Assoziationsmaße II

Katja Markert

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

July 22, 2019

Bisher und heute

1. Kollokationen
2. Assoziationmaße für Kollokationen: Kookkurrenz, Mean and Variance, Signifikanztests
3. Jetzt: Informationstheoretische Assoziationsmaße (Hintergrund Informationstheorie)
4. Wofür braucht man Informationstheorie in NLP: Kollokationen sowie embeddings, Gütemaß für ML-Modelle sowie n-gram Modelle

Übersicht

Übersicht

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.

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: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

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: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

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: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

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: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

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: $p(K) = 3/4; p(Z) = 1/4 \Longrightarrow H(X) = 0.811$

Beispiel Münzwurf

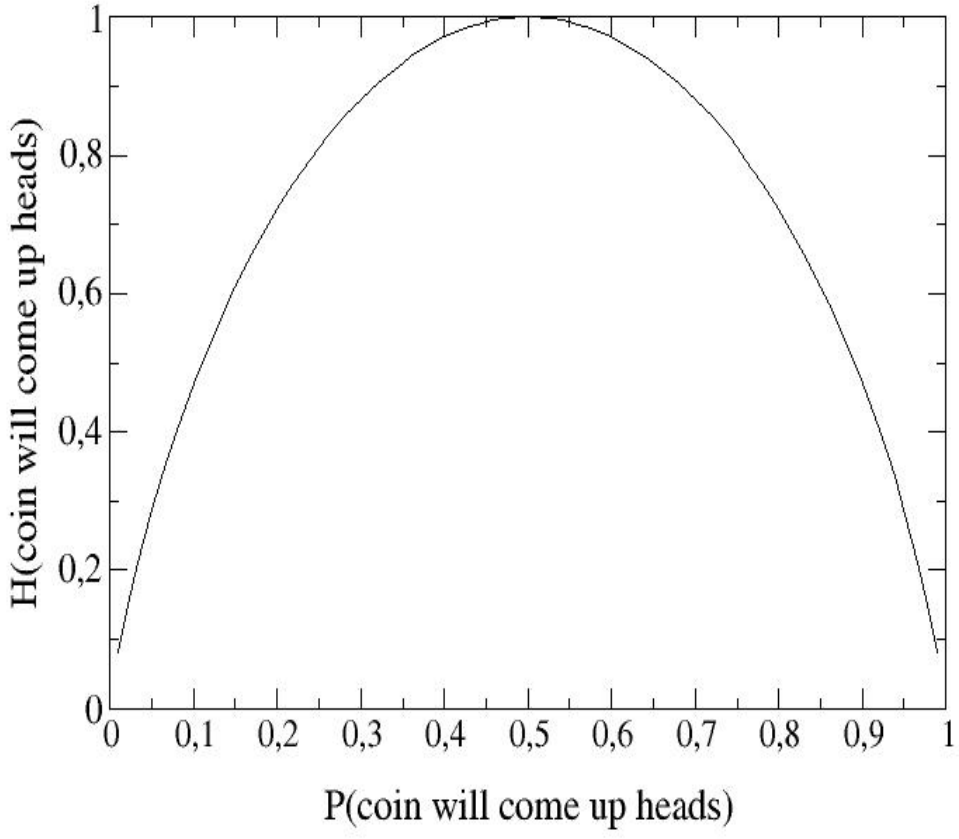
Nehmen wir an, unsere “Nachricht” ist die Vermittlung des Ergebnis eines einzigen Münzwurfs mit gewichteter Münze. Was ist die Entropie?

- 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

H(X) ist konkave Kurve.

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

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
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

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

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.

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

Simplified Polynesian ctd.

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

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

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

Dropping Conditioning: Conditioning reduces entropy

Es gilt:

$$H(X|Y) \leq H(X)$$

Beweis (braucht Jensen Ungleichung):

$$\begin{aligned}
H(X|Y) &= \sum_{y\in\Omega_y} p(y) \sum_{x\in\Omega_x} -p(x|y)\log p(x|y) \\
&= \sum_{x\in\Omega_x} p(x) \sum_{y\in\Omega_y} -p(y|x)\log p(x|y) \\
&\leq \sum_{x\in\Omega_x} p(x)\log \sum_{y\in\Omega_y} \frac{p(y|x)}{p(x|y)} \\
&= \sum_{x\in\Omega_x} p(x)\log \sum_{y\in\Omega_y} \frac{p(y)}{p(x)} \\
&= \sum_{x\in\Omega_x} -p(x)\log p(x)
\end{aligned}$$

Subadditivity

Aus Kettenregel und dropping conditioning folgt dann direkt:

$$H(X_1, \ldots X_n) \leq \sum_{i=1 \ldots n} H(X_i)$$

Interpretation: Variablen, die man zusammen “sieht”, können nicht überraschender sein, als wenn man sie getrennt sieht, da jede Abhängigkeit zwischen den Variablen die Überraschung reduziert.

Übersicht

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \ldots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \ldots X_n) = n$

Insbesondere:

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

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \ldots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \ldots X_n) = n$

Insbesondere:

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

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \ldots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \ldots X_n) = n$

Insbesondere:

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

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \ldots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \ldots X_n) = n$

Insbesondere:

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

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \ldots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \ldots X_n) = n$

Insbesondere:

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

Entropie als Grenzwert in “Questions Game”

Ein Freund wirft eine Münze n-mal. X_i spezifiziert den Ausgang des i-ten Wurfes. Wieviele Fragen $H_o(X_1, \dots X_n)$ muss man im Durschnitt stellen, um das Ergebnis zu erfahren? Hierbei soll man seine Strategie optimieren. Man darf ODER Fragen stellen (*Ist das Ergebnis Y oder Z*) und bekommt nur Ja/Nein-Antworten.

