Entropie

Mathematikreferat von Ingo Blechschmidt und Michael Hartmann am 15. November 2006

Definition

- Informationsgehalt eines Zeichens:
  $I(x) := -\log_2 P(x);$ [1 bit]
- Entropie als Erwartungswert des Informationsgehalts:
  $E(I) = \sum_{x \in \Sigma} P(x) \cdot I(x);$

wobei

x: ein bestimmtes Zeichen
Σ: Menge aller vorkommenden Zeichen
$P(x)$: Wahrscheinlichkeit des Auftretens von x

Eigenschaften

- Maximale Entropie bei $|\Sigma| = 2^n$, $n \in \mathbb{N}$:
  $E(I) = -\log_2 \frac{1}{2^n} = n\,\text{bit};$
- Maß für Überraschung; je seltener ein Zeichen, desto höher der Informationsgehalt
- Maß für die untere Schranke verlustfreier Kompression
- Maß für die Zufälligkeit von Information
- Maß für „Chaos“
- Charakteristika von Autoren und Komponisten

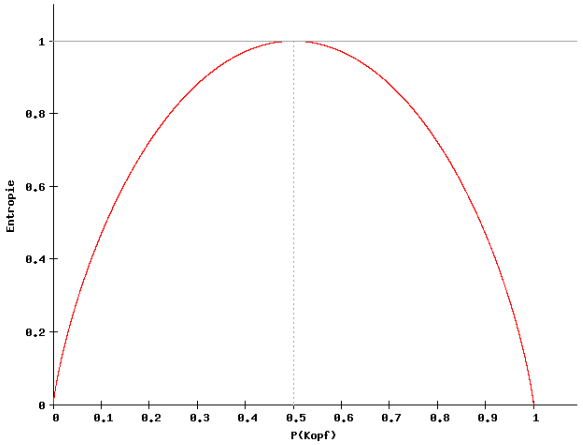
Gezinkter Münzwurf

$\Sigma = \{\text{Kopf}, \text{Zahl}\};$

$P(\text{Kopf}) = p = 1 - P(\text{Zahl});$

$$\begin{aligned} I(\text{Kopf}) &= -\log_2 p; \\ I(\text{Zahl}) &= -\log_2 (1-p); \end{aligned}$$

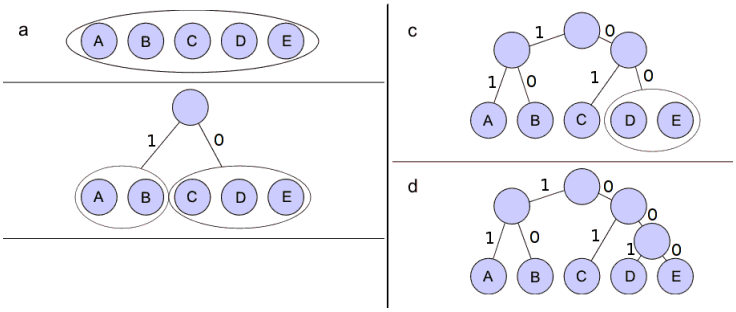
$$\begin{aligned} E(I) &= p\, I(\text{Kopf}) + (1-p)\, I(\text{Zahl}) = \\ &= p \cdot [-\log_2 p] + (1-p) \cdot [-\log_2 (1-p)]; \end{aligned}$$



Shannon–Fano-Kodierung

- Entropiekodierung („Kompressionsverfahren“)
- Darstellung *häufiger* Zeichen durch *kurze* Bitfolgen, *seltener* Zeichen durch *lange* Bitfolgen
- Eindeutigkeit der Bitfolgen („Präfixfreiheit“) notwendig

1. Sortierung der Zeichen nach rel. Häufigkeit
2. Einteilung der Zeichen in zwei Gruppen, sodass Summen der Häufigkeiten etwa gleich
3. So lange fortfahren, bis Entsprechung jedes Zeichens durch einen Pfad im Baum



Probleme

- Keine Beachtung der Reihenfolge
- Analyse nur auf syntaktischer Ebene