

Mathematische Grundlagen III

Informationstheorie

Garance PARIS

20. Juni 2011

Informationstheorie

- Ein Gerüst, um über den Informationsgehalt von (linguistischen) Ereignissen nachzudenken
- Einige Beispiele für Anwendungen:
 - Einen sparsamen Code finden, um Information zu übertragen
 - Sprachmodelle evaluieren und vergleichen
 - Die Assoziationsstärke zwischen Wörtern beurteilen, um Kollokationen zu entdecken
 - Spracherkennung, OCR, Rechtschreibkorrektur, Schrifterkennung, maschinelle Übersetzung, ...

Entropie (H oder E)

Anzahl von Bits, die benötigt werden, um den Stand eines Systems vollständig anzugeben

Metapher

- Wieviele Ja/Nein-Fragen braucht man, um das Ergebnis eines Münzwurfs herauszufinden?
- Wieviele braucht man bei zwei Münzen? ...3, 4...?

Entropie (H oder E)

Anzahl von Bits, die benötigt werden, um den Stand eines Systems vollständig anzugeben

Metapher

- Wieviele Ja/Nein-Fragen braucht man, um das Ergebnis eines Münzwurfs herauszufinden?
- Wieviele braucht man bei zwei Münzen? ...3, 4...?

Mögliche Zustände = $2^{\text{Anzahl Fragen}}$

$\Leftrightarrow \log_2 \text{Zustände} = \text{Anzahl Fragen}$

2 Zustände	1 Fragen
4 Zustände	2 Fragen
8 Zustände	3 Fragen
16 Zustände	4 Fragen

Metapher, Teil 2: Verschiedene Wurfel



$$\log_2 4 = 2$$

$$\log_2 6 = 2.585$$

$$\log_2 8 = 3$$

$$\log_2 12 = 3.585$$

$$\log_2 20 = 4.322$$

Entropie bei Gleichverteilung

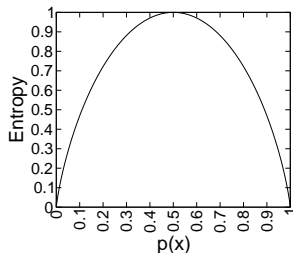
In einem System mit n gleich wahrscheinlichen Zustanden gilt:

$$\begin{aligned} H &= \log_2(n) \\ &= -\log_2(1/n) \\ &= -\log_2(P(n)) \end{aligned}$$

Andere Sichtweise

- Durchschnittliche Unsicherheit darüber, welches Ereignis als nächstes stattfindet
- Ein Maß dafür, wieviel Unordnung oder Zufälligkeit in einem System oder Zufallsvariable enthalten ist

Entropie einer gezinkten Münze



Gesamtentropie unabhängiger Elemente

z. B. Entropie dreier Würfel mit 4, 6, und 12 Seiten:

$$\begin{aligned}H &= \log_2(4 * 6 * 12) \\ &= \log_2(4) + \log_2(6) + \log_2(12) \\ &= 8.170 \text{ bits}\end{aligned}$$

Die Gesamtentropie ist also die **Summe** der unabhängigen Elemente im System

Ungleichverteilte Wahrscheinlichkeitsfunktionen

- Bei Ungleichverteilung werden die Summanden nach ihrer Wahrscheinlichkeit gewichtet; man erhält also:

$$H(x) = - \sum_x p(x) \log_2 p(x)$$

- $0 \log_2 0$ wird gleich 0 gesetzt
- Durch die allgemeinen Logarithmusrechenregeln kann die Formel auch umgeformt werden:

$$H(x) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

- Außerdem gilt bei einem Korpus oder Datensatz der Größe N :

$$\frac{N \log_2 N - \sum_x x \log_2 x}{N}$$

Bsp.: Wurf dreier Münzen

Wie oft kommt Kopf vor,
wenn ich eine Münze dreimal werfe?

