Seminararbeit:

Information einer Verteilung

Autor: Antun Topic

Betreuerin: Prof. Christina Schneider

7. Januar 2015

Inhaltsverzeichnis

1 Begriffsklärung Information-Entropie

Der Begriff *Information* geht in dieser Arbeit über das Gebiet der induktiven Statistik hinaus und beruht auf dem Verständnis der Informationstheorie, die sich mit Problem- und Fragestellungen bei Nachrichtenübertragungen beschäftigt. Zu Grunde liegt ein Sender-Empfänger-System, eine Kommunikation zwischen diesen Parteien wird entsprechend vorausgesetzt. Der Sender verschickt eine Nachricht an den Empfänger, der bei einer störungsfreien Übertragung problemlos von der empfangenen Nachricht auf die gesendete Nachricht schließen kann. Meistens jedoch verläuft die Übertragung nicht reibungslos, sodass man anhand einer empfangenen Nachricht nur mit einer gewissen Wahrscheinlichkeit die ursprünglich versendete Nachricht rekonstruieren kann. (E.Henze; 1970, Seite 1) Desweiteren herrscht beim Empfänger vor dem Erhalt der Nachricht eine Unsicherheit über die zu erhaltende Nachricht, da er nicht weiß, was ihn erwartet und ob er die Informationen wird verarbeiten können. „Das System Sender/Empfänger befindet sich in einem Zustand der Unbestimmtheit."(B.Rüger; 1998, Seite 80) Diesem Umstand der Unbestimmtheit wird ein Maß zugrunde gelegt, welches als *Entropie* bezeichnet wird. Information bezeichnet hingegen die Beseitigung dieser Ungewissheit und Unkenntnis; die gewonnene Information ist also gleich der beseitigten Entropie. Anschaulich:

$$Gewonnene\ Information = Beseitigte\ Entropie$$

Da der Beseitigung der Entropie ein Gewinn an Information entspricht, müssten diese beiden Maße eigentlich mit unterschiedlichen Vorzeichen bezeichnet werden. Allerdings ist es üblich, diese beiden Maße gleichzusetzen, sodass *Information* = *Entropie* gilt, ohne deren unterschiedliche Bedeutung zu vernachlässigen. Der Sinn und Inhalt einer Information, der im alltäglichen Gebrauch des Begriffs von zentraler Bedeutung ist, findet in der Informationstheorie keine Berücksichtigung. (B.Rüger; 1998, Seiten 79-80)

2 Information von diskreten Verteilungen

Doch wenn in der Informationstheorie der Inhalt einer Information ausser Acht gelassen wird, stellt sich nun die Frage, wie der Informationsgehalt einer Nachricht dann gemessen wird. Dies erfolgt durch die Länge einer Nachricht, also durch die Anzahl der Zeichen, aus denen eine Nachricht besteht. Da es viele verschiedene Zeichensysteme gibt und diese für die Nachrichtenlängen von höchster Bedeutung sind, hat man sich auf das minimale Zeichensystem geeinigt, das nur aus Nullen und Einsen besteht. Kann eine Variable nur die Werte 0 und 1 annehmen, so nennt man sie ein *Binärzeichen*. Somit gilt: „Die Information(smenge) einer Nachricht soll durch die Anzahl der Binärzeichen gemessen werden, die zur Formulierung der Nachricht erforderlich sind." (B.Rüger; 1998, Seite 79)

2.1 Definition von Information

Der Gehalt an Information wird also in *bit* gemessen. Im Bezug auf die Entropie beschreibt ein Bit die Ungewissheit, ob als nächstes Zeichen 0 oder 1 folgt. Die Aufgabe der Informationsübertragung ist es also, diese Ungewissheit zu mindern und währenddessen Erkenntnisse über die Nachricht zu erhöhen. Für Informationen/Entropien von diskreten Verteilungen haben sich vor allem die beiden Definitionen beziehungsweise Formeln nach Hartley und Shannon durchgesetzt. L.Hartley war sehr verantwortlich für Untersuchungen, die nachrichtentechnische Fragestellungen beantworteten und schon in den zwanziger Jahren die logarithmische Struktur des Informations- und Entropiemaßes begründeten. Dies wird sich in den folgenden beiden Definitionen zeigen. C.E. Shannon gilt mit seinen Überlegungen und Publikationen Ende der 1940er Jahre als einer der Begründer der Informationstheorie. (Hofmann; 1973, Seite 6)

