Datenkompression

Sommersemester 2018 — **Übungsblatt 2** — Prof. Ohlebusch

Institut für Theoretische Informatik — Uwe Baier

Ausgegeben am 17.05.2018 — Besprechung am 24.05.2018

Aufgabe 2.1

Bestimmen Sie für $\Sigma = \{a, b, c, d\}$ mit $P(a) = 0,7$ und $P(b) = P(c) = P(d) = 0,1$ einen Huffmancode. Wie viele Huffmancodes gibt es insgesamt für P?

Wie kann sichergestellt werden, dass der Decodierer den gleichen Huffmancode verwendet, wie der Codierer und welche Information(en) müssen hierfür zusätzlich übertragen werden?

Aufgabe 2.2

Die Zeichen des Alphabetes $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ mit der Anordnung $\mathtt{a} < \mathtt{b} < \mathtt{c}$ treten mit den Wahrscheinlichkeiten $P(\mathtt{a}) = \frac{1}{4}$, $P(\mathtt{b}) = \frac{1}{2}$ und $P(\mathtt{c}) = \frac{1}{4}$ auf.

(a) Berechnen Sie die arithmetische Kodierung des Wortes `baac`.

(b) Dekodieren Sie die arithmetische Kodierung `0100010` eines Wortes der Länge 4.

Aufgabe 2.3

Auf einem Alphabet Σ sei die Wahrscheinlichkeitsfunktion $P\colon \Sigma \to [0, 1]$ definiert. Für Zeichenketten aus Σ^m, deren Zeichen unabhängig voneinander gezogen werden, erhalten wir die Wahrscheinlichkeitsfunktion $P_m\colon \Sigma^m \to [0, 1]$, $(c_1, \dots, c_m) \mapsto P(c_1) \cdots P(c_m)$.

Beweisen Sie die Additivität der Entropie: $H(\Sigma^m, P_m) = mH(\Sigma, P)$.

Aufgabe 2.4

Gegeben sei ein Text S der Länge n über einem Alphabet Σ der Größe $|\Sigma| = \sigma$. Weiter sei die Häufigkeit des Buchstaben mit Rang i gegeben durch $\mathtt{c}_i$, es gilt $\sum_{i=1}^{\sigma} \mathtt{c}_i = n$. In der Vorlesung haben Sie den Begriff der Shannon-Entropie kennen gelernt:

$$H(S) \quad = \quad \sum_{i=1}^{\sigma} \frac{\mathtt{c}_i}{n} \log_2\left(\frac{n}{\mathtt{c}_i}\right) \quad \overset{P(i):=\frac{\mathtt{c}_i}{n}}{=} \quad \sum_{i=1}^{\sigma} P(i) \log_2\left(\frac{1}{P(i)}\right)$$

Wie Sie bereits aus der Vorlesung wissen, bildet die Shannon-Entropie eine untere Schranke für die mittlere Codewortlänge eines eindeutig dekodierbaren Codes. In dieser Aufgabe sollen Sie herleiten, warum die Shannon-Entropie eine solche untere Schranke bildet.

(a) Gegeben sei die Häufigkeitsverteilung $\mathtt{c}_1, \dots, \mathtt{c}_\sigma$ wie am Anfang der Aufgabe beschrieben. Wie viele verschiedene Texte können mithilfe dieser Buchstabenhäufigkeiten erzeugt werden? Geben Sie eine allgemeine Formel in Abhängigkeit von n und $\mathtt{c}_1, \dots, \mathtt{c}_\sigma$ an. Zeigen Sie, dass diese zu $\frac{n!}{\mathtt{c}_1! \cdot \mathtt{c}_2! \cdot \ldots \cdot \mathtt{c}_\sigma!}$ äquivalent ist.

(b) Leiten Sie mithilfe der Teilaufgabe (a), der Stirlingschen Abschätzung $\log_2(n!) \approx n \log_2(n)$ und folgendem Theorem die Shannon-Entropie her:

> **Theorem**
> Sei A eine Menge von Texten. Treten die Elemente in A mit gleicher Wahrscheinlichkeit auf, so benötigt ein beliebiger, eindeutig dekodierbarer Code für A (jedem Element wird eine Binärsequenz zugeordnet) im Mittel mindestens $\log_2(|A|)$ Bits um ein einzelnes Element aus A zu kodieren.