Statistische Methoden in der Sprachverarbeitung

Gerhard Jäger

7. Mai 2002

Informationstheorie

Der Entropiebegriff

- *Entropie:* „Chaos“, Unordung, Nicht-Vorhersagbarkeit, ...
- Begriff kommt ursprünglich aus der Physik:
  - Entropie wird nur geringer wenn Energie zugeführt wird
- Maß der *Ungewissheit*:
  - geringe Entropie bedeutet geringe Unsicherheit und umgekehrt
  - hohe Entropie bedeutet hohe „Überraschung“
  - je überraschender das Ergebnis eines Experimentes ist, desto überraschter sind wir

Die Formel

- Sei $p(X)$ die Verteilung der Zufallsvariablen X
- Entropie von X („$H(X)$“): Erwartungswert des negativen Logarithmus der Wahrscheinlichkeit

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

- Einheit: *Bit*
- Konventionen:
  - Wenn nicht explizit anders angegeben, benutzen wir den binären Logarithmus ($\log x \doteq \log_2 x$)
  - $0 \log 0 = 0$
- Notationsvarianten: $H(X) = H(p) = H_X(p) = H(p_X)$

Beispiele

- Bernoulli-Verteilung, basiert auf fairer Münze
  - $p(0) = p(1) = 0.5$
  - $\mathbf{H(p)} = -(0.5\log(0.5) + 0.5\log(0.5)) = -2 \cdot (0.5 \cdot -1) = \mathbf{1}$
- Augenzahl beim fairen Würfel
  - $\mathbf{H(p)} = -(6 \cdot (\frac{1}{6}\log\frac{1}{6})) = \log 6 \approx 2.584962$
- Unfaire Münze:
  - $p(1) = 0.2 \ldots H(p) = 0.722; p(1) = 0.01 \ldots h(p) = 0.081$

Extreme

- Wenn für ein $x : p(x) = 1$, dann $H(p) = 0$
- Keine Überraschung $\Rightarrow$ keine Information
- keine oberste Schranke, aber $H(X) \leq \log|\Omega|$
  - Wenn alle Elementarereignisse gleich wahrscheinlich sind, ist die Vorhersagbarkeit am geringsten

Kodierungs-Interpretation der Entropie

- minimale durchschnittliche Zahl von Bits, die nötig sind, um eine Botschaft (Zeichenkette, Sequenz, Serie, ...) zu kodieren (wobei jedes Element Resultat eines Zufallsprozesses mit Verteilung p ist) $:= H(p)$
- **Beispiel:**
  - 4 Symbole

    $$A\ C\ G\ T$$

  - Wahrscheinlichkeiten

    $$P(A) = 0.5,\ P(C) = 0.25,\ P(G) = 0.125,\ P(T) = 0.125$$

  - Negative binäre Logarithmen:

    A: 1 Bit, C: 2 Bit, G: 3 Bit, T: 3 Bit

- Entropie

$$\begin{aligned} H &= \sum p(x) - \log p(x) \\ &= 0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3 \\ &= 0.5 + 0.5 + 0.375 + 0.375 \\ &= 1.75 \end{aligned}$$

- Mögliche Kodierung in binärem Alphabet:

A: 00; C: 01; G: 10; T: 11

- gewichtete durchschnittliche Kode-Länge: 2 ($> H$)
- Besserer Kode:

A: 1; C: 01; G: 000; T: 001

- gewichtete durchschnittliche Kode-Länge:

$$0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.25 \cdot 3 = 1.75$$

- Optimaler Kode: Kodelänge für Symbol $S = -log(P(S))$

Perplexität: Motivation

- Gleichverteilung:
  - bei 2 gleichwahrscheinlichen Ergebnissen: $H(p) = 1$
  - bei 32 gleichwahrscheinlichen Ergebnissen: $H(p) = 5$
  - bei 4.3 Milliarden gleichwahrscheinlichen Ergebnissen: $H(p) \approx 32$
- Wenn Ergebnissen nicht gleichwahrscheinlich sind:
  - 32 Elementarereignisse, 2 davon gleichwahrscheinlich, Rest unmöglich:
    - $H(p) = 1$
  - Gibt es eine Möglichkeit, den Grad der Unsicherheit/Informationsgehalt für Zufallsvariable zu vergleichen, wenn die zugrundeliegenden Ereignisräume unterschiedlich groß sind?