2.1.1 Formel von Hartley

Die erste Formel, die wir für Information bzw. Entropie betrachten möchten, geht auf Hartley zurück und stammt aus dem Jahr 1928. Es gelten folgende Annahmen: Der Empfänger weiß, dass sich die Nachricht aus einem Element ω der endlichen Grundmenge Ω zusammensetzt. Ausserdem liegen keine weiteren Vorkenntnisse über die Wahrscheinlichkeiten des Eintreffens eines Elementes vor. Somit ergibt sich für die Information $J(\{\omega\})$ der Nachricht ω oder die Entropie $H(\Omega)$ des Zustandes der Ungewissheit vor der Sendung die *Formel von Hartley*

$$J\{(\omega)\} = H(\Omega) = \log_2(N) \quad , \quad \mathrm{N} \;=\; \text{Anzahl der Elemente von } \Omega.$$

Dies soll die erste Definition eines Maßes für Information oder Entropie in diesem Bericht sein. Es gilt hierfür allerdings noch zu beweisen, dass mindestens approximativ $\log_2(N)$ Binärzeichen notwendig sind, um eine Nachricht ω zu formulieren. Die folgenden drei Abschnitte sollen diesen Beweis darstellen und neue Impulse geben:

1) Handelt es sich bei der Anzahl N der Elemente von Ω um eine Zweierpotenz, so gilt $N = 2^{\mathrm{m}}$. Folglich benötigt man genau $m = \log_2(N)$ Binärzeichen, um die Elemente von Ω darzustellen.

2) Handelt es sich bei der Anzahl N der Elemente von Ω um keine Zweierpotenz, so kann man N näherungsweise angeben durch $2^{\mathrm{m}\,-\,1} \leq N \leq 2^{\mathrm{m}}$. Es werden also wieder m Binärzeichen zur Darstellung der Elemente von Ω benötigt. Allerdings ist $\log_2(N)$ hier nur als Näherung zu verstehen, da $m \,-\, 1 \leq N \leq m$.

3) Diese Näherung kann allerdings auch verschärft werden, im Grenzfall liegt sogar Gleichheit vor. Nachrichten seien nun wieder Elemente von Ω, doch nun betrachten wir mehrmalige Übertragungen von Nachrichten. Diese Menge von k Einzelnachrichten heißt *Nachrichtenkette* der Länge k. „Eine solche ist darstellbar als Element von Ω^{k}, dem k-fachen kartesischen Produkt der Grundmenge Ω."(B.Rüger; 1998, Seite 81) Doch wie viele Binärzeichen benötigt man hierfür?
Werden die einzelnen Nachrichten, die Elemente von Ω sind, durch Binärzeichen dargestellt, so benötigt man nach 2) dafür $k \cdot m$ Binärzeichen für die Erstellung der Nachrichtenkette von k Einzelnachrichten.
Es gibt allerdings noch eine ökonomischere Methode, indem man die N^{k} Elemente von

Ω^{k} gleich durch Binärzeichen darstellt. Dafür sind n_{k} Binärzeichen notwendig, wenn n_{k} die natürliche Zahl ist mit $2^{n_{\mathrm{k}}-1} < N^k < 2^{n_{\mathrm{k}}}$. Dies ist gleichbedeutend mit $k \cdot \log_2 N < n_{\mathrm{k}} < k \cdot \log_2 N + 1$. Teilt man nun diesen Term durch k und lässt k gegen unendlich laufen, so folgt

$$\lim_{k\to\infty} \frac{n_{\mathrm{k}}}{k} = \log_2 N$$

Dies bedeutet also: "Die durchschnittliche Anzahl n_{k}/k von Binärzeichen, die pro Einzelnachricht in einer Nachrichtenkette der Länge k erforderlich sind, konvergiert mit wachsendem k gegen $\log_2 N$."(B.Rüger; 1998, Seite 81) Im Endeffekt reichen auf Dauer also 'fast' $\log_2 N$ Binärzeichen pro Einzelnachricht. (B.Rüger; 1998, Seiten 80-81)