$$P(X) = \begin{cases} x=3 & 1/8 \\ x=2 & 3/8 \\ x=1 & 3/8 \\ x=0 & 1/8 \end{cases}$$

$$\begin{aligned} H(X) &= \sum_x p(x) \log_2 \frac{1}{p(x)} \\ &= 2 \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} + 2 \frac{3}{8} \log_2 \frac{1}{\frac{3}{8}} \\ &= \frac{1}{4} \log_2 8 + \frac{3}{4} \log_2 \frac{8}{3} \\ &= \frac{1}{4} \cdot 3 + \frac{3}{4} (\log_2 8 - \log_2 3) \\ &= \frac{3}{4} + \frac{3}{4} (3 - \log_2 3) \\ &= 1.811 \text{ bits} \end{aligned}$$

Weiter...

- Untere Schranke für Anzahl von Bits, die benötigt werden, um eine Nachricht zu senden
- Eine Angabe darüber, wieviel Information eine Nachricht enthält

Der Zustand als Nachricht

Wie kann man das Ergebnis des Wurfs eines 8-seitigen Würfels in Bits ausdrücken?

Eine Möglichkeit:

1	000		3	010		5	100		7	110
2	001		4	011		6	101		8	111

- Es werden drei Bits gebraucht, um das Ergebnis als Nachricht mitzuteilen
- In diesem Fall entspricht dies der Entropie: $\log_2(8) = 3$

Der Zustand als Nachricht, Teil 2

- Der *Informationsgehalt* eines Ereignisses, der ein Teil der Entropieformel bildet, gibt die optimale Länge des Codes für ein Ereignis

$$I(p(x)) = \log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

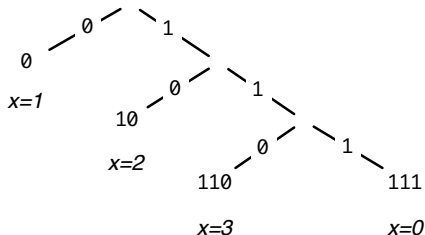
- Bei einem 8-seitigen Würfel: $8 \frac{1}{8} \log_2(8) = \log_2(8) = 3$
- Bei drei Münzen:

$$\text{Für } x = 1 \text{ oder } 2 : \log_2 \frac{1}{3/8} = 1.415$$

$$\text{Für } x = 0 \text{ oder } 3 : \log_2 \frac{1}{1/8} = 3$$

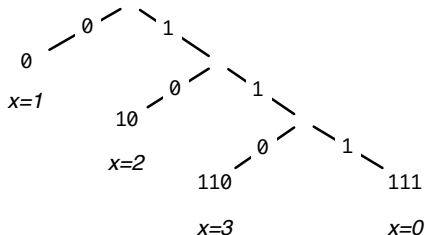
Der Zustand als Nachricht, Teil 3

An Hand des Informationsgehalts kann man mit einem Baum einen effizienten Code für diese "Sprache" entwickeln:



Der Zustand als Nachricht, Teil 3

An Hand des Informationsgehalts kann man mit einem Baum einen effizienten Code für diese "Sprache" entwickeln:



Für die "Nachricht" $x = 1$ werden weniger Bits verwendet,
für $x = 2$ etwas mehr

Der Zustand als Nachricht, Teil 4

- Der durchschnittliche Anzahl von Bits für die Übertragung einer Nachricht mit diesem Code ist höher als die von der Entropie dargestellten untere Schranke (1.811)
- Um ihn zu berechnen, wird $l(x)$ durch die Länge des Codes in der Entropieformel ersetzt:

$$\sum_x p(x) * \text{code length}(x) = 3\frac{1}{8} + 2\frac{3}{8} + \frac{3}{8} + 3\frac{1}{8} = 1.875$$

