Kolmogoroffkomplexität

Teil 3

Informationstheorie und Kodierung

Torsten Steinbrecher

Informationstheorie

Information ist …

- Δ Wahlfreiheit beim Sender
- Δ Unbestimmtheit beim Empfänger

Information ist nicht …

- Länge der Nachricht an sich
- Aussage (Semantik) der Nachricht an sich

Kodierung

Wie kompakt können wir

Nachrichten theoretisch fassen?

- Idee: messen Wahlfreiheit / Unbestimmtheit

Wie kommen wir diesem Grenzwert möglichst nahe?

- optimale Codes

Wie können wir Nachrichten *noch* kürzer fassen?

- verlustbehaftete Kompression (siehe Buch)

Beispiel

heute gabs kartoffeln und l lachs, lamm, leber, labskaus, letscho

heute gabs kartoffeln und q quark

l(p(q)) > l(p(l))

Formalisierung I

sei X Zufallsvariable über alle Buchstaben in A

und x=X(ω) der nächste Buchstabe in der Leitung

je geringer seine Wahrscheinlichkeit p(x),

desto stärker sinkt die Unbestimmtheit / Wahlfreiheit,

desto höher ist der Informationsgehalt

$$I(p(x)) := \log\left(\frac{1}{p(x)}\right) \qquad \text{P.S.: } \log = \log_2$$

Einheit: Bit

Beispiel

$$I(p) := \log\left(\frac{1}{p(x)}\right)$$

heute gabs kartoffeln und l lachs, lamm, leber, labskaus, letscho

heute gabs kartoffeln und q quark

I(1) = 0

I(0,5) = 1

I(0,25) = 2

I(0,125) = 3

Formalisierung II

Entropie (Unordnung) ist

der erwartete Informationsgehalt des nächsten Zeichens

$$H(X) = \sum_{x \in A} p(x) \cdot I(p(x))$$

Einheit: Bit pro Zeichen

Beispiel

Versuch 1

a 00000

b 00001

c 00010

d 00011

e 00100

f 00101

g 00110

h 00111

…

optimal bei p(a)=p(b)=…=1/32

spiegelt nicht die

Häufigkeitsverteilung in

deutschen Texten wider

schlecht

Beispiel

Versuch 2

e 0

n 10

i 110

s 1110

r 11110

a 111110

t 1111110

d 11111110

…

optimal bei p(e)=1/2, p(n)=1/4, p(i)=1/8, …

d.h. bei Wortlänge 6 und Satzlänge 11

nur alle vier Sätze ein d

und jeder zweite Buchstabe ein e

noch schlechter

Zwischending

Ziel: Abschätzung nach unten für die Länge der kodierten Nachricht. (als Referenz für mögliche Zwischendinger)

Annahme: kennen Länge der Nachricht k und genaue Häufigkeit der einzelnen Buchstaben

Ergebnis: müssen mind. k*H(X) Bit übertragen

(Beweis siehe nächste Folie)

Beweis

n: Anzahl verschiedener Buchstaben, k: Länge der Nachricht, k_i: abs. Häufigkeit von Buchstabe i

müssen mindestens $h(x) = \log\left(\frac{k!}{k_1!\cdot k_2!\cdots k_n!}\right)$ Bit übertragen

