Mathematische Grundlagen III

Informationstheorie

Prof Dr. Matthew Crocker

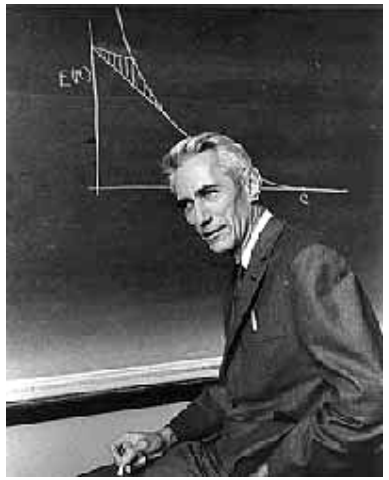
Universität des Saarlandes

22. Juni 2015

Informationstheorie

Entropie (H)

- Wie viel Information kann man über eine Leitung mit Störungen (z.B. Telefonleitung) übertragen?
- Wie stark kann eine Nachricht maximal verdichtet werden? (Wie viele bits braucht man mindestesn um eine Nachricht zu übertragen?)
- Wie schnell kann eine Nachricht durch eine bestimmte Leitung übertragen werden, ohne zu viel Information zu verlieren?



Claude Shannon

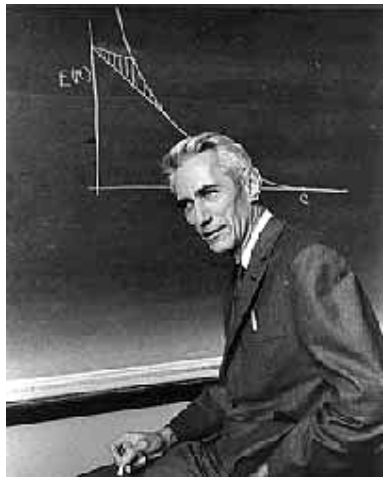
Beispiel: Münzen

- Information die wir übertragen wollen: Zustand einer Münze
- mögliche Zustände: Kopf oder Zahl
- Wie viele ja-nein Fragen (bits) braucht man um den Zustand einer bestimmten Münze zu beschreiben?

Informationstheorie

Entropie (H)

- Wie viel Information kann man über eine Leitung mit Störungen (z.B. Telefonleitung) übertragen?
- Wie stark kann eine Nachricht maximal verdichtet werden? (Wie viele bits braucht man mindestesn um eine Nachricht zu übertragen?)
- Wie schnell kann eine Nachricht durch eine bestimmte Leitung übertragen werden, ohne zu viel Information zu verlieren?



Claude Shannon

Beispiel: Münzen

- Information die wir übertragen wollen: Zustand einer Münze
- mögliche Zustände: Kopf oder Zahl
- Wie viele ja-nein Fragen (bits) braucht man um den Zustand einer bestimmten Münze zu beschreiben?

Inhaltsverzeichnis

1. Entropie eines einzelnen Ereignisses
2. Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)
3. Gegenseitige Information (Mutual Information)
4. Noisy Channel Model
5. Relative Entropie (Kullback-Leibler Divergenz)
6. Kreuz-Entropie

Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

6 Kreuz-Entropie

Entropie

Wie viele bits braucht man, um eine Zahl zwischen 1 und 8 zu kodieren?

2

Entropie

Wie viele bits braucht man, um eine Zahl zwischen 1 und 8 zu kodieren?

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

$$\text{Mögliche Zustände} = 2^{\text{Anzahl Fragen}}$$
$$\Leftrightarrow \log_2 \text{Zustände} = \text{Anzahl Fragen}$$

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

Entropie

Wie viele bits braucht man, um eine Zahl zwischen 1 und 8 zu kodieren?

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

$$\text{Mögliche Zustände} = 2^{\text{Anzahl Fragen}}$$
$$\Leftrightarrow \log_2 \text{Zustände} = \text{Anzahl Fragen}$$

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

Entropie

Verschiedene Würfel

$\log_2 4 = 2$

$\log_2 6 = 2.585$

$\log_2 8 = 3$

$\log_2 12 = 3.585$

$\log_2 20 = 4.322$

Gesamtentropie unabhängiger Elemente

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

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

Entropie

Entropie bei Gleichverteilung

In einem System mit n gleich wahrscheinlichen Zuständen gilt:

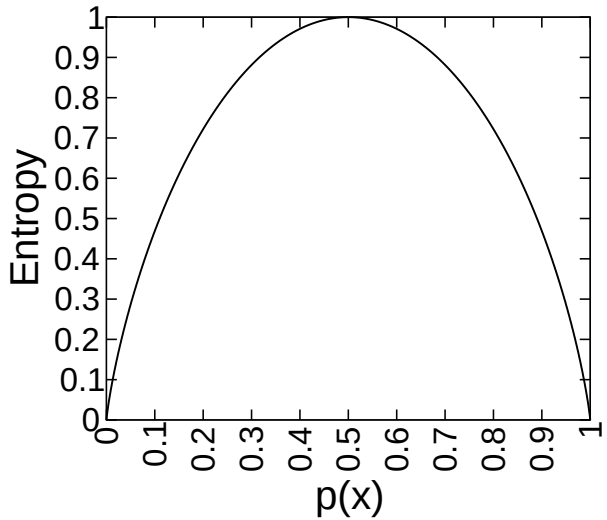
$$\begin{aligned} H &= \log_2(n) \\ &= -\log_2(1/n) \\ &= -\log_2(P(n)) \end{aligned}$$

Entropie

Andere Perspektive

- Durchschnittliche Unsicherheit darüber, welches Ereignis als nächstes stattfindet
- Ein Maß dafür, wieviel Unordnung oder Zufälligkeit in einem System oder einer Zufallsvariable enthalten ist
- Wie leicht ist es, den nächsten Zustand eines Systems zu erraten?

Entropie einer gezinkten Münze



Entropie

Nicht gleichverteilte Wahrscheinlichkeitsfunktionen:

- Summanden nach ihrer Wahrscheinlichkeit gewichtet; man erhält also:

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

- Durch die allgemeinen Logarithmusrechenregeln kann die Formel auch umgeformt werden:

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

- $0 \log_2 0$ wird gleich 0 gesetzt

Entropie

Bsp.: Wurf dreier Münzen

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

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

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

Informationstheorie in der CL

- Ermöglicht uns, den Informationsgehalt von (linguistischen) Ereignissen zu quantifizieren.
- Einige Beispiele für Anwendungen:
  - Einen sparsamen Code finden mit dem Information übertragen werden können.
  - Sprachmodelle evaluieren und vergleichen
  - Die Assoziationsstärke zwischen Wörtern beurteilen, um Kollokationen zu entdecken
  - Spracherkennung, Rechtschreibkorrektur, Schrifterkennung, maschinelle Übersetzung, ...

Entropie

Weiter...

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

Entropie

Der Zustand als Nachricht

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

Eine Möglichkeit:

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

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

Entropie

- Linguistisches Beispiel: Entropie einer Sprache
- Polynesische Sprachen wie Hawaiisch sind für ihre geringe Anzahl an Lauten bekannt.

Beispiel: "Vereinfachtes Polynesisch"

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

Vereinfachtes Polynesisch

Beispiel: "Vereinfachtes Polynesisch"

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

Entropie für vereinfachtes Polynesisch

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

- Entropy wie Fragespiel: Man kann ja/nein-Fragen stellen (z.B. "Ist es ein *a* oder ein *t*?" "Ist es ein Vokal?"
- im Durchschnitt 2.28 Fragen um einen Buchstaben für vereinfachtes Polynesisch zu erraten.

Vereinfachtes Polynesisch

Beispiel: "Vereinfachtes Polynesisch"

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

Möglicher Code

p	t	k	a	i	o
100	00	101	01	110	111

Kodierung von "pato" $= 1000100111$

(Wahrscheinlichere Ereignisse mit kürzerem Code kodiert (damit die Gesamtlänge der Nachricht kürzer ist); kann immernoch eindeutig dekodiert werden, da 2-bit Codes mit 0 anfangen und 3-bit Codes mit 1.)

- Automatische Kodefindung: Huffman Kodierung (Beispiel an Tafel)

