Proseminar Datenkompression
WS 96/97

Grundlagen der Informationstheorie

Maximilian Höfner

February 13, 1997

Betreuer: Thomas Erlebach

1 Historische Entwicklung der Informationstheorie

Um 1930 begann Hartley sich mit Kommunikation und Information mathematisch auseinander zu setzen. Gegen 1945 brachte Shannon den Begriff der Entropie und der Gegenseitigen Information auf. Die Relative Entropie wurde etwas später von Kullback und Leibler eingeführt, weswegen man manchmal auch vom Kullback-Leibler-Abstand spricht.

2 Grundbegriffe der Informationstheorie

2.1 Informationsgehalt einer Nachricht

In der Informationstheorie interessiert der semantische Wert einer Nachricht nicht. Das heißt, daß es für den Informationstheoretiker gleichgültig ist, ob die Nachrichtenquelle die Bibel oder ein Comic ist. Der Wert einer Nachricht [n] richtet sich nach der Wahrscheinlichkeit [p(n)], mit der genau diese Nachricht von der Quelle ausgesandt wird. Je geringer die Wahrscheinlichkeit ist, daß diese Nachricht gesandt wird, desto größer ist der Informationsgewinn. Man könnte sagen, je größer die Überraschung über eine Nachricht beim Empfänger ist, desto größer ist der Informationsgewinn.

Eine Funktion, die den Informationsgehalt einer Nachricht bestimmen soll, muß folgende drei Eigenschaften aufweisen:

1. Werden zwei Nachrichten unabhängig voneinander ausgesandt, so soll sich der Informationsgehalt der Nachrichten addieren.

2. Unterscheiden sich die Auswahlwahrscheinlichkeiten zweier Nachrichten nur geringfügig, so soll ihr Informationsgehalt auch annähernd gleich sein.

3. Eine Nachricht, deren Auswahlwahrscheinlichkeit genau $\frac{1}{b}$ ist, hat den Informationsgehalt 1, wobei b eine reelle Zahl größer 1 ist. Ist b gleich 2, so heißt die Maßeinheit bit.

Die resultierende Funktion zur Bestimmung des Informationsgehaltes ist demnach:

$$I(n) = -\log_b p(n)$$

2.2 Entropie

Die Entropie [H(X)] einer Nachrichtenquelle [X] ist das Maß für den mittleren Informationsgewinn, der von der Quelle ausgeht. Je höher die Entropie einer Nachrichtenquelle ist, desto unsicherer ist der Empfänger über die nächste Nachricht. Die Entropie ist aber vor allem die untere Grenze für die Anzahl der benötigten Informationseinheiten (zum Beispiel bits) um eine Nachricht genau zu identifizieren. Sie ist das gewichtete Mittel der Informationsgehalte aller Nachrichten einer Quelle und wird wie folgt berechnet:

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_b p(x)$$

2.2.1 Nichtnegativität der Entropie

Die Entropie ist nie negativ, da der Logarithmus einer Zahl zwischen Null und Eins nicht positiv ist und das Minus vor der Summe das Vorzeichen umkehrt.

2.2.2 Minimale Entropie

Die Entropie einer Nachrichtenquelle ist dann minimal, wenn sie Null ist. Das ist der Fall, wenn keine Nachrichten gesandt werden. Die zweite Möglichkeit ist, daß eine Nachricht mit der Wahrscheinlichkeit Eins, also immer ankommt. In beiden Fällen ist der Empfänger nicht überrascht und es besteht keine Unsicherheit über die nächste Nachricht.

2.2.3 Maximale Entropie

Die Frage, wann die Entropie maximal wird, ist nicht ganz so trivial. Deshalb betrachten wir zwei Nachrichtenquellen, deren Stichprobenraum gleichgroß und deren Nachrichten eine Auswahlwahrscheinlichkeit größer Null besitzen:

1. $\sum_{i=1}^{n} x_i = 1;$
2. $\sum_{i=1}^{n} y_i = 1;$
3. $x_i \cdot y_i > 0$ für alle $1 \leq i \leq n$

Unter diesen Vorraussetzungen gilt:

$\sum\limits_{i=1}^{n}(x_i \cdot \log \frac{y_i}{x_i}) =$

$\overset{\{1\}}{=} \frac{1}{\ln b} \cdot \sum\limits_{i=1}^{n}(x_i \cdot \ln \frac{y_i}{x_i}) \leq$

