GRUNDLAGEN DER INFORMATIONS-THEORIE

Information

Definition

Information ist eine räumliche oder zeitliche Abfolge physikalischer Signale, die mit bestimmten Häufigkeiten oder Wahrscheinlichkeiten auftritt. Sie lässt sich als Folge von Bits, meistens in 0 und 1 angegeben, darstellen und kann unter drei hermeneutischen Ebenen betrachtet werden:

- Syntax: die formale Struktur und Darstellung
- Semantik: die Bedeutung der Information
- Pragmatik: die Wirkung oder der Nutzen im Kontext

Informationsgehalt

Der Informationsgehalt I eines Ereignisses mit Eintrittswahrscheinlichkeit p ist definiert als:

$$I = -\log_2 p \quad [\text{Bit}]$$

Diese logarithmische Maßzahl erfüllt drei wesentliche Eigenschaften:

1. Additivität:

   Bei unabhängigen Ereignissen addieren sich die Informationsgehalte

2. Stetigkeit:

   Kleine Änderungen in p führen zu kleinen Änderungen in I

3. Normierung:

Ein Bit entspricht der Information eines gleichwahrscheinlichen Ja/Nein-Entscheids

Entropie

Definition

Die Entropie H einer diskreten Quelle Q wird definiert als der Erwartungswert des Informationsgehalts ihrer Zeichen:

$$H(Q) = \sum_{s \in Q} p(s) \cdot \log_2 \left(\frac{1}{p(s)} \right)$$

Wobei: $p(s)$ die Wahrscheinlichkeit des Auftretens des Symbols s. Die Basis 2 des Logarithmus gibt die Einheit "Bit" pro Zeichen. Für $p(s) = 0$ wird der Beitrag als 0 definiert.

Bedeutung der Entropie

Die Entropie quantifiziert die mittlere Unsicherheit über die nächste Nachricht einer Quelle. Sie ist maximal, wenn alle Zeichen gleichwahrscheinlich sind, und minimal (null), wenn ein Zeichen sicher eintritt.

Kompressionsprinzip

Die Entropie stellt eine fundamentale untere Schranke für die verlustlose Kompression dar: Kein Code kann im Mittel weniger als H Bit pro Zeichen benötigen. Effiziente Kompressionsverfahren weisen häufigen Zeichen kurze Codewörter und seltenen Zeichen längere Codewörter zu.

Shannon-Fano Kodierung

Die Shannon-Fano-Kodierung ist ein verlustloser Kompressionsalgorithmus. Das Ziel der Shannon-Fano-Kodierung ist die Zuordnung von optimalen

Präfixcodes zu Symbolen basierend auf ihren Auftrittswahrscheinlichkeiten. Häufig auftretende Symbole erhalten kurze Codes, seltene Symbole längere Codes.

Huffman-Kodierung

Die Huffman-Kodierung ist ein verlustloser Kompressionsalgorithmus. Das Ziel der Huffman-Kodierung ist die Zuordnung von optimalen Präfixcodes zu Symbolen basierend auf ihren Auftrittswahrscheinlichkeiten. Häufige Symbole erhalten kurze Bitfolgen, seltene Symbole längere Bitfolgen, wodurch die durchschnittliche Codelänge minimiert wird.

Optimalität des Huffman-Verfahren

Theorem

Der Huffman-Algorithmus erzeugt einen optimalen Präfixcode, das heißt einen Code mit minimaler durchschnittlicher Codewortlänge.

Beweis

Für zwei Symbole erzeugt der Huffman-Algorithmus zwei Codewörter der Länge 1 (z.B. '0' und '1'). Dies ist offensichtlich optimal, da kürzere Codewörter nicht möglich sind.

Annahme: Der Huffman-Algorithmus erzeugt für jedes Alphabet mit k Symbolen ($k \leq n$) einen optimalen Präfixcode.

Seien a_x und a_y die beiden Symbole mit den kleinsten Wahrscheinlichkeiten. Der Huffman-Algorithmus kombiniert diese zu einem neuen Symbol z mit $p_z = p_x + p_y$. Wir erhalten ein reduziertes Alphabet mit n Symbolen.

Nach Induktionsvoraussetzung erzeugt der Huffman-Algorithmus für das reduzierte Alphabet mit n Symbolen einen optimalen Präfixcode C'.