$$
\begin{aligned}
h(x) &= \log k! - \log k_1! - \cdots - \log k_n! \\
&\approx \log\left(\sqrt{2\pi k}\left(\frac{k}{e}\right)^k\right) - \log\left(\sqrt{2\pi k_1}\left(\frac{k_1}{e}\right)^{k_1}\right) - \cdots - \log\left(\sqrt{2\pi k_n}\left(\frac{k_n}{e}\right)^{k_n}\right) \\
&\approx \log\left(\sqrt{2\pi k}\right) - \log\left(\sqrt{2\pi k_1}\right) - \cdots - \log\left(\sqrt{2\pi k_n}\right) \\
&\quad + k\log k - k_1\log k_1 - \cdots - k_n\log k_n \\
&\quad - k\log e + k_1\log e + \cdots + k_n\log e \\
&\approx \log\left(\sqrt{2\pi k}\right) - \log\left(\sqrt{2\pi k_1}\right) - \cdots - \log\left(\sqrt{2\pi k_n}\right) \\
&\quad + k\log k - k_1\log k_1 - \cdots - k_n\log k_n \\
&\geq k\log k - k_1\log k_1 - \cdots - k_n\log k_n \\
&\geq \log k^k - \log k_1^{k_1} - \cdots - \log k_n^{k_n} \\
&\geq \log k^{k_1+\cdots+k_n} - \log k_1^{k_1} - \cdots - \log k_n^{k_n} \\
&\geq \log\left(k^{k_1}\cdot k^{k_2}\cdots k^{k_n}\right) - \log k_1^{k_1} - \cdots - \log k_n^{k_n} \\
&\geq \log k^{k_1} + \cdots + \log k^{k_n} - \log k_1^{k_1} - \cdots - \log k_n^{k_n} \\
&\geq \sum\left(\log k^{k_i} - \log k_i^{k_i}\right) \\
&\geq \sum k_i\left(\log k - \log k_i\right) \\
&\geq k\sum\frac{k_i}{k}\left(\log k - \log k_i\right) \\
&\geq k\sum\frac{k_i}{k}\log\frac{k}{k_i} \\
&\geq k\sum p_i\log\frac{1}{p_i} \\
&\geq k\cdot H(X)
\end{aligned}
$$

stirlingsche Approximation:

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

aus Konkavität von $\sqrt{\ }$ und log folgt:

$$\log\sqrt{a+b} \leq \log\sqrt{a} + \log\sqrt{b}$$

Shannon-Fano-Code

- sortiere Buchstaben absteigend nach Wahrscheinlichkeit

  $p_e=0,3$; $p_n=0,25$; $p_i=0,2$; $p_s=0,15$; $p_r=0,1$

- berechne kumulative Wahrscheinlichkeiten in binär

  $P_e=0$; $P_n=0,3$; $P_i=0,55$; $P_s=0,75$; $P_r=0,9$ ← dezimal

  $P_e=0$; $P_n=0,0100\ldots$; $P_i=0,1000\ldots$; $P_s=0,11$; $P_r=0,1110\ldots$ ← binär

- berechne aufgerundete Informationsgehalte

  $I'(p_e)=2$; $I'(p_n)=2$; $I'(p_i)=3$; $I'(p_s)=3$; $I'(p_r)=4$ $\quad I(p) := \log\left(\frac{1}{p(x)}\right)$

- schneide kumul. Wahrschlk. entspr. Informationsgehalt ab

  E(e)=00; E(n)=01; E(i)=100; E(s)=110; E(r)=1110

Shannon-Fano-Code

$$\begin{aligned}\sum_x p_x \cdot |E(x)| &= \sum_x p_x \cdot l'(p_x) \\ &< \sum_x p_x \cdot (l(p_x)+1) \\ &< 1+\sum_x p_x \cdot l(p_x) \\ &< 1+H(X)\end{aligned}$$

erwartete Länge des nächsten Codewortes weicht um max. 1 von Entropie der Nachrichtenquelle ab

asymptotisch gut

per Konstruktion präfixfrei

Quellenkodierungssatz

noiseless coding theorem

$\forall$ Wahrscheinlichkeitsmassenfunktion p auf Quellwörtern

$\exists$ präfixfreier Code, so dass

durchschnittliche Codewortlänge $\sum_x p_x \cdot |E(x)|$ ungefähr

gleich der Entropie der Nachrichtenquelle $H(X) = \sum_x p_x \cdot \log\left(\frac{1}{p_x}\right)$

Beweis:

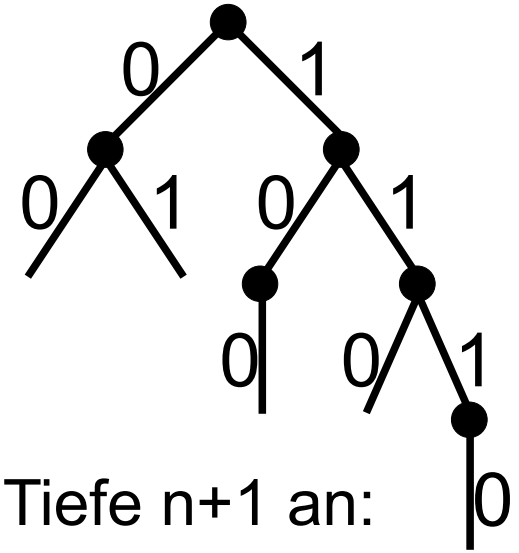
- können Shannon-Fano-Code konstruieren mit $\sum_x p_x \cdot |E(x)| < 1 + H(X)$ (Folie 13)
- können keinen Code konstruieren mit $\sum_x p_x \cdot |E(x)| < H(X)$ (Folie 11)

kraftsche Ungleichung

∃ präfixfreier Code mit Codewortlängen l_1, l_2, …

gdw.

$$\sum_i 2^{-l_i} \leq 1$$



Beweis ⇒

- PFC als Baum mit Codewörtern in Blättern
- Baum mit Tiefe 0: $2^{-0} = 1$
- hänge an Blatt mit Tiefe n max. 2 Blätter mit Tiefe n+1 an:
  - Blatt wird zu innerem Knoten: ziehe 2^{-n} ab
  - max. 2 Blätter dazu: addiere max. $2*2^{-(n+1)} = 2^{-n}$

Beweis ⇐

- gibt maximal 2^n Knoten mit Tiefe n
- ein Blatt mit Tiefe l_i verbraucht deshalb 2^{-l_i} der Breite des Baumes
- lt. Vorausstzg. verbrauchen alle Blätter max. 1 der Breite des Baumes
- also ist ein Baum mit Blättern der Tiefe l_i konstruierbar

kraftsche Ungleichung

daraus folgt:

- wenn $\sum_i 2^{-l_i} < 1$, werden Codewörter verschwendet
  (der Baum ist nicht voll, *unvollständiger Code*)

aber daraus folgt nicht:

- wenn für gegebenen Code $\sum_i 2^{-l_i} \leq 1$ gilt, ist es ein PFC
  (stimmt ja auch nicht)

kraftsche Ungleichung gilt nicht nur für PFC,
sondern für alle eindeutig dekodierbaren Codes
(Beweis nächste Folie)

daraus folgt:

- jeder EDC kann durch einen PFC mit
  gleichen Codewortlängen ersetzt werden

Beweis

Voraussetzung: $l_1 \ldots l_n$ sind die Längen der n Codewörter eines EDC

zu zeigen: es gilt die kraftsche Ungleichung $\sum_{i=1}^{n} 2^{-l_i} \leq 1$

setzen die linke Seite ins Quadrat und erhalten:

$$\left(\sum_{i=1}^{n} 2^{-l_i}\right)^2 = (2^{-l_1} + \cdots + 2^{-l_n})^2$$

$$= 2^{-l_1} \cdot 2^{-l_1} + 2^{-l_1} \cdot 2^{-l_2} + \cdots + 2^{-l_n} \cdot 2^{-l_n}$$

$$= 2^{-l_1 - l_1} + 2^{-l_1 - l_2} + \cdots + 2^{-l_n - l_n}$$

$$= m_1 \cdot 2^{-1} + m_2 \cdot 2^{-2} + \cdots m_{2l} \cdot 2^{-2l} \text{ mit } l = \max\{l_i | i = 1 \ldots n\} \text{ und } m_i = \text{Anzahl Paare } (l_a, l_b) \text{ mit } l_a + l_b = i$$

