Značíme

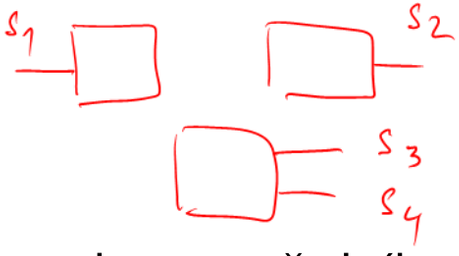
$$G_i \prec G_j$$



Bezprostřední zjemnění $\overset{*}{\prec}$



G_i je bezprostředním zjemněním G_j jestliže neexistuje G_k takové, že



$$G_i \prec G_k \quad \text{a} \quad G_k \prec G_j$$

Relace bezprostředního zjemnění tvoří svaz na množině všech rekonstrukčních hypotéz.

Pozn: nejde o zjemňování rozkladu, rekonstrukční hypotézy netvoří rozklad (proměnné se opakují).

Generátor rekonstrukčních hypotéz

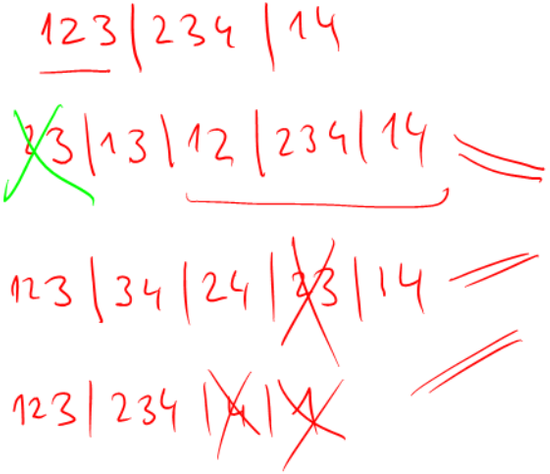
Vstup: Rekonstrukční hypotéza $G = \{{}^1\mathbf{S}, {}^2\mathbf{S}, \ldots, {}^q\mathbf{S}\}$

Výstup: Všechna bezprostřední zjemnění G

Procedura (prohledávání svazu $(G, \overset{*}{\prec})$)

1. Pro $i = 1, 2, \ldots, q$ dělej kroky 2,3.
2. Jestliže $|{}^i\mathbf{S}| \geq 2$ potom nahraď ${}^i\mathbf{S}$ množinou všech podmnožin ${}^i\mathbf{S}$ o velikosti $|{}^i\mathbf{S}| - 1$.
3. Po odstranění redundantních podmnožin v každém rozkladu dostaneme seznam bezprostředních zjemnění G.

Pozn: může se opakovat na stejné úrovni zjemnění.



$$G = \{\{1,2,3\},\{2,3,4\},\{1,4\}\} = \{{}^1\mathbf{S}, {}^2\mathbf{S}, {}^3\mathbf{S}\}$$

$$G_1 = \{\{1,2\},\{1,3\},\{2,3,4\},\{1,4\}\}$$

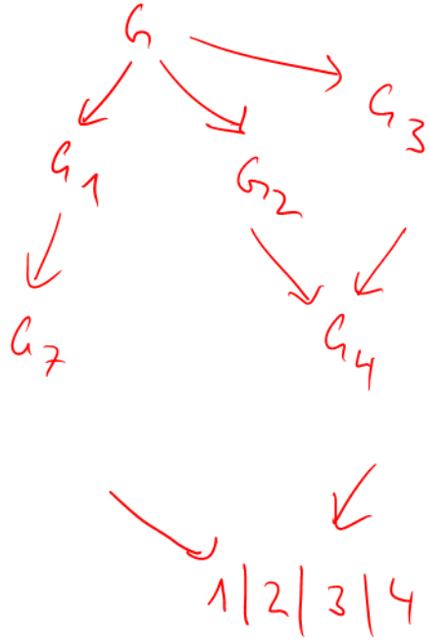
$$G_2 = \{\{1,2,3\},\{2,4\},\{3,4\},\{1,4\}\}$$

$$G_3 = \{\{1,2,3\},\{2,3,4\}\}$$

$$G_4 = \{\{1,2\},\{1,3\},\{2,3\},\{3,4\},\{2,4\},\{1,4\}\}$$

$$G_5 = \{\{1,2\},\{1,3\},\{2,3,4\}\}$$

$$G_6 = \{\{1,2,3\},\{2,4\},\{3,4\}\}$$



Prohledávání svazu zjemnění

- Rekonstrukční chyba Δ monotonně roste podél každé větve svazu zjemnění: Je-li $G_i \succ G_k \succ G_j$ potom $\Delta(G_i) \leq \Delta(G_k) \leq \Delta(G_j)$
- **Nejjednodušší postup:** Hledáme cestu, podle které Δ roste nejpomaleji
- **Složitější postup:** Na každé úrovni nás zajímá množina řešení s nejmenším Δ
  - Není nutno expandovat všechny větve: postačí metoda větví a mezí

[Narenda&Fukunaga 1977]