2.1.2 Formel von Shannon

Bei der nun betrachtenden *Formel von Shannon* erhalten die möglichen Nachrichten, die Elemente der Grundmenge $\Omega = \{\omega_1, ..., \omega_{\mathrm{N}}\}$ sind, zusätzlich noch eine Wahrscheinlichkeitsverteilung $\pi = \{p_1, ..., p_{\mathrm{N}}\}$ auf Ω, die dem Empfänger ebenso bekannt ist. p_{i} ist dann die Wahrscheinlichkeit, mit der die Einzelnachricht ω_{i} gesendet wird $(i = 1, ..., k)$. Betrachten wir nun, wie sich die Definition von Information mit diesen zusätzlichen Vorkenntnissen ändert. Wir gehen nun wieder davon aus, dass es über die Wahrscheinlichkeiten, die einzelnen/bestimmten Nachrichten zu erhalten, keine genaueren Kenntnisse gibt. Das bedeutet, dass jedes Element von Ω der selben Wahrscheinlichkeit unterliegt, also gilt $p_{\mathrm{i}} = 1/N$ für jedes Element. Nach der Formel von Hartley gilt dann also für die Information der Nachricht ω_{i} folgendes: $J\{(\omega_{\mathrm{i}})\} = \log_2(N) = \log_2(1/p_{\mathrm{i}})$. Diesen Gedanken wollen wir nun auf beliebige Verteilungen $\pi = \{p_1, ..., p_{\mathrm{N}}\}$ übertragen. Das bedeutet nun, dass verschiedene Einzelnachrichten $\omega_1, ..., \omega_{\mathrm{N}}$ auch jeweils verschiedene Informationen erhalten: $J\{(\omega_1)\} = \log_2(1/p_1), ..., J\{(\omega_{\mathrm{N}})\} = \log_2(1/p_{\mathrm{N}})$. Je kleiner p_{i} ist, also je geringer die Wahrscheinlichkeit einer Nachricht ist, desto größer ist ihre Information. Um es nochmal in den Bezug zu den einführenden Begriffsklärungen am Anfang dieser Arbeit zu setzen: Die Information einer Nachricht wird also umso größer, je unwahrscheinlicher das Verschicken dieser Nachricht ist, da der Empfänger nicht mit einem Erhalt dieser Nachricht rechnet. Zwangsläufig herrscht eine große Entropie bezüglich dieser Nachricht, die dann um den selben Betrag reduziert wird um den Information gewonnen wird. Da wir uns in einer a priori Situation befinden (Situation vor der Nachrichtenübertragung) wollen wir allerdings noch einen Schritt weiter gehen und *eine* Mazahl für Information definieren, „die der Empfänger mit der Sendung einer Nachricht zu bekommen *erwarten*

kann"(B.Rüger; 1998, Seite 82) und mithilfe derer man das Sender-Empfänger-System bewerten kann. Dafür eignet sich der Erwartungswert der Informationen der einzelnen Nachrichten $J\{(\omega_1)\}, ..., J\{(\omega_N)\}$. Diesen definieren wir als Information $J(\pi)$ bzw. Entropie $H(\pi)$ der Verteilung π und erhalten somit die *Formel von Shannon*:

$$J(\pi) = H(\pi) = \sum p_i \log_2 \frac{1}{p_i} = -\sum p_i \log_2 p_i$$

Wenn $p_i = \frac{1}{N}$ gilt (Gleichverteilung von $\pi = (p_1, ...p_N)$ auf Ω), dann stimmt die Formel von Shannon mit der Formel von Hartley überein. Um die Formel von Shannon zu rechtfertigen, muss der Zusammenhang zwischen Information/Entropie und der Anzahl der Binärzeichen nachgewiesen werden. Auf Grund der verschiedenen Wahrscheinlichkeiten von Einzelnachrichten macht es wenig Sinn, jede Nachricht mit der gleichen Anzahl von Zeichen zu erstellen. Viel besser ist hingegen, für Einzelnachrichten eine Anzahl von Binärzeichen zu verwenden, die sehr nah (im Idealfall identisch) an $\log_2 \frac{1}{p_i}$ liegt $(i = 1, ..., N)$. Dafür müssen wir uns im nächsten Kapitel mit Codierungen beschäftigen. (B.Rüger; 1998, Seiten 81-82)

