Informationstheorie und Entropie

Katja Markert

Institut für Computerlinguistik
Uni Heidelberg
markert@cl.uni-heidelberg.de

November 13, 2019

Bisher und heute

1. Bisher: Wahrscheinlichkeiten und Zufallsvariablen
2. Jetzt: Informationstheorie und Entropie
3. Jetzt: Informationstheoretische Assoziationsmaße
4. Wofür braucht man Informationstheorie in NLP: Gütemaß für ML-Modelle sowie n-gram Modelle, Assoziationsmaße für Merkmalsselektion in ML

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Motivation

Gehen von einem Wahrscheinlichkeitsraum aus!

- Wieviel durchschnittliche Information steckt in einer Zufallsvariable? Wieviel Unsicherheit steckt in einer Zufallsvariable? $\longrightarrow$ Entropie
- Wie überrascht bin ich von einem Ereignis? $\longrightarrow$ Entropie
- Wie stark wird mein Wissen über eine Variable Y durch das Wissen über eine andere Variable X beeinflusst? $\longrightarrow$ Joint Entropy, Condiitional Entropy, Mutual Information

Wir brauchen Maße für **Information** und **Überraschung**

Information

Gegeben ein diskreter Wahrscheinlichkeitsraum (diskrete Ergebnismenge Ω mit Wahrscheinlichkeitsmaß). X sei diskrete Zufallsvariable mit Wahrscheinlichkeitsverteilung p.

In NLP/Nachrichtentheorie/Codierungstheorie ist X eine Quelle von "Nachrichten"

Wieviel **Information** $I(x)$ beinhaltet eine Nachricht x?

Gesucht Funktion $I : X \to \mathbb{R}_0^+$

Information

Annahme: Information eines Ereignisses hängt nur von seiner Wahrscheinlichkeit ab.

Damit gesucht $I : [0,1] \to \mathbb{R}_0^+$ mit den folgenden Eigenschaften

- Je unwahrscheinlicher eine Nachricht ist, desto größer ist deren Information. Also: Wenn $p(x_1) < p(x_2)$, dann $I(x_1) > I(x_2)$
- $I(1) = 0$
- $p(x)$ ähnlich zu $p(y) \Longrightarrow I(x)$ ähnlich zu $I(y)$: stetige Funktion
- Gegeben seien zwei voneinander unabhängige Nachrichten x_1, x_2. Dann sollte gelten $I(x_1, x_2) = I(x_1) + I(x_2)$
- $I(0.5) = 1$

Einzige Möglichkeit (Beweis in Ross (2009))

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

Alle Logarithmen haben Basis 2 in dieser Vorlesung.

Entropie

Entropie misst die **durchschnittliche/erwartete** Menge an Information in einer Zufallsvariable.

$$H(X) := \sum_{x \in \Omega_X} p(x) I(x)$$

$$H(X) := -\sum_{x \in \Omega_X} p(x) \log_2 p(x)$$

$$H(X) := \sum_{x \in \Omega_X} p(x) \log_2 \frac{1}{p(x)}$$

- $\Omega_X = \Omega(X)$ ist Menge der Werte, die die Zufallsvariable annehmen kann
- Gemessen in Bits.
- $0 \log 0 := 0$
- Notation: $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