$$= \sum_{i=1}^{2l} m_i \cdot 2^{-i}$$

analoge Überlegungen führen zu:

$$\left(\sum_{i=1}^{n} 2^{-l_i}\right)^r = \sum_{i=1}^{r \cdot l} m_i \cdot 2^{-i}$$

$$\leq \sum_{i=1}^{r \cdot l} 2^i \cdot 2^{-i} \text{ (nach Voraussetzung: } m_i \leq |\{0,1\}^i| = 2^i \text{ (für jede Bitkette höchstens ein Codewortpaar))}$$

$$\leq \sum_{i=1}^{r \cdot l} 1$$

$$\leq r \cdot l$$

setzen $r \to \infty$

$$\sum_{i=1}^{n} 2^{-l_i} \leq \sqrt[r]{r \cdot l}$$

$$\leq 1$$

optimale Codes

Quellenkodierungssatz und kraftsche Ungleichung liefern Konstruktion eines relativ guten PFC

„gut“ = Erwartungswert der Codewortlänge ist klein

Further Reading: Huffman-Code liefert Konstruktion eines besten (optimalen) PFC

A ist optimaler Code (PFC, EDC) $\Leftrightarrow \nexists$ Code (PFC, EDC) B, sodass $\sum_x p_x |B(x)| < \sum_x p_x |A(x)|$

Achtung: vollständig ≠ optimal

universelle Codes

kodieren die Nachrichtenquelle

in Unkenntnis der Wahrscheinlichkeitsverteilung p,

sodass die erwartete Codewortlänge

linear von der Entropie abhängt

$$\sum_x p_x \cdot |E(x)| \leq c \cdot \max\{H(X), 1\}$$

sind zusätzlich *asymptotisch optimal,* wenn

$$\sum_x p_x \cdot |E(x)| \leq f(H(X)) \cdot \max\{H(X), 1\} \text{ mit } \lim_{H(X)\to\infty} f(H(X)) = 1$$

Beispiel

$A : n \to \bar{n}$

ist universell

(da die Codewortlänge nur linear mit der Entropie zunimmt

$B : n \to \overline{l(n)}\, n$

ist universell

und asymptotisch optimal

(da die Codewortlänge nur logarithmisch mit der Entropie zunimmt)

Fragen?

- Ming Li, Paul Vitány.

  An Introduction to Kolmogorov Complexity and Its Applications.

  Springer. 2008

- Claude Shannon.

  A Mathematical Theory of Communication. 1948

  http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

- Alexander May.

  Diskrete Mathematik 2.

  Vorlesungsskript Ruhr-Universität Bochum. 2009

  http://www.cits.rub.de/lehre/dismath2ss09.html

Glossar

complete code	vollständiger Code
conditional	bedingt
countable	abzählbar
dependent	abhängig
distribution	Verteilung
domain	Definitionsbereich
efficient	effizient
event	Ereignis
fixed-length code	Blockcode
independent	unabhängig
joint	gemeinsam
marginal	rand-
mutual	gemeinsam
noiseless	rauschfrei
prefix-code	präfixfreier Code
random variable	Zufallsvariable
range	Wertebereich
sample	Ergebnis
uncertain	unbestimmt
uncountable	überabzählbar
uniform	gleich-
unique	eindeutig

abhängig	dependent
abzählbar	countable
bedingt	conditional
Blockcode	fixed-length code
Definitionsbereich	domain
effizient	efficient
eindeutig	unique
Ereignis	event
Ergebnis	sample
gemeinsam	joint, mutual
gleich-	uniform
präfixfreier Code	prefix-code
rand-	marginal
rauschfrei	noiseless
überabzählbar	uncountable
unabhängig	independent
unbestimmt	uncertain
Verteilung	distribution
vollständiger Code	complete code
Wertebereich	range
Zufallsvariable	random variable