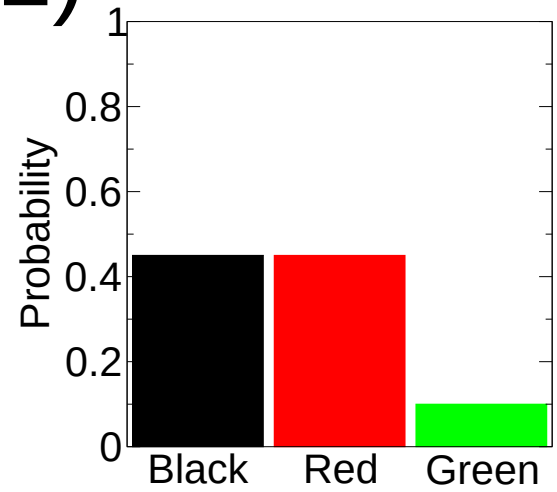
Entropie

Einige Eigenschaften

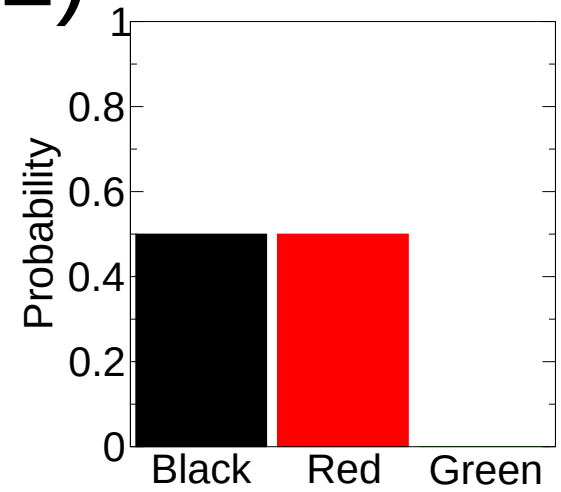
- Bestenfalls hat ein Ereignis die Wahrscheinlichkeit 1
  Dann ist $H = 0$
- Der ungünstigste Fall tritt bei Gleichverteilung ein
  (Wenn die Variable binär ist, ist dann $H = 1$)
- Flache, breite Verteilungen haben eine hohe Entropie
- Spitze, schmale, kompakte Verteilungen haben eine niedrige Entropie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

Quiz: Welche Verteilung hat die höchste Entropie?

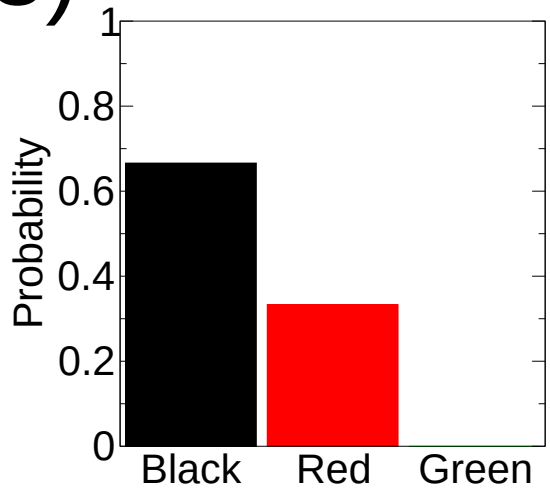
1)



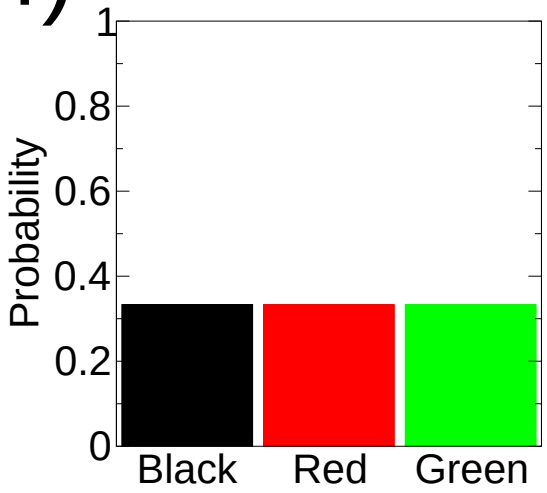
2)



3)



4)



Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

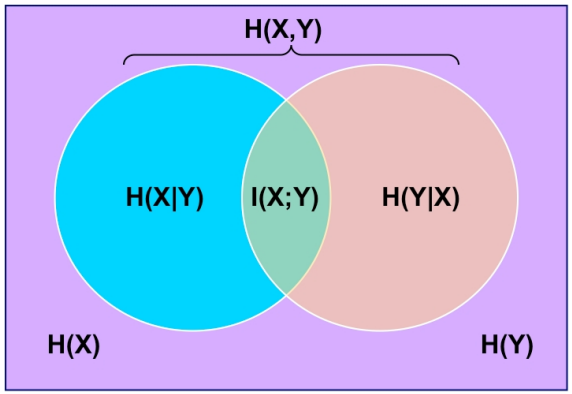
6 Kreuz-Entropie

Gemeinsame Entropie (Joint Entropy)

Gemeinsame Entropie (Joint Entropy)

Wieviel Information benötigt wird, um den Wert zweier Zufallsvariablen anzugeben

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



Wir wollen den kürzesten Code für zwei Ereignisse angeben – wenn diese Ereignisse nicht unabhängig voneinander sind, enthält das eine Ereignis Information über das andere. Dann können wir kompakter kodieren!

Gemeinsame und Bedingte Entropie

Gemeinsame Entropie (Joint Entropy)

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

Bedingte Entropie (Conditional Entropy)

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

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

Gemeinsame Entropie

Vereinfachtes 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	

$$H(C,V) = 4 * \frac{1}{16}\log_2 16 + 2 * \frac{3}{16}\log_2 \frac{16}{3} + \frac{3}{8}\log_2 \frac{8}{3}$$
$$= 2.436$$

pro Silbe

Entropie

Entropie Rate

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

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

Im Polynesischen...

