Anhang B

Entropie

Der Begriff Entropie wurde von Shannon (1948) auf die Informationstheorie übertragen. Ursprünglich stammt er aus der Thermodynamik, wo Entropie – etwas vereinfacht gesagt – das Maß an Ordnung oder Unordnung (je nachdem, was man als Ordnung definiert) bezeichnet. In der Informationstheorie steht die Entropie in unmittelbarem Zusammenhang mit dem Informationsgehalt. Der Informationsgehalt b einer beliebigen Einheit eines Zeichensystems I, ausgedrückt in Bits, ist der negative binäre Logarithmus der Auftrittswahrscheinlichkeit $p(I)$ dieses Zeichens:

$$b(I) = -log_2(p(I)) \tag{B.1}$$

Die Entropie h_1 ist nun ein Maß für den mittleren Informationsgehalt H eines Zeichensystems mit n unterschiedlichen Ereignissen I. Sie wird über die Summe der Produkte aller Wahrscheinlichkeiten von I mit dem Informationsgehalt aller I ermittelt:

$$h_1 = H(I) = \sum_{I=1}^{n} p(I) * b(I) \tag{B.2}$$

Nach Einsetzen der Formel B.1 in B.2 ergibt sich dann

$$h_1 = H(I) = -\sum_{I=1}^{n} p(I) * log_2(p(I)) \tag{B.3}$$

B.1 Interpretation der Entropie

Setzen wir als Zeichensystem einen Text, so liegt der Informationsgehalt eines Zeichens innerhalb dieses Textes umso höher, je seltener das Zeichen im Text vorkommt. Wenn der Text ausschließlich aus einem Zeichen besteht, ist der Informationsgehalt dieses Zeichens (und des gesamten Textes) 0. Die Entropie als mittlerer Informationsgehalt eines Textes wäre in diesem Fall 0. Was aber ist der höchste mittlere Informationsgehalt, also die maximale Entropie H_{max} eines Textes? H_{max}, auch als h_0 bezeichnet, wird erreicht, wenn alle Zeichen gleich oft vorkommen. Der Wert

hängt von der Anzahl der unterschiedlichen Zeichen ab, da

$$h_0 = H_{max} = -\sum_{I=1}^{n} \frac{1}{N} * log_2 \frac{1}{N} = log_2 N \tag{B.4}$$

Für Texte mit 30 verschiedenen Einheiten (wie dem deutschen Alphabet inklusive Umlauten und ß) liegt die maximale Entropie bei 4,91, in Zeichensystemen mit nur 10 Einheiten (z.B. Ziffernfolgen) liegt sie bei 3,32. Die maximale Entropie kann dazu genutzt werden, um unterschiedliche Texte mit einem unterschiedlich umfangreichen Zeicheninventar miteinander zu vergleichen, indem man die ermittelten Entropien der Texte durch die maximalen Entropiewerte für diese Texte teilt. Der damit ermittelte Entropiewert bleibt immer kleiner als 1.

Die Entropie eines Zeichensystems liegt demnach immer zwischen 0 (wenn der Text nur aus einem Zeichen besteht) und H_{max}(Wenn alle Zeichen gleich oft vorkommen). Die Entropiewerte sind also, wenn man nur einzelne Zeichen betrachtet, lediglich ein Maß für die Gleichverteilung von Zeichen eines Textes. Wenn nun aber nicht nur Einzelzeichen, sondern auch Zeichenkombinationen als Zeichen zugelassen werden, kann die Entropie als ein Maß für die Repetitivität von Texten angesehen werden, d.h. als Maß für die Wiederkehr von Mustern in Texten. Zur Ermittlung der Entropiewerte für Zeichenkombinationen muss die Verbundentropie, auch Blockentropie genannt, errechnet werden.

B.2 Verbund- oder Blockentropie

Die Blockentropie H_k, wobei k die Anzahl der Zeichen von Blöcken angibt, basiert auf der Verbundwahrscheinlichkeit von Zeichen $p(I, J)$. Diese ergibt sich aus der Beziehung:

$$p(I, J) = \frac{p(J|I)}{(p(I))} \tag{B.5}$$

Dabei ist $p(J|I)$ die sogenannte *Bedingte Wahrscheinlichkeit*, also die Wahrscheinlichkeit für das Auftreten von I unter der Bedingung, dass J gegeben ist. $p(\text{“}a\text{“}|\text{“}\_\text{“})$, verbalisiert “Wahrscheinlichkeit für das Zeichen a unter der Bedingung, dass ein Leerzeichen vorangeht” wäre mithin die Wahrscheinlichkeit, dass ein Wort mit “a” beginnt.

Der Informationsgehalt eines Blocks der Länge 2 mit J als erstem und J als zweitem Element ist nun folgendermaßen definiert:

$$b(I, J) = -log(p(I, J)) \tag{B.6}$$

Daraus ergibt sich für die Verbundentropie von Zweierblöcken h_2, in diesem Fall auch die bedingte

Entropie oder Entropie zweiter Ordnung:

$$h_2 = \sum_{I=1}^{n} p(I,J) * b(I,J) = -\sum_{I=1}^{n}\sum_{J=1}^{n} p(x_I) * p(y_J|x_I) * log_2(p(y_J|x_I)) \tag{B.7}$$

Entropien höherer Ordnung ermittelt man entsprechend, wobei unschwer zu erkennen ist, dass schon die Berechnung von h_3 sehr kostenintensiv ist:

$$h_3 = \sum_{I} p(I) \sum_{J} p(I,J) \sum_{K} P(I,J,K) - log_2(P(I,J,K)) \tag{B.8}$$