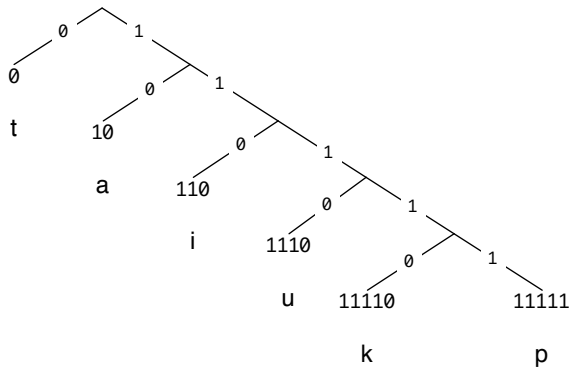
Linguistisches Beispiel: "Simplified Polynesian"

p	t	k	a	i	u
1/16	3/8	1/16	1/4	1/8	1/8

$$\begin{aligned}H(X) &= \sum_x p(x) \log_2 \frac{1}{p(x)} \\ &= 2 * \frac{1}{16} \log_2 16 + 2 * \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 + \frac{3}{8} \log_2 \frac{8}{3} \\ &= 2.28 \text{ bits}\end{aligned}$$

Beispiel: "Simplified Polynesian", Teil 2

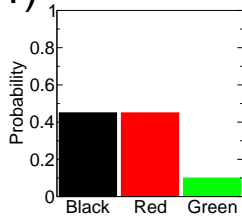
	p	t	k	a	i	u
$P(x)$	$1/16$	$3/8$	$1/16$	$1/4$	$1/8$	$1/8$
$l(x)$	4	1.415	4	2	3	3



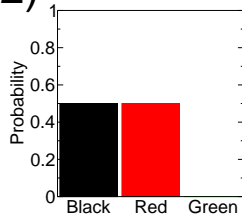
Einige Eigenschaften

- Bestenfalls hat ein Ereignis die Wahrscheinlichkeit 1
Dann ist $H = 0$
- Der ungünstigster Fall tritt bei Gleichverteilung ein
(Wenn die Variable binär ist, ist dann $H = 1$)
- Flache, breite Verteilungen haben eine hohe Entropie
- Spitze, schmale, kompakte Verteilungen
haben eine niedrige Entropie
- Da die Entropie unabhängiger Elemente addiert wird,
wächst die Entropie mit der Anzahl der Dimensionen
einer Verteilung

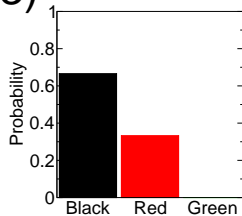
1)



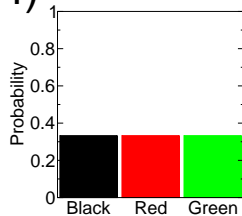
2)



3)



4)



Gemeinsame Entropie

Wieviel Information benötigt wird, um den Wert zweier Zufallsvariablen anzugeben

$$H(X, Y) = H(p(x, y)) = - \sum_x \sum_y p(x, y) \log_2 p(x, y)$$

Bedingte Entropie

Wieviel Information benötigt wird, um Y mitzuteilen, wenn X bekannt ist

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= \sum_x p(x) \left[- \sum_y p(y|x) \log_2 p(y|x) \right] \\ &= - \sum_x \sum_y p(x, y) \log_2 p(y|x) \end{aligned}$$

Polynesisch, Teil 2

Wir finden heraus, dass alle Wörter der oben betrachteten Sprache aus CV-Folgen bestehen, mit folgenden Wahrscheinlichkeiten:

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

$$\begin{aligned}
 H(C, V) &= 4 * \frac{1}{16} \log_2 16 + 6 * \frac{1}{16} \log_2 \frac{16}{3} + \frac{3}{8} \log_2 \frac{8}{3} \\
 &= 2.436
 \end{aligned}$$

pro Silbe

Entropie Rate

Da die Entropie von der Länge der Nachricht abhängt, ist es häufig sinnvoll zu normalisieren

$$H(x) = -\frac{1}{n} \sum_x p(x_{1\dots n}) \log_2 p(x_{1\dots n})$$

Im Polynesischen...