$2.436/2 = 1.218$ bits **pro Buchstabe**
(zu vergleichen mit 2.28 ohne Wissen über Silben)

Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

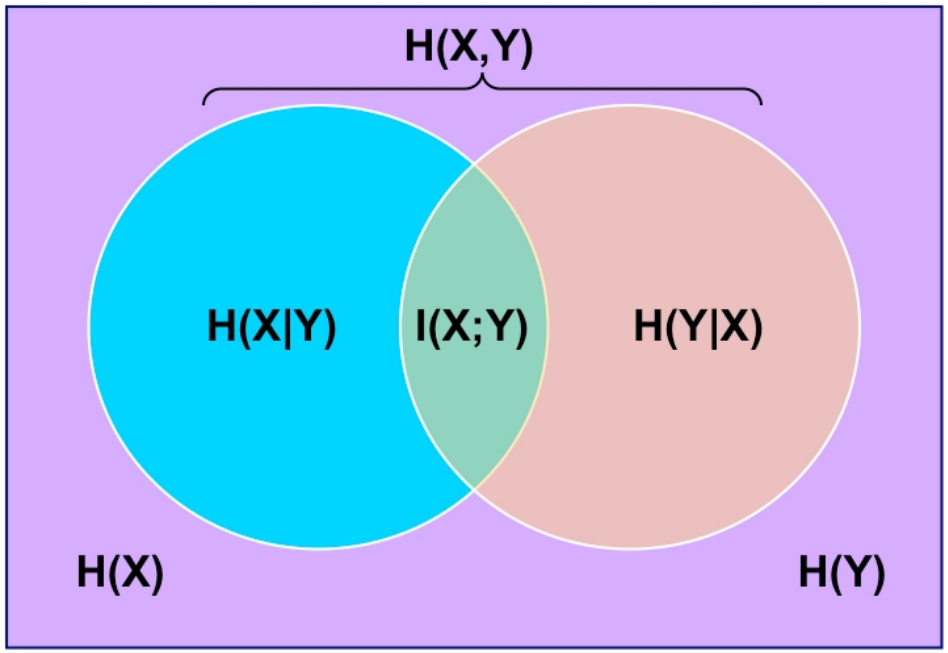
4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

6 Kreuz-Entropie

Mutual Information

- Wenn zwei Variablen nicht unabhängig voneinander sind, lässt sich von der einen Variable mit gewisser Wahrscheinlichkeit auf die andere schliessen.
- Die eine Variable enthält Information über die andere Variable.



Mutual Information

Die Kettenregel

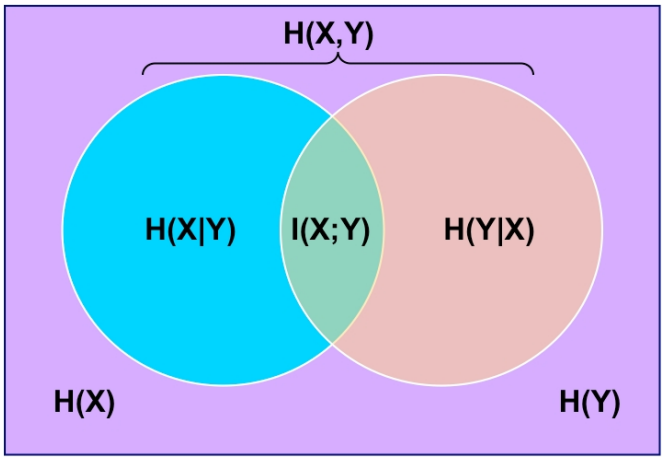
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

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

Mutual Information I(X;Y)

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

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



Mutual Information

MI berechnen

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= \sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_y p(y) \log_2 \frac{1}{p(y)} + \sum_{xy} p(x,y) \log_2 p(x,y) \\
&= \sum_{xy} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

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

Mutual Information

Pointwise Mutual Information

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

Pointwise Mutual Information

$$I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$
$$= \log_2 \frac{p(x|y)}{p(x)}$$
$$= \log_2 \frac{p(y|x)}{p(y)}$$