$\overset{\{2\}}{\leq} \frac{1}{\ln b} \cdot \sum\limits_{i=1}^{n}(x_i \cdot (\frac{y_i}{x_i} - 1)) =$

$= \frac{1}{\ln b} \cdot \sum\limits_{i=1}^{n}(y_i - x_i) =$

$\overset{\{3\}}{=} \frac{1}{\ln b} \cdot (\sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i) = 0$

{1}Basiswechsel des Logarithmus
{2}für jedes positive, reelle b ist $\ln b \leq b - 1$, Gleichheit gilt bei $b = 1$
{3}da sich die x_i und y_i jeweils zu1aufaddieren
Das heißt:

$$\sum_{i=1}^{n}(x_i \cdot \log \frac{y_i}{x_i}) \leq 0, \text{ Gleichheit gilt, wenn jedes } x_i = y_i$$

Behauptung:
$H(X) \leq \log_b n$, wobei n die Mächtigkeit des Stichprobenraumes der Zufallsvariablen ist.

$$\begin{aligned} H(X) - \log_b n &\overset{\{1\}}{=} -\sum_{i=1}^{n} p(i) \cdot \log_b p(i) - \sum_{i=1}^{n} p(i) \cdot \log_b n = \\ &\overset{\{2\}}{=} \sum_{i=1}^{n} p(i) \cdot \log_b \tfrac{1}{p(i)} + \sum_{i=1}^{n} p(i) \cdot \log_b \tfrac{1}{n} = \\ &= \sum_{i=1}^{n} p(i) \cdot (\log_b \tfrac{1}{p(i)} + \log_b \tfrac{1}{n}) = \\ &= \sum_{i=1}^{n} p(i) \cdot \log_b \tfrac{1}{p(i) \cdot n} \leq 0 \end{aligned}$$

{1}$\log_b n$ wurde mit $\sum_{i=1}^{n} p(i) = 1$ erweitert
{2}dasVorzeichen wurde beidesmal in den Logarithmus hineingezogen
Das heißt:

$$\sum_{i=1}^{n} p(i) \cdot \log_b \frac{1}{p(i) \cdot n} \leq 0, \text{ Gleichheit gilt, wenn } p(i) = \frac{1}{n}$$

Damit ist bewiesen, daß die Entropie maximal ist, wenn die Zufallsvariable gleichverteilt ist.

2.3 Verbundentropie

Die Verbundentropie ist einfach als die Entropie eines Zufallsvektors aufzufassen, dessen Komponenten Zufallsvariablen sind. Betrachtet man zum Beispiel die Zufallsvariablen X und Y, so ist die Verbundentropie von X und Y [H(X,Y)] :

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(xy)\cdot \log_b p(xy)$$

2.4 Bedingte Entropie

Die Bedingte Entropie ist die Unsicherheit über eine Zufallsvariable [Y], die verbleibt, wenn eine zweite Zufallsvariable [X] bereits bekannt ist. Sind die Zufallsvariablen voneinander unabhängig, so bleibt natürlich die Entropie der Unbekannten [Y] vollständig erhalten. Sind beide aber abhängig voneinander, das heißt, daß die Eine die Andere beeinflußt, so kann die Unsicherheit schon abnehmen.

$$\begin{aligned} H(Y|X) = &\sum_{x\in X} p(x) \overset{\{1\}}{\cdot} H(Y|X=x) = \\ = &\sum_{x\in X} p(x)\cdot(-\overset{\{2\}}{\sum_{y\in Y}} p(y|x)\cdot\log_b p(y|x)) = \\ =- &\sum_{x\in X} p(x)\cdot \overset{\{3\}}{\sum_{y\in Y}} \tfrac{p(yx)}{p(x)}\cdot\log_b p(y|x) = \\ =- &\sum_{x\in X}\sum_{y\in Y} \overset{\{4\}}{p(yx)}\cdot\log_b p(y|x) \end{aligned}$$

{1}es werden die nach der Wahrscheinlichkeit von x gewichteten Bedingten Entropien von [Y], unter X=x, über alle x aufsummiert

{2}für H(Y|X=x) wird die Formel, die die Entropie von [Y] unter der Vorbedingung, daß x empfangen wurde, berechnet, eingesetzt