- Münze fair, $n = 1$.
  Man muss immer genau eine Frage stellen. $H_0(X) = 1$
- Münze fair, $n = 2$.
  Ist es KK oder KZ?. Je nach Antwort, *Ist es KZ* oder *Ist es ZZ?*
  $H_0(X, X) = 2$
- Münze fair, Generelles n
  $H_o(X_1, \dots X_n) = n$

Insbesondere:

$$\lim_{n \to \infty} \frac{1}{n} H_0(X_1, \dots X_n) = 1 = H(X_i)$$

Entropie als Grenzwert in Questions Game

An der Tafel:

- Münze gibt immer Zahl.
- Münze hat Verteilung $p(K) = 0.75$, $p(Z) = 0.25$. $n = 1$.
- Münze hat Verteilung $p(K) = 0.75$, $p(Z) = 0.25$. $n = 2$.

Interpretation der Entropie als Limes im “Questions Game”

Man kann zeigen:

$$H_0(X_1) \geq \frac{1}{2} H_0(X_1, X_2) \geq \frac{1}{3} H_0(X_1, X_2, X_3) \geq \dots$$

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H_0(X_1, \dots X_n)$$

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 $H(\text{English})$ zwischen 1.25 und 1.35.

`http://math.ucsd.edu/~crypto/java/ENTROPY/` (at the moment defunct)

Entropie des Englischen (Shannon)

Lückentext: (englisches Alphabet plus “space”)

_____

(Cross-)Entropie des Englischen

Wie gut sind Modelle des Englischen:

Model	(Cross-)Entropie
Uniform	4.76 (= $\log_2 27$)
unigram	4.03
bigram	2.8
Shannon's Experiment	1.34

Alles per-Buchstabenentropie! Wie sähe wohl die per-Wort Entropie aus?

Übersicht

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

- $I(X;Y) \geq 0$, da $I(X;Y) = H(X) - H(X|Y)$ sowie dropping conditioning
- $I(X;Y) = I(Y;X)$;
- $I(X;Y)$ 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 $H(Y)$;
- $I(X;X) = H(X) - H(X|X) = H(X)$

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

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

- $-\infty \leq pmi(x;y) \leq min(-\log p(x), -\log p(y))$
- Dies heisst bei perfekter Abhängigkeit, steigt pmi für seltene Wörter/Bigramme
- Perfekte Unabhängigkeit: $pmi(x,y) = \log \frac{p(x,y)}{p(x)p(y)} = \log 1 = 0$
- pmi gutes Maß für Unabhängigkeit!

Pointwise Mutual Information für Kollokationen

Table 5. 16 aus Manning und Schütze (links mit 1000 Dokumenten, rechts mit 23,000 Dokumenten)

I_{1000}	w^1	w^2	w^1w^2	Bigram	I_{23000}	w^1	w^2	w^1w^2	Bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Variationen von Pointwise Mutual Information

- Benutze PMI nur für Wörter, die mindestens dreimal im Korpus vorkommen
- Positive Pointwise Mutual Information:

$$ppmi(x, y) = max(0, pmi(x, y))$$

Übersicht

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 als gemittelte Anzahl an Fragen, die man stellen muss, um nächste Nachricht zu erraten
- Pointwise Mutual Information als Assoziationsmaß zwischen zwei Wörtern
- Probleme PMI: Data Sparseness

Ausblick

- Von Kollokationen zu fensterbasierten Darstellungen von Wörtern als Vektoren
- Vektorräume und Vektoroperationen
- Distanzen und Ähnlichkeiten zwischen Vektoren
- Benutzung zur Wortähnlichkeitsberechnung
- Evaluation: Korrelationen zu menschlichen Ähnlichkeiten

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.

Aufgaben

Übungsblatt 1: Aufgabe 1

Übersicht

Interpretation der Entropie in Codierungstheorie (Optional)

Suche nach einem binären Präfixcode, der die Nachricht möglichst effizient übermittelt.

Präfixcode

Ein Code, bei dem kein Codewort Präfix eines anderen Codewortes ist. Dies erlaubt konkatenierte Codewörter ohne Space. Beispiel (nicht binär): Ländervorwahlen.

Präfixcode für einen 8-seitigen fairer Würfel (Entropie = 3 Bits).

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

Brauche 3 Bit = Entropie!

Interpretation der Entropie in Codierungstheorie (Optional)

Suche nach einem binären Präfixcode, der die Nachricht möglichst effizient übermittelt.

Präfixcode

Ein Code, bei dem kein Codewort Präfix eines anderen Codewortes ist. Dies erlaubt konkatenierte Codewörter ohne Space. Beispiel (nicht binär): Ländervorwahlen.

Präfixcode für einen 8-seitigen fairer Würfel (Entropie = 3 Bits).

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

Brauche 3 Bit = Entropie!

Codierung für Simplified Polynesian (Optional)

Berechne die Buchstabenentropie der folgenden Sprache mit nur 6 Buchstaben:

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

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

Das heisst es gibt einen binären Präfixcode der nur 2.5 bits braucht, um einen Buchstaben zu übermitteln.

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

Codierung für Simplified Polynesian (Optional)

Berechne die Buchstabenentropie der folgenden Sprache mit nur 6 Buchstaben:

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

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

Das heisst es gibt einen binären Präfixcode der nur 2.5 bits braucht, um einen Buchstaben zu übermitteln.

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

Codierung für Simplified Polynesian (Optional)

Berechne die Buchstabenentropie der folgenden Sprache mit nur 6 Buchstaben:

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

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

Das heisst es gibt einen binären Präfixcode der nur 2.5 bits braucht, um einen Buchstaben zu übermitteln.

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