Mutual Information

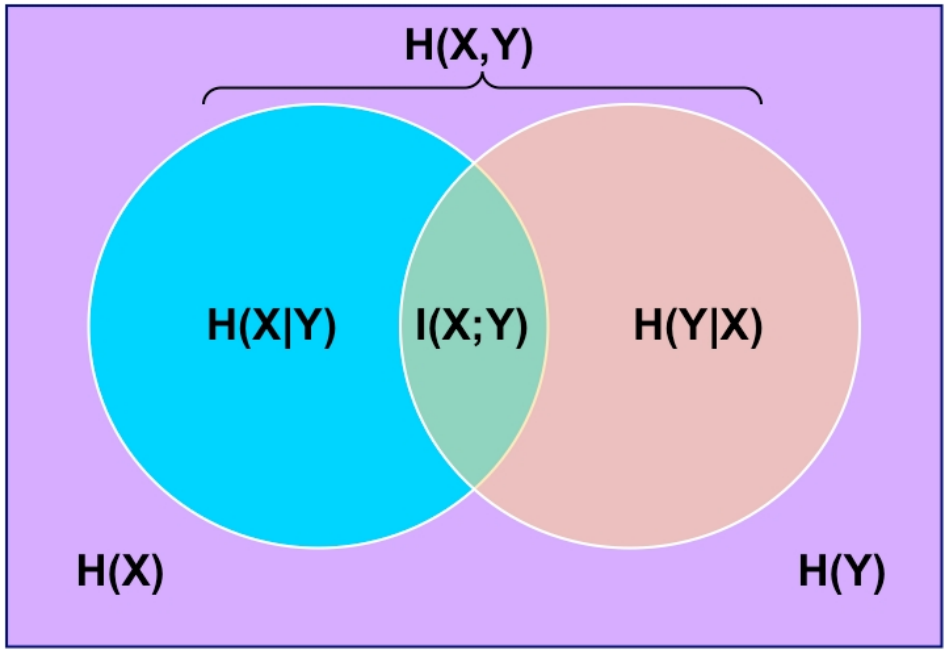
$$I(X;Y) = \sum_{xy} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Verwendung in NLP z.B. für Bedeutungsdisambiguierung von Wörtern, Kollokationsfindung, Clustering.

Zwischenbilanz

Bislang haben wir behandelt:

- Entropie H(X) einer Variablen X
- Joint Entropy H(X;Y)
- Conditional Entropy $H(X|Y)$
- Mutual Information I(X;Y)



Zweiter Teil der Vorlesung:

- The noisy channel model
- Relative Entropie (Kullback-Leibler Divergence)
- Cross Entropy

Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

6 Kreuz-Entropie

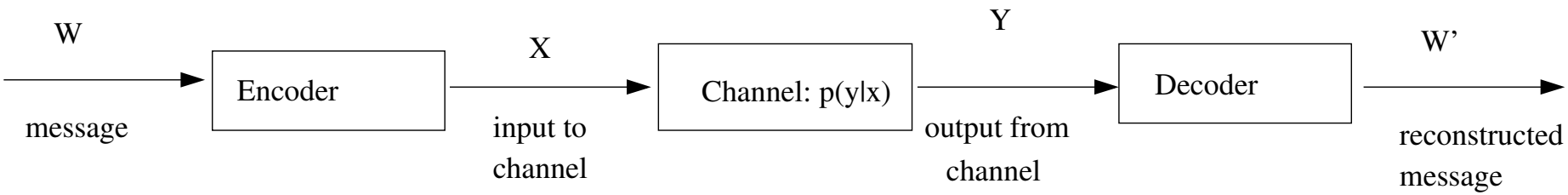
Noisy channel model

- Entropy beschreibt, wie man eine Nachricht effizient kodieren kann.
- Aber wie können wir die Nachricht effizient und korrekt übertragen?

Kompression vs. Redundanz

- Kompression für effiziente Übertragung
- Redundanz um für Verrauschung des Signals bei der Übertragung zu kompensieren.

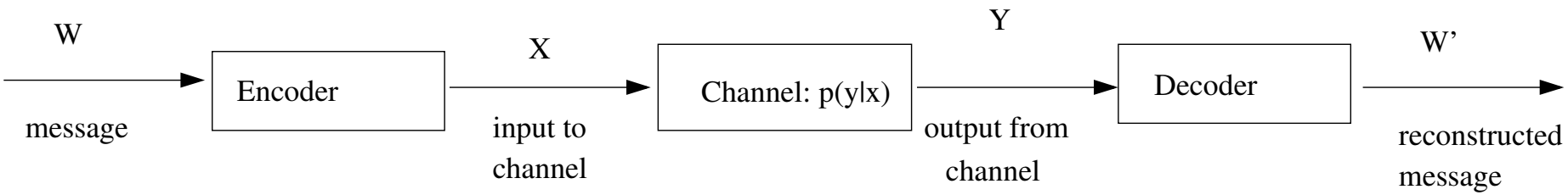
Noisy channel model



Die Übertragung kann durch einen verrauschten Kanal modelliert werden:

- Eine Nachricht W wird (z.B. als binäre Folge von 0en und 1en) X verschlüsselt.
- X wird durch einen verrauschten Kanal übertragen, Resultat ist das verrauschte Signal Y.
- Verrauschtes Signal Y wird entschlüsselt, wir erhalten bestmögliche Schätzung des Ausgangassignals W

Noisy channel model



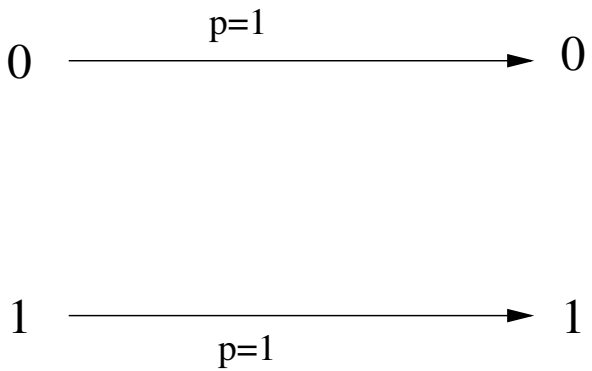
- Die **Kapazität eines Kanals (channel capacity)** ist die Anzahl von Bits die im Durchschnitt übertragen werden können, in Abhängigkeit davon, wie verrauscht der Kanal ist.
- Eingabealphabet X, Ausgabealphabet Y, Wahrscheinlichkeitsverteilung $P(y|x)$ die ausdrückt mit welcher Wahrscheinlichkeit Y ausgegeben wird, wenn X gesendet wurde.
- Kapazität eines Kanals:

$$C = \max_{p(X)} I(X;Y)$$

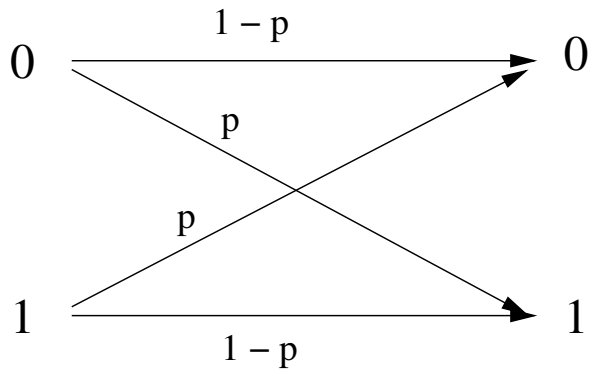
- Die Kapazität eines Kanals ist also die maximale gegenseitige Information (mutual information) von X und Y über alle Eingabeverteilungen p(x)

Noisy channel model

Optimale vs. verrauschte Übertragung



perfect channel



noisy channel which flips 0 to 1 with probability p

Kapazität eines Kanals

$$\begin{aligned} C &= \max_{p(X)} I(X;Y) \\ C &= \max_{p(X)} H(Y) - H(p) \\ &= 1 - H(p) \end{aligned}$$

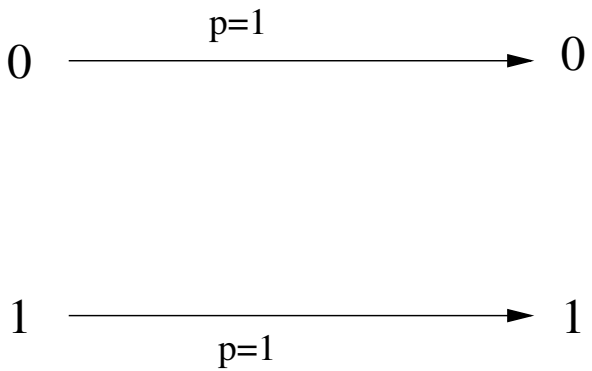
Remember: MI

$$\begin{aligned} &I(X;Y) = \\ &H(Y) - H(Y|X) = \\ &H(Y) - H(p) \end{aligned}$$

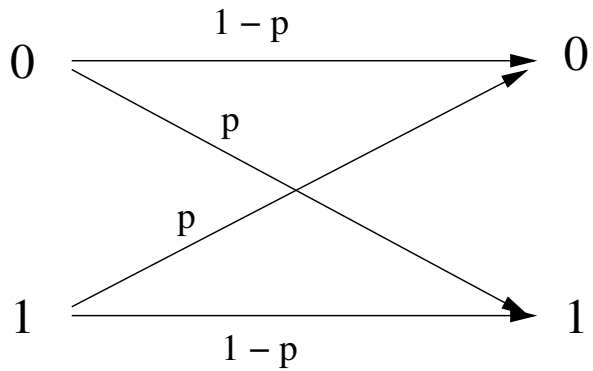
(denn $\max_{p(X)} H(X)$ wenn X uniform verteilt, und wenn X uniform, dann auch Y uniform)

