Vorlesung 13

Relative Entropie

Definition: Seien ρ und π Wahrscheinlichkeitsverteilungen mit Gewichten $\rho(a)$ und $\pi(a)$, $a \in S$. Dann ist die *relative Entropie* von ρ bzgl. π definiert als

$$D(\rho\|\pi) := \sum_{a\in S} \rho(a) \log \frac{\rho(a)}{\pi(a)} ,$$

wobei die Summanden mit $\rho(a) = 0$ gleich 0 gesetzt werden.

Eine Interpretation der relativen Entropie:

Man denke sich einen zufälligen Buchstaben mit Verteilung ρ
mit einem Shannon-Code codiert, der nicht der Verteilung ρ,
sondern der Verteilung π angepasst ist,
also mit Codewortlängen

$$-\log \pi(a) \leq \ell(a) < -\log \pi(a) + 1.$$

Dann ändert sich die erwarteten Codelänge
im Vergleich zu dem an ρ angepassten Shannon-Code
(bis auf höchstens 1) um

$$-\sum_a \rho(a) \log \pi(a) - \Big(-\sum_a \rho(a) \log \rho(a) \Big) = D(\rho \| \pi).$$

Satz: ("Informationsungleichung") $D(\rho\|\pi) \geq 0$.

Beweis: Wieder verwenden wir die Abschätzung $\log x \leq c \cdot (x-1)$ mit passendem $c > 0$:

$$D(\rho\|\pi) = -\sum_{a:\rho(a)>0} \rho(a) \log \frac{\pi(a)}{\rho(a)}$$

$$\geq -\sum_{a:\rho(a)>0} \rho(a)\, c \cdot \left(\frac{\pi(a)}{\rho(a)} - 1\right)$$

$$= -c\Big(\sum_{a:\rho(a)>0} \pi(a) - \sum_a \rho(a)\Big) \geq 0. \quad \square$$

Bemerkung: Aus $D(\rho||\pi) = 0$ folgt $\rho = \pi$.

In der Tat: In der Ungleichung $\log x \leq c(x-1)$

besteht abgesehen für $x = 1$ strikte Ungleichung.

Also folgt aus

$$-\sum_{a:\rho(a)>0} \rho(a) \log \frac{\pi(a)}{\rho(a)} = -c \sum_{a:\rho(a)>0} \rho(a)\Big(\frac{\pi(a)}{\rho(a)} - 1\Big),$$

dass $\pi(a) = \rho(a)$ für alle a mit $\rho(a) > 0$.

Daraus folgt $\sum_{a:\rho(a)>0} \pi(a) = 1,$

also auch $\pi(a) = \rho(a)$ für alle a mit $\rho(a) = 0$. $\square$

Zusammenfassend ergibt sich der

Satz (von der relativen Entropie):

Die relative Entropie $D(\rho||\pi)$ ist nichtnegativ,

und verschwindet genau für $\rho = \pi$.

In den folgenden Beispielen

benutzen wir den Satz in der Gestalt

$$(*) \quad -\sum_a \rho(a) \log \rho(a) \leq -\sum_a \rho(a) \log \pi(a)$$

mit Gleichheit genau für $\rho = \pi$.

Wir sehen:

Hat X die Verteilung ρ,

so liefert jede Wahl von π eine Schranke für $\mathbf{H}[X]$,

mit Gleichheit genau für $\rho = \pi$.

Beispiel: Vegleich mit der uniformen Verteilung:

Sei S endlich mit n Elementen

und sei $\pi(a) = 1/n$ für alle $a \in S$.

Dann folgt aus $(*)$:

$$\mathbf{H}[X] \leq -\sum_a \rho(a) \log(\frac{1}{n}) = \log n \, .$$

$$\boxed{\mathbf{H}[X] \leq \log n.}$$

Gleichheit gilt genau im Fall der uniformen Verteilung,

sie maximiert auf S die Entropie. $\square$

Beispiel: Vergleich mit (verschobener) geom. Verteilung:

Sei nun $S = \{0, 1, 2, \ldots\}$, und $\pi(k) := 2^{-k-1}$

Dann folgt aus $(*)$:

$$\mathbf{H_2}[X] \leq \sum_a \rho(a) \log_2(-2^{-k-1})$$

$$= \sum_{k=0}^{\infty} (k+1)\rho(k) = (\mathbf{E}[X] + 1) \; .$$

Gleichheit gilt für $\rho(k) = 2^{-k-1}$, dann ist $\mathbf{H}_2[X] = 2$. Also:

Unter allen $\mathbb{N}_0$-wertigen ZV'en mit EW ≤ 1

hat das X mit Gewichten 2^{-k-1} die größte Entropie,

nämlich $\mathbf{H}_2[X] = 2$. $\square$

Beispiel: Vergleich mit einer “Gibbsverteilung”:

Gegeben sei $u : S \to \mathbb{R},\ \beta \geq 0$.

