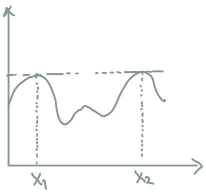
EM-Algorithmus

Freitag, 12. Januar 2018 10:04

- Zufallsexperiment: $(\mathcal{X}, p)$
  - $\mathcal{X}$ endliche Menge (Ergebnismenge)
  - $p: X \to [0,1]$ mit $\sum_{x \in X} p(x) = 1$ (Wahrscheinlichkeitsverteilung)
- $\mathcal{M}(X)$: Menge aller Wahrscheinlichkeitsverteilungen über X
- $\mathcal{M} \subseteq \mathcal{M}(X)$: Wahrscheinlichkeitsmodell

Korpora & Korpuswahrscheinlichkeiten

- Korpus $h: X \to \mathbb{R}_{\geq 0}$ Bsp. X = (engl. Sätze × sp. Sätze)
  - h(x) auch als Frequenz bezeichnet — anschaulich: absolute Häufigkeit
  - $\{x \mid h(x) > 0\}$ ist endlich und $\sum h(x) \neq 0$ (mind. 1 Satzpaar soll drin sein)
- Likelihood: Maß wie gut Korpus zu Wahrscheinlichkeitsverteilung passt
  - 2 Argumente: meine Wahrscheinlichkeitsverteilung, Korpus
  - $L(h,p) := \prod_{x \in X} p(x)^{h(x)}$
- Maximum-Likelihood Schätzer: — argmax



$\Rightarrow \operatorname{argmax} f(x) = \{x_1, x_2\}$

  - $\mathcal{M} \subsetneq \mathcal{M}(X)$ — nur ein Element, keine Menge wie rechts
  - $\text{mle}(h, \mathcal{M}) := \operatorname{argmax}_{p \in \mathcal{M}} L(h,p) \in \mathcal{M}$
  - mle ist Wahrscheinlichkeitsverteilung, für die Likelihood maximal wird
  - $L(h, \text{mle}(h,\mathcal{M})) \geq L(h,p)$

- relative Häufigkeit:
  - $\text{rfe}(h): X \to [0,1], \quad x \mapsto \frac{h(x)}{|h|}$ mit $|h| = \sum_{x \in X} h(x)$
  - $\sum \text{rfe}(h)(x) = \sum \frac{h(x)}{|h|} = \frac{\sum h(x)}{\sum h(x)} = 1 \Rightarrow$ rfe(h) ist Wahrscheinlichkeitsverteilung
- Satz: Sei X eine Ergebnismenge und h ein X-Korpus.
  - rfe(h) ist Wahrscheinlichkeitsverteilung über X, d.h. $\text{rfe}(h) \in \mathcal{M}(X)$
  - $\text{rfe}(h) = \text{mle}(h, \mathcal{M}(X))$ (unbrauchbar, da $\mathcal{M} \subsetneq \mathcal{M}(X)$)
- $\mathcal{M} \subsetneq \mathcal{M}(X)$: Sei $\mathcal{M} = \{ p_1 \times p_2 \mid p_1, p_2 \in \mathcal{M}(\{K,Z\}) \}$.

  $\Rightarrow \mathcal{M}(X_1 \times X_2) \setminus \mathcal{M} \neq \emptyset$, wenn $p(K,K) = p(Z,Z) = 0$ und $p(K,Z) = p(Z,K) = 0{,}5$

Beweis: $\exists p_1, p_2 \in \mathcal{M}(\{K,Z\}): \quad p = p_1 \times p_2$

$p(K,K) = (p_1 \times p_2)(K,K) = p_1(K,K) \cdot p_2(K,K) = 0$

⋮

- Bsp.: 30 maliger Münzwurf

Korpora:

ⓗ	K	Z	h_1
K	5	10	15
Z	5	10	15
h_2	10	20	30

$30 = |h_1| = |h_2|$

$rfe(h_1)(K) = \frac{15}{30} = \frac{1}{2}$

$rfe(h_2)(K) = \frac{10}{30} = \frac{1}{3}$

Korpora mit unvollständigen Daten

Bsp.: Person A wirft Münze 2mal

Ergebnismenge: $X = \{(K,K),(K,Z),(Z,K),(Z,Z)\}$

A teilt absolute Häufigkeit von Kopf B mit

B beobachtet 0,1,2

Beobachtungsmenge $Y = \{0,1,2\}$

Beobachtungsfunktion: **yield**: $X \to Y$

Bsp.: $A(\text{Gewinn}) = \{KK, ZZ\}$
$\text{yield}(KK) = \text{Gewinn}$

Analysator: $\text{yield}^{-1} = A: Y \to \mathcal{P}(X)$

$$A(y) = \{x \in X \mid \text{yield}(x) = y\}$$

Korpuswahrscheinlichkeit von h unter p:

$$L(h,p) = \prod_{y \in Y} \Big(\sum_{x \in A(y)} p(x)\Big)^{h(y)}$$

$$\text{mle}(h,\mathcal{M}) = \underset{p \in \mathcal{M}}{\text{argmax}}\ L(h,p)$$

Satz: Sei $q_0, q_1, \dots$ eine durch den EM-Algorithmus berechnete Sequenz von Wahrscheinlichkeitsverteilung über X. Dann gilt:

$$L(h,q_0) \leq L(h,q_1) \leq \dots \leq L(h, \text{mle}(h,\mathcal{M}))$$

EM- Algorithmus:

Input: Y- Korpus h
Analysator $A: Y \to \mathcal{P}(X)$
Wahrscheinlichkeitsmodell $\mathcal{M} \subseteq \mathcal{M}(X)$ über X
$q_0 \in \mathcal{M}$ mit $q_0(x) > 0 \ \forall x \in X$

Output: Sequenz $q_1, q_2, \dots$ von Elementen aus $\mathcal{M}$

für jedes $i = 1,2,3,\dots$

E-Schritt: berechne den Y-Korpus h_i

$$h_i(x) = h(\text{yield}(x)) \cdot \frac{q_{i-1}(x)}{\sum\limits_{x' \in A(\text{yield}(x))} q_{i-1}(x')}$$

M-Schritt: berechne den Maximum-Likelihood-Schätzer von h_i und $\mathcal{M}$

$$q_i = \operatorname*{argmax}_{p \in \mathcal{M}} L(h_i, p) = \text{mle}(h_i, \mathcal{M})$$

print q_i

praktisches Verfahren:

1. Angabe des Analysators

Beispiel (Übung AGS 10.16)

$A(\text{Gewinn}) = \{(r,g),(r,b),(g,r),(g,b),(b,r),(b,g)\}$
$A(\text{kein Gewinn}) = \{(r,r),(g,g),(b,b)\}$

2. Angabe der Korpora (absolute Häufigkeiten)

$h(\text{Gewinn}) = 16$ (gegeben)
$h(\text{kein Gewinn}) = 24$ (berechnet)

3. Angabe von q_0 (vollständig)
   - initiale Wahrscheinlichkeit nahezu vollständig gegeben (fehlende ergänzen zu 1)
   - unabhängiges Produkt nutzen

$\left.\begin{array}{l} q_0^1(r) = 1/2 \\ q_0^1(g) = 1/3 \end{array}\right\} \Rightarrow q_0^1(b) = 1 - 1/2 - 1/3 = 1/6$

$\left.\begin{array}{l} q_0^1(r) = 1/4 \\ q_0^1(g) = 1/2 \end{array}\right\} \Rightarrow q_0^1(b) = 1 - 1/4 - 1/2 = 1/4$

$q_0^1(r,r) = 1/2 \cdot 1/4 = 1/8$ $\quad q_0^1(r,g) = 1/4$ $\quad q_0^1(r,b) = 1/8$
$q_0^1(g,r) = 1/3 \cdot 1/4 = 1/12$ $\quad q_0^1(g,g) = 1/6$ $\quad q_0^1(g,b) = 1/12$
$q_0^1(b,r) = 1/6 \cdot 1/4 = 1/24$ $\quad q_0^1(b,g) = 1/12$ $\quad q_0^1(b,b) = 1/24$

4. E-Schritt:

Berechne Korpora h_1 mit

$$h_1(x) = h(\text{yield}(x)) \cdot \frac{q_0(x)}{\sum_{x' \in A(\text{yield}(x))} q_0(x')}$$

$$h_1(r,r) = h(\text{kein Gewinn}) \cdot \frac{q_0(r,r)}{\sum_{x' \in A(\text{yield}(x))} q_0(x')}$$
$$= h(\text{kein Gewinn}) \cdot \frac{q_0(r,r)}{q_0(r,r) + q_0(g,g) + q_0(b,b)}$$
$$= 24 \cdot \frac{1/8}{1/8 + 1/6 + 1/24} = 9$$

$h_1(g,r) = 2$ $\quad h_1(r,g) = 6$ $\quad h_1(r,b) = 3$
$h_1(b,r) = 1$ $\quad h_1(g,g) = 12$ $\quad h_1(g,b) = 2$
$\quad h_1(b,g) = 2$ $\quad h_1(b,b) = 3$

5. M-Schritt: Berechne die Teilkorpora h_1^1 und h_1^2

$$h_1^1(r) = h_1(r,r) + h_1(r,g) + h_1(r,b) = 9 + 6 + 3 = 18$$

$$h_1^2(r) = h_1(r,r) + h_1(g,r) + h_1(b,r) = 9 + 2 + 1 = 12$$

→ Marginalisierung in Matrixschreibweise

$x_1 \backslash x_2$	r	g	b	Σ	
r	9	6	3	18	$= h_1^1(r)$
g	2	12	2	16	$= h_1^1(g)$
b	1	2	3	6	$= h_1^1(b)$
Σ	12	20	8	40	
	$= h_1^2(r)$		$= h_1^2(b)$		

6. Schätzung der Wahrscheinlichkeitsverteilungen q_1^1 und q_1^2

$$q_1^1(r) = \frac{h_1^1(r)}{h_1^1(r) + h_1^1(g) + h_1^1(b)} = \frac{18}{18+16+6} = \frac{9}{20}$$

$q_1^1(g) = \frac{16}{40} = \frac{2}{5}$ $\quad q_1^1(b) = \frac{6}{40} = \frac{3}{20}$

$q_1^2(r) = \frac{12}{40} = \frac{3}{10}$ $\quad q_1^2(g) = \frac{20}{40} = \frac{1}{2}$ $\quad q_1^2(b) = \frac{8}{40} = \frac{1}{5}$