2.2 Binärcodierung

2.2.1 Codewortbildung

Sei $\Omega = \{\omega_1, ..., \omega_N\}$ das Alphabet und $\omega_1, ..., \omega_N$ seine Buchstaben. Die Buchstaben $\omega_1, ..., \omega_N$ bilden Wörter, aus denen dann die Sätze entstehen, die Nachrichten bilden. Die Buchstaben kommen mit den relativen Häufigkeiten, die der gegebenen Verteilung $\pi = (p_1, ...p_N)$ entsprechen, in der Nachricht vor. "Unter einer *Codierung* (Binärcodierung) von Ω, genauer von (Ω, π), versteht man eine eineindeutige Darstellung der Buchstaben von Ω durch Kombinationen, genauer: Tupel aus Nullen und Einsen."(B.Rüger; 1998, Seiten 82-83) Das Codewort oder der Code von ω_i bezeichnet dann das m_i-Tupel, das innerhalb der Codierung den Buchstaben ω_i darstellt. Die Länge des Codeworts wird dann mit m_i bezeichnet. Bildet man folgenden Erwartungswert aus der Länge des Codewortes und dessen Wahrscheinlichkeit

$$m = \sum m_i p_i \ ,$$

so wird dieser die *mittlere Codewortlänge* (Codierungslänge) der jeweiligen Codierung genannt. Je kleiner dieses m ist, desto ökonomischer (folglich auch günstiger) ist die Codierung. Eine Codierung, die vom Empfänger erfolgreich und fehlerfrei decodiert werden kann, nennt man *entzifferbar*. Dies ist nicht immer der Fall, da die Null-Eins-Tupel ohne Abstände aneinander gereiht werden und der Empfänger somit nicht immer erkennen kann, wann ein Codewort aufhört und wann das nächste bereits anfängt. Dieses Problem kann man umgehen, „wenn innerhalb der Codierung kein Codewort mit dem *Anfang* eines anderen, längeren Codewortes übereinstimmt. Eine solche Codierung heißt *irreduzibel*." (B.Rüger; 1998, Seiten 83) Es ist ersichtlich, dass jede irreduzible Codierung entzifferbar ist, jedoch muss nicht jede entzifferbare Codierung irreduzibel sein.
Betrachten wir dazu folgendes Beispiel:
Gegeben sei das Alphabet $\Omega = \{A, B, C, D, E, F, G, H\}$ mit seiner Verteilung $\pi = (p_1, ...p_N) = (\frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{4})$. Nach der Formel von Shannon erhält man für die Information bzw. Entropie $J(\pi) = H(\pi) = 2.75$. Folgende Tabelle soll die Codierungen I bis IV und deren Codewortbildungen anschaulich erklären:

		I		II		III		IV	
i	ω_i	Codewort	Länge	Codewort	Länge	Codewort	Länge	Codewort	Länge
1	A	000	3	100	3	001	3	000	3
2	B	001	3	1100	4	0011	4	001	3
3	C	010	3	1101	4	1011	4	010	3
4	D	100	3	101	3	101	3	100	3
5	E	011	3	00	2	00	2	01	2
6	F	101	3	1110	4	0111	4	101	3
7	G	110	3	1111	4	1111	4	110	3
8	H	111	3	01	2	10	2	11	2
		$m = 3$		$m = 2.75$		$m = 2.75$		$m = 2.5$	
Entzifferbar?		Ja		Ja		Ja		Nein	
Irreduzibel?		Ja		Ja		Nein		Nein	

(B.Rüger; 1998, Seiten 83-84)