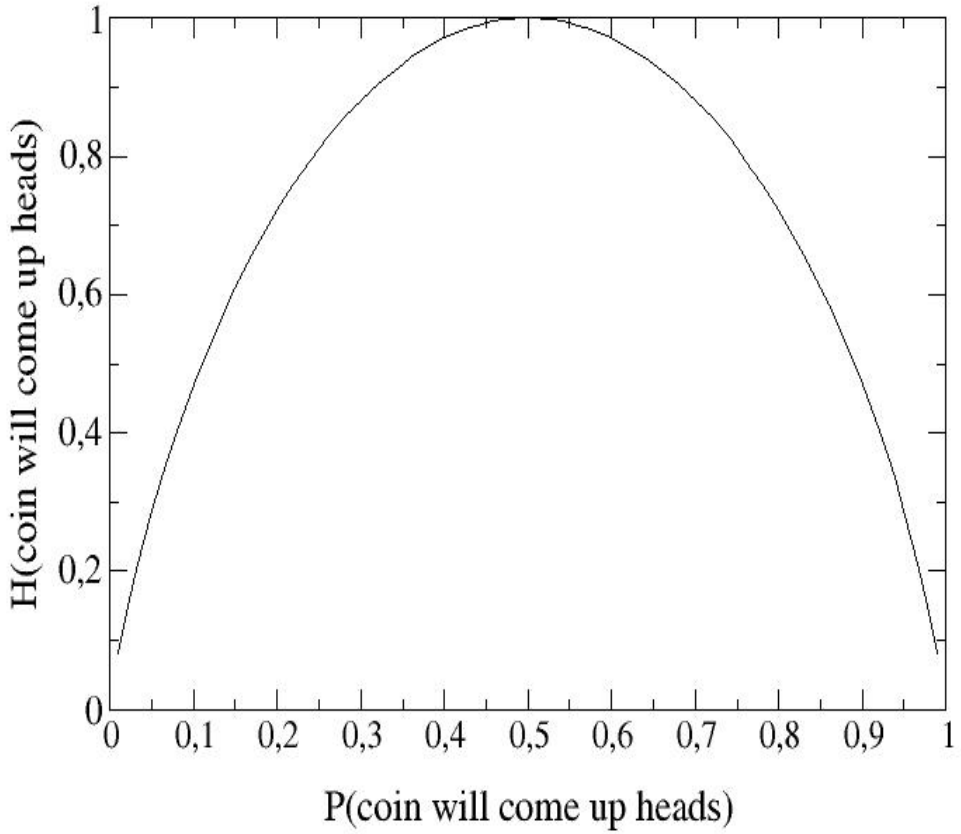
Beispiel Münzwurf

Nehmen wir an, unsere “Nachricht” ist die Vermittlung des Ergebnis eines einzigen Münzwurfs mit gewichteter Münze. Was ist die Entropie?

- Münze 1: $p(K) = p(Z) = 0.5 \Longrightarrow$
  $H(X) = -0.5 \cdot \log 0.5 + (-0.5 \log 0.5) = 2 \cdot 0.5 = 1$
- Münze 2: $p(K) = 0; p(Z) = 1 \Longrightarrow H(X) = 0$
- Münze 3: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

Eigenschaften der Entropie

Kurve für binäre Zufallsvariable mit Wahrscheinlichkeiten p und $1 - p$ (Münzwurf):



Beispiel: n-seitiger Würfel

Entropie eines n-seitigen fairen Würfels. X Ergebnis eines einzigen Wurfes.

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p(i) \log p(i) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\ &= -\log \frac{1}{n} \\ &= \log n \end{aligned}$$

Beispiel: 8-seitiger Würfel Entropie 3 Bits.

Eigenschaften der Entropie

- $H(X) \geq 0$ (Erinnerung: $H(X) = -\sum_{x \in \Omega_X} p(x) \log_2 p(x)$)
- Bei Ergebnismenge Ω mit Größe n und p Gleichverteilung: $H(p) = \log n$
- Bei Ergebnismenge Ω mit n Werten ungleich Null: $H(p) \leq \log n$
- $H(X) = 0$, dann und nur dann wenn ein $x \in \Omega$ Wahrscheinlichkeit 1 hat (und die anderen damit notwendigerweise alle die Wahrscheinlichkeit Null).