2.436/2 = 1.218 bits **pro Buchstabe**

(zu vergleichen mit 2.28 ohne Wissen über Silben)

Entropie kann auch verwendet werden, um die Qualität eines Modells zu beurteilen

Relative Entropy

- Die Relative Entropie, auch *Kullback-Leibler (KL) Divergenz* genannt, vergleicht die Entropie zweier Verteilungen
- Intuitiv: Durchschnittliche Anzahl Bits, die verschwendet werden, wenn eine Verteilung p mit einem für q entwickelten Code enkodiert wird
- Nicht symmetrisch!

$$D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

Kreuz-Entropie

- Wenn wir ein Modell m eines Phänomens p bauen, wollen wir $D(p||m)$ niedrig halten
- Kreuzentropie:

$$\begin{aligned}H(X, m) &= H(X) + D(p||m) \\ &= \sum_x p(x) \log_2 \frac{1}{m(x)}\end{aligned}$$

- Problem: Um die Kreuzentropie berechnen zu können, brauchen wir die Verteilung von $p(x)$
- Es wird angenommen, dass Sprache *ergodisch* ist, d.h. dass jede Stichprobe repräsentativ des Ganzen ist

Die Entropie von Englisch

ASCII	8	$\log_2 256$
Uniform	4.76	$\log_2 27$
Unigram	4.03	
Bigram	2.8	
Gzip	2.5	
Trigram	1.76	Brown et al., 1992
Human	1.3	Shannon (<i>Shannon guessing game</i>)
	1.34	Cover & Thomas (using gambling)

Die Kettenregel

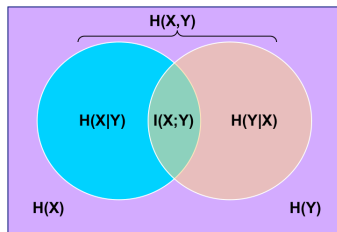
$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1, \dots, X_{n-1})$$

Mutual Information

Die Reduzierung der Unsicherheit in einer Variable, dadurch dass man über einer anderen Variable Information hat

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



MI berechnen

$$I(X; Y) =$$

$$H(X) - H(X|Y)$$

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

$$\sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_y p(y) \log_2 \frac{1}{p(y)} + \sum_{xy} p(x, y) \log_2 p(x, y)$$

$$\sum_{xy} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Eigenschaften

- Symmetrisch, immer positiv
- Maß für die Abhängigkeit zweier Verteilungen:
 $I(X; Y) = 0$ wenn X, Y unabhängig
- Wächst sowohl mit Abhängigkeit als auch mit Entropie

Pointwise Mutual Information

- Maß für die Assoziationsstärke zweier Elemente
- Information,
die in einem Ereignis x über ein Ereignis y enthalten ist

$$\begin{aligned}I(x, y) &= \log_2 \frac{p(x, y)}{p(x)p(y)} \\ &= \log_2 \frac{p(x|y)}{p(x)} \\ &= \log_2 \frac{p(y|x)}{p(y)}\end{aligned}$$

Beispiel 1: Kollokationen

Word 1	Word 2	C(w1)	C(w2)	C(w1 w2)	I(w1,w2)
Ayatollah	Khomeini	42	20	20	18.38
Agatha	Christie	30	117	20	16.31
cassette	recorder	77	59	20	15.94
unsalted	butter	24	320	20	15.19
over	many	13484	10570	20	1.01
like	people	14093	14776	20	0.46
time	last	15019	1569	20	0.29

(N = 14307688)

- Pointwise MI kann verwendet werden, um mögliche Kollokationen zu bewerten:

$$I(\text{cassette}, \text{recorder}) = \log_2 \frac{\frac{20}{N}}{\frac{77}{N} \frac{59}{N}} = 15.94$$

- Durch das Ergebnis werden die Wortpaaren geordnet, wie bei statistischen Tests (allerdings ohne Theorie darüber, wo das Threshold zu setzen ist)