2.2.2 Zusammenhang zwischen Aufwand und Information der Codierung

Den Zusammenhang zwischen dem Aufwand/der Länge einer Codierung und der Information/Entropie zeigen die folgenden drei Aussagen. Gleichzeitig rechtfertigen wir hiermit die Formel von Shannon.
1) Für jede entzifferbare Codierung gilt: $m \geq H(\pi)$
2) Es gibt genau dann eine entzifferbare Codierung mit $m = H(\pi)$, wenn jedes p_i in π eine ganzzahlige Potenz von $\frac{1}{2}$ ist.
3) Es gibt stets eine irreduzible (mithin auch entzifferbare) Codierung mit $H(\pi) \leq m < H(\pi) + 1$.

$H(\pi)$ ist also, nach Aussage 1), gemessen mit der mittleren Codewortlänge einer entzifferbaren Codierung, eine untere Schranke für den Codierungsaufwand. Nach Aussage 2) ist diese Schranke scharf, allerdings nur unter den angegeben Voraussetzungen. Aussage 3) besagt, dass es zu jeder beliebigen Verteilung π eine passende Codierung gibt, deren mittlere Codewortlänge maximal in einem Intervall der Länge 1 über $H(\pi)$ liegt. Die Grenzwertaussage 3) der Hartley-Information $J\{(\omega)\}$ lässt sich mit Aussage 3) auch auf die Shannon-Information $H(\pi)$ anwenden: $\lim_{k\to\infty} \frac{m(k)}{k} = H(\pi)$
$m(k)$ ist also die mittlere Codewortlänge einer Codierung von Nachrichten, die aus k Buchstaben des Alphabets Ω bestehen. Als Verteilung auf Ω^k liegt π^k zu Grunde. Mit wachsendem k entsteht also ein Folge von Codierungen, unter denen es dann solche gibt, die auf lange Sicht sehr günstig sind, deren mittlere Codewortlänge pro Buchstabe im Grenzwert gegen $H(\pi)$ läuft. (B.Rüger; 1998, Seiten 84-85)

3 Information von beliebigen Verteilungen

In diesem Kapitel soll der Entropie- bzw. Informationsbegriff von diskreten Verteilungen auf beliebige Verteilungen ausgedehnt werden. Wir legen eine Zufallsgröße X und ihre Verteilung P mit v fest. Für diskrete Fälle ist nun $f(x)$ die Wahrscheinlichkeitsfunktion von X und v das Zählmaß auf dem Träger von P. Für stetige Fälle ist v das Lebesgue-Maß und $f(x)$ die (Lebesgue-)Dichte von X. Weiterhin bezeichnet $H(X)$ die Unsicherheit/Unbestimmtheit über die zukünftige, durch die Verteilung P gegebene, Beobachtung x von X. Die Information $H(X)$ stellt wieder die Beseitigung dieser Unkenntnis dar, also die durch die Beobachtung x von X gewonnene Kenntnis. Um Information/Entropie von stetigen Verteilungen herzuleiten, müssen wir zunächst noch einen Blick auf die diskrete Ein- und Zweipunktverteilung werfen.

3.1 Herleitung der Entropie für beliebige Verteilungen

Die vereinfachte, bisher behandelte Vorstellung, dass $H(X)$ die Anzahl der Binärzeichen oder die mittlere Codewortlänge erfasst, eignet sich nicht dazu, die Entropie/Information für beliebige Verteilungen einzuführen. Allgemein kann X nämlich unendlich viele verschiedene Werte annehmen, was zur Folge hätte, dass unendlich viele Binärzeichen benötigt wären und $H(X) = \infty$ gelten würde. Dann hätten jedoch alle Verteilungen mit einer stetigen Verteilungsfunktion die Entropie 'Unendlich'. Um die Entropie als ein vergleichbares Maß beizubehalten, hat man sich in der Folge entschieden, einen anderen Nullpunkt für die Entropie zu wählen, sollte X unendlich viele verschiedene Werte annehmen. Daraus folgt, dass die *Entropie der Einpunktverteilung den Wert Null hat.*
Leider hat diese Vorstellung der Entropie im allgemeinen Fall keine Substanz mehr. Somit ergibt sich für die *Entropie der stetigen Gleichverteilung auf dem Einheitsintervall der Wert Null.* Folglich haben Verteilungen, die informativer sind als die Gleichverteilung, eine negative Entropie. Nun werden die Entropien allerdings auf verschiedenen