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Joint und Conditional Entropy

Joint Entropy: durchschnittliche Information eines Paares von diskreten Zufallsvariablen

$$H(X,Y) := -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\log p(x,y)$$

Conditional Entropy: Wieviel Extrainformation bekommt man von Y, wenn man X schon kennt?

$$H(Y|X) := -\sum_{x\in\Omega_X} p(x)\sum_{y\in\Omega_Y} p(y|x)\log p(y|x) = -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\log p(y|x)$$

Beispiel: Simplified Polynesian

Berechne die Buchstabenentropie der folgenden Sprache mit nur 6 Buchstaben:

p	t	k	a	i	u
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

$$
\begin{aligned}
H(P) &= -\sum_{i \in \{p,t,k,a,i,u\}} p(i) \log p(i) \\
&= -[4\tfrac{1}{8} \log \tfrac{1}{8} + 2\tfrac{1}{4} \log \tfrac{1}{4}] \\
&= 2\tfrac{1}{2} \text{ bits}
\end{aligned}
$$

Simplified Polynesian ctd.

Erster Buchstabe in Spalte:

	p	t	k	
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
u	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	

$$H(V|C) = -\sum_{c \in C} \sum_{v \in V} p(c,v) \log p(v|c)$$

Simplified Polynesian ctd.

$$
\begin{aligned}
H(V|C) &= -\sum_{c\in C}\sum_{v\in V} p(c,v)\log p(v|c) \\
&= -(p(a,p)\log p(a|p)+p(a,t)\log p(a|t)+p(a,k)\log p(a|k)+ \\
&\quad p(i,p)\log p(i|p)+p(i,t)\log p(i|t)+p(i,k)\log p(i|k)+ \\
&\quad p(u,p)\log p(u|p)+p(u,t)\log p(u|t)+p(u,k)\log p(u|k))
\end{aligned}
$$

$$
\begin{aligned}
H(V|C) &= -\sum_{v\in C}\sum_{v\in V} p(c,v)\log p(v|c) \\
&= -(p(a,p)\log p(a|p)+p(a,t)\log p(a|t)+p(a,k)\log p(a|k)+ \\
&\quad p(i,p)\log p(i|p)+p(i,t)\log p(i|t)+p(i,k)\log p(i|k)+ \\
&\quad p(u,p)\log p(u|p)+p(u,t)\log p(u|t)+p(u,k)\log p(u|k)) \\
&= -(\tfrac{1}{16}\log\tfrac{\frac{1}{16}}{\frac{1}{8}}+\tfrac{3}{8}\log\tfrac{\frac{3}{8}}{\frac{3}{4}}+\tfrac{1}{16}\log\tfrac{\frac{1}{16}}{\frac{1}{8}}+ \\
&\quad \tfrac{1}{16}\log\tfrac{\frac{1}{16}}{\frac{1}{8}}+\tfrac{3}{16}\log\tfrac{\frac{3}{16}}{\frac{3}{4}}+0+ \\
&\quad 0+\tfrac{3}{16}\log\tfrac{\frac{3}{16}}{\frac{3}{4}}+\tfrac{1}{16}\log\tfrac{\frac{1}{16}}{\frac{1}{8}}) \\
&= \tfrac{11}{8} = 1.375 \text{ bits}
\end{aligned}
$$

Kettenregel für Entropie

Erinnerung an Kettenregel für Ereignisse in Wahrscheinlichkeitstheorie:

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\ldots P(A_n|A_1 \ldots A_{n-1})$$

Kettenregel für Entropie:

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

Symmetrie

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

Verallgemeinert

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

Beweis

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

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Entropie von Sprachen

$$H(L) = \lim_{n \to \infty} \frac{1}{n} H_0(X_1, X_2 \ldots X_n)$$

- $H(X_i)$ sind eine Serie von Buchstaben.
- $H(X_i)$ sind eine Serie von Wörtern.

Entropie des Englischen (Shannon)

Claude Shannon ließ Menschen den nächsten Buchstaben in einem Text erraten. Er benutzte die bedingten Wahrscheinlichkeiten, um die (per-Buchstaben)Entropie des Englischen zu bestimmen.

# Fragen	1	2	3	4	5	> 5
Wahrscheinlichkeit	.79	.08	.03	.02	.02	.05

- Resultat Menschen $H(\mathrm{English})$ zwischen 1.25 und 1.35.