Noisy channel model

Optimale vs. verrauschte Übertragung



perfect channel



noisy channel which flips 0 to 1 with probability p

Kapazität eines Kanals

$$
\begin{aligned}
C &= \max_{p(X)} I(X;Y) \\
C &= \max_{p(X)} H(Y) - H(p) \\
&= 1 - H(p)
\end{aligned}
$$

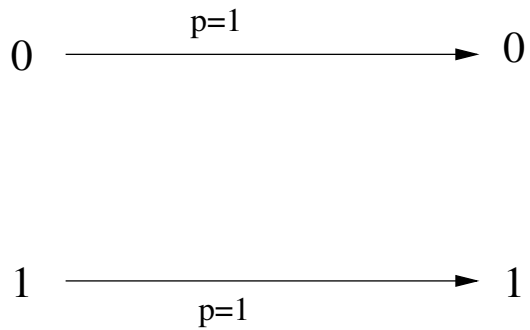
Remember: MI

$$
\begin{aligned}
&I(X;Y) = \\
&H(Y) - H(Y|X) = \\
&H(Y) - H(p)
\end{aligned}
$$

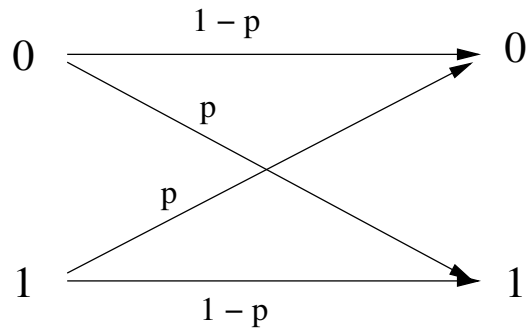
(denn $\max_{p(X)} H(X)$ wenn X uniform verteilt, und wenn X uniform, dann auch Y uniform)

Noisy channel model

Optimale vs. verrauschte Übertragung



perfect channel



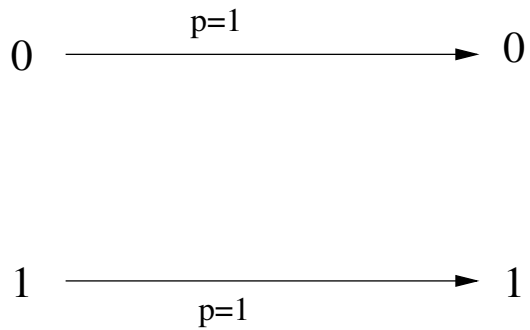
noisy channel which flips 0 to 1 with probability p

Kapazität $C = 1 - H(p)$

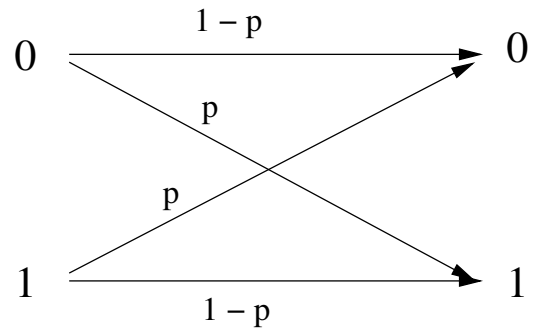
- perfekte Übertragung (p=0): C = 1 bit
- immer garantiert falsch (p=1): C = 1 bit
- schlimmste Störung: 50% Chance dass 0 als 1 oder 1 als 0 übertragen wird: C = 0 bits
- Kanal mit C = 0 bits kann keine Information übertragen. Eingabe und Ausgabe unabhängig voneinander.

Noisy channel model

Optimale vs. verrauschte Übertragung



perfect channel

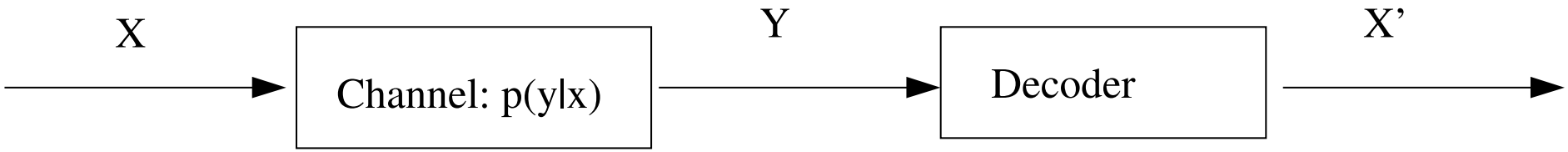


noisy channel which flips 0 to 1 with probability p

Kapazität $C = 1 - H(p)$

- perfekte Übertragung (p=0): C = 1 bit
- immer garantiert falsch (p=1): C = 1 bit
- schlimmste Störung: 50% Chance dass 0 als 1 oder 1 als 0 übertragen wird: C = 0 bits
- Kanal mit C = 0 bits kann keine Information übertragen. Eingabe und Ausgabe unabhängig voneinander.