Skalen gemessen und sind dadurch nicht vergleichbar. Es gibt aber auch zwei Vorteile dieser Darstellung:
1) Die Entropie vieler stetiger Verteilungen ist nun endlich.
2) Es liegt eine einheitliche Form der Entropie *H(X)* vor, basierend auf der Formel von Shannon. Für eine Zufallsgröße X mit der v-Dichte *f(x)* definieren wir für beliebige Verteilungen als *Entropie* oder *(Shannon-)Information (der Verteilung) von* X die Größe

$$H(X) = -\int f(x)\log_2(f(x))dv = -E\log_2(f(X)).$$

Diese Gleichung ist unsere *allgemeine Entropiedefinition für beliebige Verteilungen.* Hierdurch wird die a priori Situation beschrieben. Für die a posteriori Situation definiert man $J(x) := -\log_2(f(X))$ als die (Shannon-)Information der Beobachtung x von X, folglich wird dann *H(X)* zum Erwartungswert von *J(X).*

Da wir nun auf den Zusammenhang zwischen Codierung und Entropie verzichten, können wir im nächsten Schritt den Logarithmus dualis $'\log_2'$ durch den natürlichen Logarithmus $'log'$ ersetzen.
H(X) hängt nur von der Verteilung von X ab, nicht aber von den Momenten der Verteilung, wie zum Beispiel *EX* und *VarX*. „In die Parameter *EX* und *VarX* gehen nämlich die Werte x, die von X angenommen werden, auf entscheidende Weise mit ein (*EX* soll ja die Lage und *VarX* die Streuung dieser Werte beschreiben), während das für den Parameter *H(X)* ... nicht der Fall ist."(B.Rüger; 1998, Seite 87)
Sei X eine Zufallsgröße, die N verschiedene Werte annehmen kann. Dann hängt *H(X)* nach der Formel von Shannon nur von N und den Wahrscheinlichkeiten $p_1, ..., p_N$ der Werte ab, nicht aber von den Werten selbst. Somit ist es für *H(X)* im Gegensatz zur *VarX* unbedeutend, wie groß die Abstände zwischen den Werten von X sind; es besteht also keine Monotoniebeziehung zwischen beiden. (B.Rüger; 1998, Seiten 85-87)

Dazu nun der Beweis anhand einer Zweipunktverteilung.
Sei X eine reelwertige Zufallsgröße, welche die beiden Werte x_1 und x_2 mit den Wahrscheinlichkeiten p bzw. *1 - p* annehmen kann. Ihre Entropie *H(X)* ist unabhängig von x_1 und x_2 und lautet:

$$H(X) = -[p \log_2(p) + (1-p) \log_2(1-p)].$$

Ihre Varianz von X ist abhängig von x_1 und x_2 und lautet:

$$VarX = -p(1-p)(x_2 - x_1)^2.$$

Durch Variation von p, x_1 und x_2 lassen sich Verteilungen bilden, für die gilt: Je größer die Entropie, desto kleiner die Varianz.
Da bei allgemeinen (stetigen) Verteilungen durch die Bildung des v-Integrals auch die Abstände zwischen den x-Werten einen Einfluss erhalten (diese werden mit dem dominierenden Maß v gemessen), verwischen die Unterschiede zwischen Varianz und Entropie. Für stetige Verteilungen geht die Entropie von X über in

$$H(X) = -\int f(x) \log_2(f(x)) dx$$

und bildet für ein reelwertiges X also eine Fläche ab, welche die Abstände der x-Werte durchaus beeinflusst. „Daher ist hier im Unterschied zum Fall einer diskreten Verteilung mit endlichem Träger die *Entropie nicht invariant gegenüber einer eineindeutigen Transformation von X.*"(B.Rüger; 1998, Seite 88) Dies kann als Defizit des Entropiebegriffs angesehen werden. Welche Verteilungen können dann mit dem Entropiemaß verglichen werden? „Der Vergleich zweier Entropien ist *nur dann* sinnvoll, wenn in der Entropiedefinition für beliebige Verteilungen mit demselben dominierenden Maß v integriert wird."(B.Rüger; 1998, Seite 88) Der Vergleich von Entropien innerhalb der Klasse aller (Lebesgue-)stetigen Verteilungen ist stets angebracht. (B.Rüger; 1998, Seiten 87-88)