Aus C' konstruieren wir C, indem wir das Codewort für z durch zwei Codewörter ersetzen:

$$c_x = c_z \parallel 0 \quad \text{und} \quad c_y = c_z \parallel 1$$

Die durchschnittliche Länge ändert sich um:

$$L(C) = L(C') + p_x + p_y$$

Lemma 1

In jedem optimalen Präfixcode für das ursprüngliche Alphabet haben die beiden Symbole mit kleinsten Wahrscheinlichkeiten Codewörter maximaler Länge, und diese können so modifiziert werden, dass sie Geschwisterknoten im Codebaum sind.

Lemma 2

Wenn wir in C^* die beiden Symbole mit kleinsten Wahrscheinlichkeiten zu einem Symbol zusammenfassen, erhalten wir einen Code $C^{*'}$ für das reduzierte Alphabet mit:

$$L(C^{*'}) = L(C^*) - p_x - p_y$$

Da C' optimal für das reduzierte Alphabet ist:

$$L(C') \leq L(C^{*'}) = L(C^*) - p_x - p_y$$

Daraus folgt:

$$L(C) = L(C') + p_x + p_y \leq L(C^*)$$

Also kann C^* nicht besser sein als C.

Da der Huffman-Algorithmus einen Code erzeugt, der mindestens so gut ist wie

jeder andere Präfixcode, ist er optimal.

□

Struktureigenschaft optimaler Codes

Lemma

Es gibt einen optimalen Präfixcode, in dem die beiden Symbole mit den geringsten Wahrscheinlichkeiten s_{n-1} und s_n denselben Vaterknoten haben und die längsten Codewörter besitzen.

Beweis

Sei T ein optimaler Codebaum für Q. In T gibt es mindestens zwei Blätter auf der tiefsten Ebene. Seien s_i und s_j zwei Blätter auf der tiefsten Ebene, die denselben Vaterknoten haben. Wir vertauschen s_n mit s_j und erhalten einen neuen Baum T'.

$$\begin{aligned} Z(T') - Z(T) &= p(s_n)k(s_j) + p(s_j)k(s_n) - p(s_j)k(s_j) - p(s_n)k(s_n) \\ &= (p(s_n) - p(s_j))(k(s_j) - k(s_n)) \end{aligned}$$

Da $p(s_n) \leq p(s_j)$ (weil s_n die kleinste Wahrscheinlichkeit hat), gilt $p(s_n) - p(s_j) \leq 0$. Da s_j auf der tiefsten Ebene liegt und s_n irgendein Blatt ist, gilt $k(s_j) \geq k(s_n)$, also $k(s_j) - k(s_n) \geq 0$
Somit:

$$(p(s_n) - p(s_j))(k(s_j) - k(s_n)) \leq 0$$

$Z(T') \leq Z(T)$, also ist T' mindestens so gut wie T. Durch iterative Anwendung können wir s_{n-1} und s_n an die tiefste Stelle mit gemeinsamem Vater bringen.

□

Optimale Codeverfahren

- Shannon-Fano-Codierung:

  Top-down-Verfahren mit schrittweiser Teilung der Wahrscheinlichkeiten

- Huffman-Codierung:

  Bottom-up-Verfahren durch wiederholtes Zusammenfassen der unwahrscheinlichsten Symbole

Der Huffman-Code erzeugt immer einen optimalen präfixfreien Code und erreicht eine mittlere Codewortlänge L, die die Entropiegrenze erfüllt: $H \leq L < H + 1$.

Quellencodierungssatz, Shannons erster Hauptsatz

Der Quellencodierungssatz besagt, dass eine Quelle mit Entropie H verlustlos mit durchschnittlich H Bits pro Zeichen kodiert werden kann, wenn Blockkodierung verwendet wird.

Effizienz und Quellenkodierungssatz

Die Effizienz eines Codes T ist definiert als $\mathrm{EFF}(T) = H(Q)/Z(T)$. Ein Code heißt ideal bei $\mathrm{EFF}(T) = 1$ und kompakt oder optimal, wenn keine andere Präfixcodierung eine kürzere mittlere Länge $Z(T)$ erreicht. Der fundamentale Quellenkodierungssatz von Shannon besagt, dass durch Blockcodierung (Kodierung mehrerer Zeichen zusammen) die Effizienz beliebig nahe an 1 gebracht werden kann: $\lim_{r\to\infty} \mathrm{EFF}(T^r) = 1$.