Entropie des Englischen (Shannon)

Lückentext: (englisches Alphabet plus “space”)

_____

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Mutual information (MI)

Es seien X und Y diskrete Zufallsvariablen mit $p(X, Y)$ gemeinsamer Wahrscheinlichkeitsverteilung , und $p(X)$ und $p(Y)$ die Randverteilungen von X und Y. Dann ist die **Mutual Information** zwischen X und Y definiert als:

$$I(X;Y) := \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information is the amount of information that one random variable contains about another random variable.

Mutual Information

Es gilt für zwei diskrete Zufallsvariablen X und Y

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

Beweis

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x|y)}{p(x)} \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x|y) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Mutual Information

- $I(X;Y) \geq 0$
- $I(X;Y) = I(Y;X)$;
- $I(X;Y)$ ist ein Maß für **Abhängigkeit** zwischen X und Y:
  - $I(X;Y) = 0$ genau dann wenn X und Y unabhängig sind;
  - $I(X;Y)$ wächst aber nicht nur mit Abhängigkeit von X und Y, sondern auch mit $H(X)$ und $H(Y)$;

Mutual Information: Beispiel

Wieder simplified Polynesian

$p(x,y)$	p	t	k	$p(y)$
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
u	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$p(x)$	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	

Was ist die Mutual Information zwischen der Konsonantenvariable und der Vokalvariable:

$$I(V;C) = H(V) - H(V|C)$$

Mutual information

Berechne Entropie der Vokalvariable:

$$\begin{aligned} H(V) &= -\sum_{y \in V} p(y) \log p(y) \\ &= -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4}) \\ &= 1.5 \text{ bits} \end{aligned}$$

Wir haben vorher schon berechnet $H(V|C) = 1.375$ bits, also folgt

$$I(V;C) = H(V) - H(V|C) = 0.125 \text{ bits}$$

Pointwise Mutual Information

Es seien X and Y diskrete Zufallsvariablen mit gemeinsamer Verteilung $p(X, Y)$ und Randverteilungen $p(X)$ and $p(Y)$. Dann ist die pointwise mutual information bei x, y definiert als:

$$pmi(x; y) := \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

Pointwise Mutual Information für Kollokationen

Kollokationen: Worte, die gewohnheitsmäßig, zuammen auftreten.
Meist nicht ersetzbar oder modizierbar (*strong tea* vs. *powerful tea*)
Table 5. 14 aus Manning und Schütze:

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

Beispiel:

$$pmi(Ayatollah, Ruhollah) = \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \cdot \frac{20}{14307668}}$$

Eigenschaften von PMI

- Perfekte Unabhängigkeit: $pmi(x,y) = \log \frac{p(x,y)}{p(x)p(y)} = \log 1 = 0$
- pmi gutes Maß für Unabhängigkeit!
- Was passiert allerdings, wenn wir x nur einmal gesehen haben und das genau mit y?
  Dann gilt:

$$pmi(x,y) = \log \frac{\frac{1}{n}}{\frac{1}{n} \cdot p(y)} = \log \frac{1}{p(y)} = -\log p(y)$$

  Dies heisst, dass hier die PMI überschätzt wird.

Übersicht

1. Information und Entropie
2. Joint und Conditional Entropy
3. Entropie von Sprachen
4. Mutual Information und Pointwise Mutual Information
5. Zusammenfassung und Ausblick

Zusammenfassung

- Information hängt nur von Wahrscheinlichkeit ab
- Entropie: Erwartungswert der Information einer Variable, auch Maß der Unsicherheit in einer Variable
- Conditional Entropy: Extrainformation in einer Variable, wenn man eine andere schon kennt
- Conditional Entropy: Kettenregel, dropping condition
- Entropie von Sprachen
- Pointwise Mutual Information als Assoziationsmaß zwischen zwei Wörtern
- Probleme PMI: Data Sparseness

Grundlegende Literatur

- Manning and Schuetze (1999). *Introduction to Statistical Natural Language Processing.* Kapitel 2.2 sowie 5.4
- S. Ross (2009). *A first course in probability*. Pearson Prentice Hall.