3.2 Entropie einiger stetiger Verteilungen

a) Für eine auf dem Intervall [a;b] *gleichverteilte* Zufallsvariable X beträgt die Entropie:

$$H(X) = -E\log_2(1/(b-a)) = \log_2(b-a)$$

$\rightarrow$ maximale Entropie unter allen (Lebesgue-)stetigen eindimensionalen Verteilungen

b) Für eine mit dem Parameter λ *exponentialverteilte* Zufallsvariable X beträgt die Entropie:

$$H(X) = -E\log_2(\lambda e^{-\lambda X}) = -\log_2\lambda + \lambda EX\log_2 e = \log_2(e/\lambda)$$

$\rightarrow$ maximale Entropie auf der positiven Halbachse unter allen stetigen Verteilungen mit dem vorgegebenen Erwartungswert $\mu = 1/\lambda$

c) Für eine mit den Parametern μ und σ^2 *normalverteilte* Zufallsvariable X beträgt die Entropie:

$$\begin{aligned} H(X) &= -E\log_2\left(\frac{1}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{1}{2}\left\{\frac{X-\mu}{\sigma}\right\}^2\right\}\right) \\ &= \log_2\sqrt{2\pi\sigma^2} + \frac{1}{2}E\left\{\frac{X-\mu}{\sigma}\right\}^2\log_2 e \\ &= \log_2\sqrt{2\pi e\sigma^2} \end{aligned}$$

$\rightarrow$ maximale Entropie unter allen stetigen Verteilungen mit dem vorgegebenen Erwartungswert μ und vorgegebener Varianz σ^2 (B.Rüger; 1998, Seiten 88-89)

3.3 Aussagen über gemeinsame und endliche Entropien

Seien X und Y zwei Zufallsgrößen im selben Wahrscheinlichkeitsraum. Durch die allgemeine Entropiedefinition ist dann auch die *gemeinsame Entropie H(X,Y)* von X und Y erklärt. Hierzu muss man anstatt *f(x)* die gemeinsame Dichte *f(x,y)* von X und Y einsetzen und folglich auch den Erwartungswert mit der gemeinsamen Verteilung von X und Y bilden.
Auch die *bedingte Entropie* $H(X \mid Y)$ von X und Y wird durch die allgemeine Entropie-

definition erklärt. In diesem Fall muss man die bedingte $f(x \mid y)$ einsetzen, den Erwartungswert unter der Bedingung $Y = y$ bilden und abschließend die Erwartungsbildung mit der Randverteilung Y vornehmen.
Ist diese gemeinsame Entropie $H(X, Y)$ endlich, dann sind es auch die Entropien $H(X)$, $H(Y)$ und $H(X \mid Y)$. Folgende Aussagen treffen gelten:
a)

$$H(X \mid Y) = H(X, Y) - H(Y)$$

b)

$$H(X \mid Y) \leq H(X),$$

wobei das Gleichheitszeichen genau dann gilt, wenn X und Y unabhängig sind.
c)

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

falls X und Y unabhängig sind. (Additivität der Entropie)

Die Definition von Entropie für beliebige Verteilungen bedarf noch einer strikteren Rechtfertigung, als nur die Begründung der Verallgemeinerung der Formel von Shannon. Diese ist nun durch die Additivität der Entropie gegeben. (B.Rüger; 1998, Seiten 89-90)

Literaturverzeichnis

B.Rüger (1998). Test- und Schätztheorie.

E.Henze (1970). Einführung in die Informationstheorie, Braunschweig.

Hofmann, K.-D. (1973). Einführung in die Informationstheorie.