Das noisy channel model für NLP



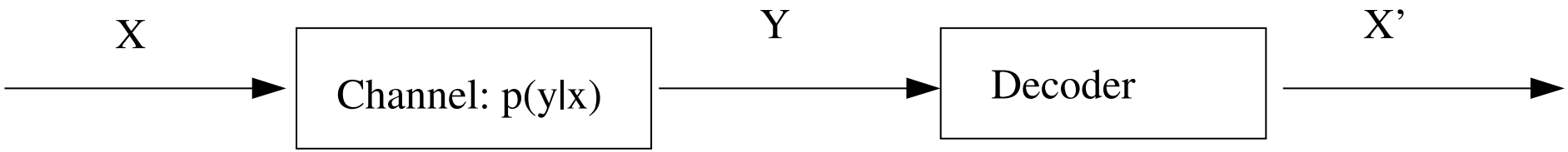
Anwendungen

Viele natürlichsprachliche Anwendungen können an Hand des Kanalmodells modelliert werden:

- Allerdings modellieren wir nicht die Verschlüsselung, sondern nur das Rauschen.

Anwendung	Eingabe	Ausgabe
Maschinelle Übersetzung	L1 Wortfolge	L2 Wortfolge
Rechtschreibkorrektur	Originaler Text	Text mit Fehlern
PoS Tagging	PoS Tags	Wörter
Spracherkennung	Wortfolge	Audiosignal

Das noisy channel model



Anwendungen, Teil 2

Es wird ein Modell gebaut, an Hand dessen man von der (bekannten) Ausgabe auf die (unbekannte) Eingabe schließen kann

$$X' = \operatorname*{argmax}_x p(x|y) = \operatorname*{argmax}_x \frac{p(x)p(y|x)}{p(y)} = \operatorname*{argmax}_x p(x)p(y|x)$$

Zwei Wahrscheinlichkeitsverteilungen müssen geschätzt werden:

- Das Sprachmodell:
  Verteilung der Zeichen in der Eingabesprache $p(x)$
- Das Kanalmodell $p(y|x)$

Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

6 Kreuz-Entropie

Relative Entropie

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

Relative Entropy

- Die Relative Entropie, auch *Kullback-Leibler (KL) Divergenz* gennant, 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!

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

Relative Entropie

Relation von **Relativer Entropie (Kullback-Leibler Divergenz)** und **Mutual Information**

- Mutual information misst in wie fern sich eine gemeinsame Verteilung von einer unabhängigen Verteilung unterscheiden.

$$I(X;Y) = D(p(x,y)||p(x)p(y))$$

Inhaltsverzeichnis

1 Entropie eines einzelnen Ereignisses

2 Entropie des Auftretens mehrerer Ereignisse (Joint Entropy)

3 Gegenseitige Information (Mutual Information)

4 Noisy Channel Model

5 Relative Entropie (Kullback-Leibler Divergenz)

6 Kreuz-Entropie

Kreuz-Entropie

Kreuz-Entropie

- Wenn wir ein Modell m eines Phänomens p (z.B. "englische Sprache") 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

Entropie

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 *(Shannon guessing game)*
	1.34	Cover & Thomas (using gambling)

Bessere Modelle haben weniger Entropie, da sie mehr über die Sprache wissen, sind sie weniger überrascht über den Input.

Perplexität ist das gleiche wie Kreuzentropie, nur anders notiert:

$$perplexity(x_{1n}, m) = 2^{H(x_{1n}, m)}$$

Zusammenfassung

- Entropie
  - Misst die durchschnittliche Unsicherheit einer Zufallsvariable
  - Mehr Information = weniger Unsicherheit / Entropie
  - Repräsentiert als Anzahl von Bits, die durchschnittlich benötigt werden um eine Nachricht zu übertragen.
- Mutual Information
  - used in NLP to calculate similarity between e.g. words
- Noisy channel model
  - Definiert die maximale Kapazität eines Kanals
  - Kann auch in NLP verwendet werden: finde die Eingabe für eine beobachtete Ausgabe
- Entropie and Sprachmodellierung
  - Modelle mit kleinerer Entropie sind besser
  - Relative Entropy: $D(p||q)$, die Distanz zwischen zwei Wahrscheinlichkeitsverteilungen
  - Cross entropy $H(X, m) = H(X) + D(p||m)$
  - Aufgabe: Model m mit kleinster Kreuzentropie finden