Wir definieren die Gewichte $\pi(a) := e^{-\beta u(a)}/z$ mit

$$z := \sum_{a \in S} e^{-\beta u(a)}$$ (Annahme: $z < \infty$.)

Die Abschätzung $(*)$ ergibt

$$\mathbf{H}_e[X] \leq \beta\, \mathbf{E}[u(X)] + \ln z\ .$$

Ist $\mathbf{E}[u(X)] \leq \sum_a u(a)\pi(a)$, so folgt

$$\mathbf{H}_e[X] \leq \beta \sum_a u(a)\pi(a) + \ln z = -\sum_a \pi(a) \ln \pi(a),$$

mit Gleichheit genau dann, wenn X die Verteilung π hat.

Zusammengefasst:

Unter allen Zufallsvariablen mit

Erwartungswert $\leq \sum_a u(a) e^{-\beta u(a)} / z$

hat diejenige die größte Entropie,

die die Verteilungsgewichte $e^{-\beta u(a)} / z$ hat.

Die Verteilung mit diesen Gewichten heißt *Gibbsverteilung*

zum Potenzial u mit Parameter β. $\square$

Beispiel:

n-maliges Würfeln zu den Gewichten $p = (p_1, \ldots, p_r)$.

Die Besetzungszahlen $Z^{(n)} = (Z_1^{(n)}, \ldots, Z_r^{(n)})$

sind multinomial$(n; p)$-verteilt.

Wie (un-)wahrscheinlich ist das Ereignis $\{Z^{(n)} = b\}$

für großes n und untypisches k?

$$\mathbf{P}_p(Z^{(n)} = b) \approx ?$$

Idee: Vergleiche dies mit der Wahrscheinlichkeit von b

unter derjenigen Gewichtung w, für die b typisch ist:

$$w_j := \frac{b_j}{n}, \quad j = 1, \ldots, r.$$

$$\frac{\mathbf{P}_p(Z^{(n)} = b)}{\mathbf{P}_w(Z^{(n)} = b)} = \prod_{j=1}^{r} \left(\frac{p_j}{w_j}\right)^{b_j} = 2^{-\sum_{j=1}^{r} b_j \log_2 \frac{w_j}{p_j}} = 2^{-nD(w\|p)}$$

Unter der Annahme $w = b/n \to \alpha$ gilt (siehe Nachtrag unten):
$\mathbf{P}_w(Z^{(n)} = b)$ nimmt nur wie eine Potenz von $1/n$ ab. Damit wächst
$\log_2 \mathbf{P}_w(Z^{(n)} = b)$ nicht mehr als logarithmisch, und es folgt:

$$\boxed{\log_2 \mathbf{P}_p(Z^{(n)} = b) \sim -nD(\tfrac{b}{n}\|p)}$$

Nachtrag: Mit der Stirling-Formel $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$ folgt

$$\frac{n!}{b_1! \cdots b_r!} \sim (2\pi)^{\frac{1-r}{2}} \left(\frac{n}{b_1}\right)^{b_1} \cdots \left(\frac{n}{b_r}\right)^{b_r} \sqrt{\frac{n}{b_1 \cdots b_r}}$$

$$= (2\pi)^{\frac{1-r}{2}} w_1^{-b_1} \cdots w_r^{-b_r} \sqrt{\frac{1}{w_1 \cdots w_r}}\, n^{\frac{1-r}{2}}$$

Für $w = w(n) \to \alpha$ ist somit
$\mathbf{P}_w(Z^{(n)} = b) \sim \mathrm{const}(\alpha) \cdot \mathrm{n}^{\frac{1-\mathrm{r}}{2}}$

$S = k \log W$

Entropie =
k mal
Logarithmus der
Wahrscheinlichkeit

Ludwig Boltzmann
1844-1906

Grabmal am
Wiener
Zentralfriedhof