{3}die Formel für die Bedingte Wahrscheinlichkeit wird eingesetzt

{4}da p(x) in der Summe über alle y eine Konstante ist, wird der Bruch mit p(x) gekürzt

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(yx)\cdot\log_b p(y|x)$$

Sind die Zufallsvariablen unabhängig, so ergibt sich:

$$\begin{aligned} H(Y|X) &= -\sum_{x\in X}\sum_{y\in Y} p(yx)\cdot\log_b p(y|x) = \\ &= -\sum_{x\in X}\sum_{y\in Y} p(yx)\cdot\log_b \tfrac{p(yx)}{p(x)} = \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y)\cdot p(x)\cdot\log_b \tfrac{p(y)\cdot p(x)}{p(x)} = \\ &= -\sum_{x\in X}\sum_{y\in Y} p(y)\cdot p(x)\cdot\log_b p(y) = \end{aligned}$$

$$=-\sum_{x\in X} p(x) \overset{\{1\}}{\sum_{y\in Y}} p(y)\cdot\log_b p(y)=$$
$$=-\sum_{y\in Y} \overset{\{2\}}{p(y)}\cdot\log_b p(y)= H(Y)$$

{1}p(x) ist auch hier einen Konstante, die sich aus der Summe entfernen läßt
{2}da $\sum_{x\in X} p(x) = 1$ kann man es ebenfalls verschwinden lassen

2.5 Relative Entropie

Die Relative Entropie läßt sich am ehesten als eine Art Abstand zwischen zwei Wahrscheinlichkeitsverteilungen beschreiben, sie ist also ein Maß für den Unterschied der Beiden. Sie ist kein richtiger Abstand, da sie weder die Dreiecksungleichung erfüllt noch symmetrisch ist.

Die Relative Entropie ist nie negativ und nur dann Null, wenn beide Verteilungen identisch sind. Ein anschauliches Beispiel wäre der Versuch, einen optimalen Kodierer für eine Quelle [X] zu entwerfen, ohne jedoch die genaue Wahrscheinlichkeitsverteilung der Zufallsvariablen [X] zu kennen. Um voranzukommen implementiert man den Kodierer auf der Basis der Zufallsvariablen [Y], deren Stichprobenraum die gleiche Mächtigkeit hat, wie der von [X]. Vorrausgesetzt, daß keine Ereignisse die Wahrscheinlichkeit Null haben. Die Relative Entropie gibt nun an, wieviele bits man durchschnittlich verschwendet, wenn man die genäherte Verteilung für [X] annimmt.

Ist nun p(x) die echte Verteilung von [X], und q(x) die Verteilung von [Y], so ist die Relative Entropie:

$$D(p||q) = \sum_{x\in X} p(x)\log_2 \frac{p(x)}{q(x)}$$

2.6 Gegenseitige Information

Der Begriff der Gegeseitigen Information ist komplementär zur Bedingten Entropie aufzufassen. Ist die Bedingte Entropie ein Maß für die verbleibende Unsicherheit über das nächste Zeichen einer Zufallsgröße [Y], wenn das Ergebnis von [X] bereits bekannt ist, so drückt die Gegenseitige Information den Zuwachs an Sicherheit über die noch unbekannte Größe [Y] aus.

$$I(X;Y) = H(Y) - H(Y|X)$$

Nach dieser Formel folgt:

$$I(X;Y) = -\sum_{y\in Y} p(y)\log_b p(y) - (-\sum_{y\in Y}\sum_{x\in X} p(yx)\cdot\log_b p(y|x)) =$$
$$= \sum_{y\in Y} p(y)\cdot\log_b \tfrac{1}{p(y)} + \sum_{y\in Y}\sum_{x\in X} p(yx)\cdot\log_b p(y|x) =$$
$$= \sum_{y\in Y}\sum_{x\in X} p(yx)\cdot\log_b \tfrac{1}{p(y)} \overset{\{1\}}{+} \sum_{y\in Y}\sum_{x\in X} p(yx)\cdot\log_b p(y|x)=$$