Perplexität

- Perplexität
  - $G(p) = 2^{H(p)}$

- D.h. wir sind wieder bei 32 (für 32 gleichwahrscheinliche Ergebnisse), 2 für Münzwurf usw.
- intuitiv leichter vorstellbar:
  - $G(p) = n$ bedeutet, dass p so gut vorhersagbar ist wie ein Experiment mit n gleichwahrscheinlichen Ausgängen
- Je mehr eine Verteilung von der Gleichverteilung abweicht, desto geringer sind Entropie und Perplexität

Gemeinsame und bedingte Entropie

- Zwei Zufallsvariable: X und Y
- Gemeinsame Entropie:
  - (Paare von X- und Y-Werten werden als ein Ereignis betrachtet)

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

- Bedingte Entropie:

$$H(Y|X) = -\sum_x \sum_y p(x,y) \cdot \log p(y|x)$$

- Alternative Definition:

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ & \sum_x p(x)(-\sum_y p(y|x) \log p(y|x)) \\ & -\sum_x \sum_y p(y|x)p(x) \log p(y|x) \\ & -\sum_x \sum_y p(x,y) \log p(y|x) \end{aligned}$$

Eigenschaften der Entropie

- Entropie ist nicht-negative
  - $H(x) \geq 0$
  - Beweis:
    - $\log x \leq 0$ für $0 < x \leq 1$
    - $p(x)$ ist nicht-negativ, also ist $p(x) \log p(x)$ nicht-positiv
    - Summe nicht-positiver Werte ist nicht-positiv

    - Negation eines nicht-positiven Wertes ist nicht-positiv

- Kettenregel (analog zu Wahrscheinlichkeiten; statt Multiplikation aber Addition)
  - $H(X,Y) = H(X|Y) + H(Y)$, und
  - $H(X,Y) = H(Y|X) + H(X)$, da $H(X,Y) = H(Y,X)$
- Bedingte Entropie ist besser
$$H(Y|X) \leq H(Y)$$
- $H(X,Y) \leq H(X) + H(Y)$ (Gleichheit wenn X und Y stochastisch unabhängig sind)

Relative Entropie und gemeinsame Information

- Effiziente Kodierung eines stocahstischen Prozesses setzt Kenntnis seiner Wahrscheinlichkeitsverteilung voraus
- Annahme falscher Verteilung führt zu Verschwendung von Bits
- *Relative Entropie* zweier Verteilungen p und q misst die minimale durchschnittliche Anzahl von verschwendeten Bits, wenn man einen Prozess mit Verteilung p auf der Basis von q kodiert

  Definition 1 (Relative Entropie) *Die* relative Entropie *(oder der* Kullback-Leibler-Abstand*) zwischen zwei Wahrscheinlichkeitsfunktionen p und q ist definiert als*

$$\begin{aligned} D(p\|q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= E_p \log \frac{p(x)}{q(x)} \end{aligned}$$

- Relative Entropie ist immer nicht-negativ
- $D(p\|q) = 0$ gdw. $p = q$
- Kann als „Abstand“ zwischen zwei Verteilungen aufgefasst werden
- Aber Vorsicht: Die relative Entropie (im Unterschied zum Abstand im geometrischen Sinn)
  - ist nicht immer symmetrisch
  - erfüllt die Dreiecksungleichung nicht

Gemeinsame Information

- X und Y sind stochastisch abhängig $\Rightarrow$ X enthält Information über Y (und umgekehrt)
- diese Information kann bei simultaner Kodierung von X und Y genutzt werden
- bei fälschlicher Annahme von stochastischer Unabhängigkeit werden Bits verschwendet
- *Gemeinsame Information* von X und Y ist demnach die relative Entropie zwischen der tatsächlichen gemeinsamen Verteilung von X und Y und ihrer gemeinsamen Verteilung unter Annahme ihrer stochastischen Unabhängigkeit

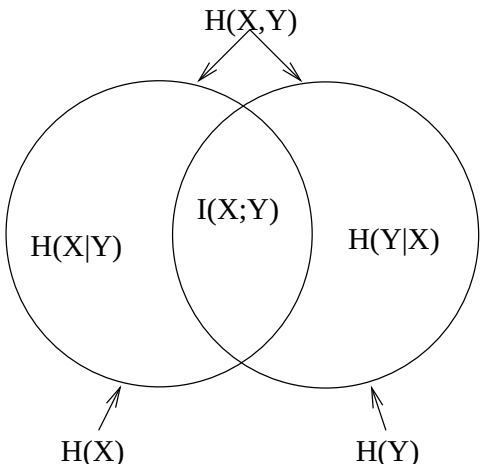
Definition 2 (Gemeinsame Information)

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)\|p(x)p(y)) \\ &= E_{p(x,y)} \log \frac{p(X,Y)}{p(Y)p(Y)} \end{aligned}$$

Verhältnis von Entropie und gemeinsamer Information

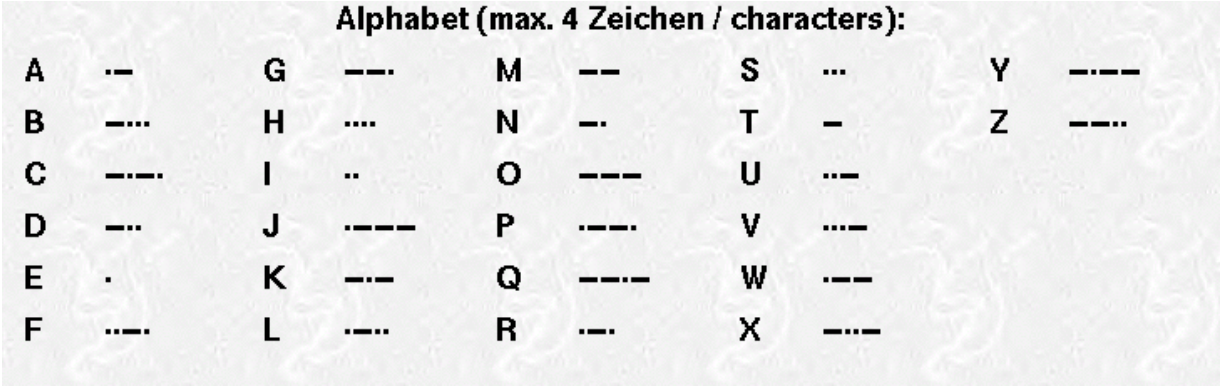
Theorem 1

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



Hausaufgaben

Das Morsealphabet kodiert Buchstaben des lateinischen Alphabets im Binärkode (der Einfachheit halber ignorieren wir das Leerzeichen zwischen den kodierten Buchstaben).



In deutschen Texten sind die einzelnen Buchstaben folgendermaßen verteilt:

1.	E	17.40%	14.	M	2.53%
2.	N	9.78%	15.	O	2.51%
3.	I	7.55%	16.	B	1.89%
4.	S	7.27%	17.	W	1.89%
5.	R	7.00%	18.	F	1.66%
6.	A	6.51%	19.	K	1.21%
7.	T	6.15%	20.	Z	1.13%
8.	D	5.08%	21.	P	0.79%
9.	H	4.76%	22.	V	0.67%
10.	U	4.35%	23.	J	0.27%
11.	L	3.44%	24.	Y	0.04%
12.	C	3.06%	25.	X	0.03%
13.	G	3.01%	26.	Q	0.02%

1. Was ist die Entropie eines Buchstabens in einem deutschen Text, wenn man Groß-/Kleinschreibung, Interpunktion und Leerzeichen ignoriert und die Wahrscheinlichkeit eines Buchstaben mit seiner relativen Häufigkeit gleichsetzt?

2. Was ist die erwartete Länge des Kodes eines deutschen Buchstaben im Morsealphabet?

3. Was ist die erwartete Länge des Kodes eines deutschen Buchstaben im ASCII-Code?

4. (freiwillig) Benutzen Sie ein Komprimierungsprogramm wie „gzip“ und einen selbstgewählten deutschen Text[1], um eine obere Schranke für die tatsächliche Entropie eines Buchstaben im Deutschen zu ermitteln.

[1] Zum Beispiel das Negra-Corpus; zugänglich von /home/jaeger/NEGRA/negra-corpus.sent auf dem Institutsserver