$= \sum_{y\in Y} \sum_{x\in X} p(yx) \cdot \log_b \frac{1}{p(y)} + p(yx) \cdot \log_b \frac{p(yx)}{p(x)} =$
$= \sum_{y\in Y} \sum_{x\in X} p(yx) \log_b \frac{p(yx)}{p(y)\cdot p(x)}$
{1}ergibt sich mit $p(y) = \sum_{x\in X} p(yx)$

$$I(X;Y) = \sum_{y\in Y} \sum_{x\in X} p(yx) \log_b \frac{p(yx)}{p(y)\cdot p(x)}$$

3 Kettenregel für die Verbundentropie

Die Verbundentropie zweier Zufallsvariablen [X] und [Y] entspricht der Entropie der Einen [H(X)] addiert zu der Bedingten Entropie der Anderen [H(Y|X)]. Sind die Zufallsvariablen unabhängig, so entspricht die Verbundentropie der Summe der beiden einzelnen Entropien.

$$H(X,Y) = H(X) + H(Y|X)$$

$H(X) + H(Y|X) =$
$= -\sum_{x\in X} p(x) \cdot \log_b p(x) + (-\sum_{x\in X} \sum_{y\in Y} p(xy) \log_b p(y|x)) =$
$\overset{\{1\}}{=} -\sum_{x\in X} \sum_{y\in Y} p(xy) \cdot \log_b p(x) - \sum_{x\in X} \sum_{y\in Y} p(xy) \log_b p(y|x) =$
$= -\sum_{x\in X} \sum_{y\in Y} (p(xy) \cdot \log_b p(x) + p(xy) \log_b p(y|x)) =$
$= -\sum_{x\in X} \sum_{y\in Y} p(xy) \cdot \log_b (p(x) \cdot p(y|x)) =$
$\overset{\{2\}}{=} -\sum_{x\in X} \sum_{y\in Y} p(xy) \cdot \log_b p(xy) =$
$= H(X,Y)$
{1}mit $p(x) = \sum_{y\in Y} p(xy)$ und $\log_b p(x)$ konstant in $\sum_{y\in Y}$
{2}mit $p(xy) = p(x) \cdot p(y|x)$

4 Anwendung der Informationstheorie

4.1 Elektrotechnik

Bevor Shannon mit der Entropie und deren Anwendung auf Nachrichtenübertragung bekannt wurde, war man allgemein der Meinung, daß mit steigender Datenübertragungsrate auch die Wahrscheinlichkeit einer fehlerhaften Übertragung stiege. Shannon zeigte aber, daß sich dieses Problem erst ergab, wenn die Entropie der übertragenen Nachricht die Kanalkapazität überstieg.

4.2 Physik

Auch in der Physik gibt es eine Größe, die Entropie heißt. Sie gibt an, wie groß die Unordung in einem physikalischen System ist. Befinden sich in einem geschlossenen Behälter Gasmoleküle, so ist die Entropie —die Unordnug der Gasmoleküle— von der Temperatur und dem Druck in dem Behälter abhängig. Mit steigender Temperatur bewegen sich die Moleküle immer schneller und die Unsicherheit über den Aufenthaltsort einzelner Moleküle steigt. Die Physik gebraucht den Begriff Entropie also analog.

4.3 Mathematik

Wie oben erwähnt, ist es möglich, mit Hilfe der Relativen Entropie Wahrscheinlichkeitsverteilungen untereinander zu vergleichen. Vor allem in der Statistik überdecken sich die Fachbereiche von Mathematik und Informationstheorie häufig.

4.4 Informatik

In der Informatik bezeichnet man das kleinste Programm, das einen String wiedergeben kann, als die Kolmogorow Komplexität dieses Strings. Interessanterweise nähert sich die Kolmogorow Komplexität eines Strings, der von einer Zufallsvariablen der Entropie Q stammt, mit wachsender Länge des Strings, immer weiter an die Entropie Q an.

Ein noch engerer Kontakt zwischen Informationstheorie und Informatik zeigt sich, wenn es um die Frage geht, ob es prinzipiell möglich wäre, immer schnellere Computer zu bauen. Diese Frage stößt nämlich dann auf die Entropie von Datenströmen, wenn man sich bewußt macht, daß mit steigender Datenverarbeitungsgeschwindigkeit auch der Datenstrom steigt. Dieser ist aber durch die Kanalkapazität in einem Bussystem eingeschränkt. Eine definitive Antwort auf diese Frage ist noch nicht gefunden.