Wie misst man Information?

Vom Bit zum maschinellen Lernen

Stephan Eckstein

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Tag der Mathematik 2026

Machine *Learning* (ML)

Lernen aus Daten $\widehat{=}$ Information über die Welt aufnehmen

Machine *Learning* (ML)

Lernen aus Daten $\widehat{=}$ Information über die Welt aufnehmen

Ziel des Vortrags:
Mathematische Formalisierung von Information.

Plan:

1. Intuition: Zahlenraten und Shannons Ratespiel
2. Information ohne Kontext: Entropie
3. Information mit Kontext: Bedingte Entropie
4. Informationsdiskrepanz: Relative Entropie und ML, oder *wie ChatGPT sprechen lernte*

1 **Intuition**

Zahlenraten und Shannons Ratespiel

2 Information ohne Kontext

Entropie

3 Information mit Kontext

Bedingte Entropie

4 Informations-diskrepanz

Relative Entropie und Lernziele in ML

Zahlenraten: Wie viele Fragen braucht man?

Spiel: Ein Computer generiert eine zufällige Zahl zwischen 1 und 8.
Wie können wir die Zahl mit möglichst wenigen Ja/Nein-Fragen herausfinden?

Zahlenraten: Wie viele Fragen braucht man?

Spiel: Ein Computer generiert eine zufällige Zahl zwischen 1 und 8.
Wie können wir die Zahl mit möglichst wenigen Ja/Nein-Fragen herausfinden?

1 2 3 4 **5 6 7 8**

Frage 1: "Ist die Zahl > 4?" $\rightarrow$ Ja! Es bleiben 4 Zahlen.

Zahlenraten: Wie viele Fragen braucht man?

Spiel: Ein Computer generiert eine zufällige Zahl zwischen 1 und 8.
Wie können wir die Zahl mit möglichst wenigen Ja/Nein-Fragen herausfinden?

1 2 3 4 **5 6** 7 8

Frage 1: "Ist die Zahl > 4?" $\rightarrow$ Ja! Es bleiben 4 Zahlen.

Frage 2: "Ist die Zahl > 6?" $\rightarrow$ Nein! Es bleiben 2 Zahlen.

Zahlenraten: Wie viele Fragen braucht man?

Spiel: Ein Computer generiert eine zufällige Zahl zwischen 1 und 8.
Wie können wir die Zahl mit möglichst wenigen Ja/Nein-Fragen herausfinden?

1 2 3 4 5 **6** 7 8

Frage 1: "Ist die Zahl > 4?" → Ja! Es bleiben 4 Zahlen.

Frage 2: "Ist die Zahl > 6?" → Nein! Es bleiben 2 Zahlen.

Frage 3: "Ist die Zahl > 5?" → Ja! Die Zahl ist **6**!

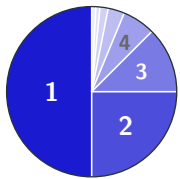
Möglichkeiten halbieren = eine Ja/Nein-Frage beantworten = 1 Bit Information.

Um das Spiel zu gewinnen brauchen wir $\log_2(8) = 3$ Fragen, bzw. 3 Bit Information.

Zahlenraten: Zweite Variante

Neues Spiel: Zahlen sind nicht gleich häufig. Aus einem Beutel mit 128 Zetteln (auf jedem Zettel steht eine Zahl zwischen 1-8) wird einer zufällig gezogen.

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1

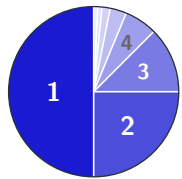


Ziel: Bei mehrfachem Spielen die Zahl mit *im Schnitt* wenigen Fragen erraten.

Zahlenraten: Zweite Variante

Neues Spiel: Zahlen sind nicht gleich häufig. Aus einem Beutel mit 128 Zetteln (auf jedem Zettel steht eine Zahl zwischen 1-8) wird einer zufällig gezogen.

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1



Ziel: Bei mehrfachem Spielen die Zahl mit *im Schnitt* wenigen Fragen erraten.

Naive Strategie: Möglichkeiten halbieren

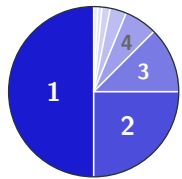
"Ist die Zahl > 4?" $\longrightarrow$ $\underbrace{6\,\%}_{\text{Ja}}$ / $\underbrace{94\,\%}_{\text{Nein}}$

$\rightarrow$ Im Schnitt 3 Fragen

Zahlenraten: Zweite Variante

Neues Spiel: Zahlen sind nicht gleich häufig. Aus einem Beutel mit 128 Zetteln (auf jedem Zettel steht eine Zahl zwischen 1-8) wird einer zufällig gezogen.

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1



Ziel: Bei mehrfachem Spielen die Zahl mit *im Schnitt* wenigen Fragen erraten.

Naive Strategie: Möglichkeiten halbieren

"Ist die Zahl > 4?" $\longrightarrow$ $\underbrace{6\,\%}_{\text{Ja}}$ / $\underbrace{94\,\%}_{\text{Nein}}$

$\rightarrow$ Im Schnitt 3 Fragen

Schlaue Strategie: W'keiten halbieren

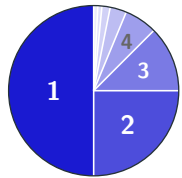
"Ist es die 1?" $\longrightarrow$ $\underbrace{50\,\%}_{\text{Ja}}$ / $\underbrace{50\,\%}_{\text{Nein}}$

$\rightarrow$ Im Schnitt nur 2 Fragen

Zahlenraten: Zweite Variante

Neues Spiel: Zahlen sind nicht gleich häufig. Aus einem Beutel mit 128 Zetteln (auf jedem Zettel steht eine Zahl zwischen 1-8) wird einer zufällig gezogen.

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1



Ziel: Bei mehrfachem Spielen die Zahl mit *im Schnitt* wenigen Fragen erraten.

Naive Strategie: Möglichkeiten halbieren

"Ist die Zahl > 4?" $\longrightarrow$ $\underbrace{6\,\%}_{\text{Ja}}$ / $\underbrace{94\,\%}_{\text{Nein}}$

$\rightarrow$ Im Schnitt 3 Fragen

Schlaue Strategie: W'keiten halbieren

"Ist es die 1?" $\longrightarrow$ $\underbrace{50\,\%}_{\text{Ja}}$ / $\underbrace{50\,\%}_{\text{Nein}}$

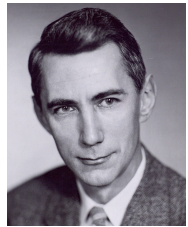
$\rightarrow$ Im Schnitt nur 2 Fragen

Eine gute Frage ist eine, bei der beide Antworten gleich wahrscheinlich sind.

Shannons Ratespiel (1951)

Spiel:

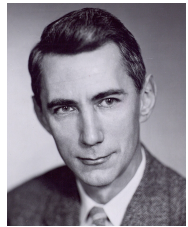
- Ein Computer wählt ein zufälliges deutsches Wort (ohne Sonderzeichen) aus allen Büchern einer Bibliothek aus.
- Wir wollen das Wort erraten, und zwar Buchstabe für Buchstabe von links nach rechts mit Ja/Nein-Fragen.
- Ziel: Das Wort mit im Schnitt möglichst wenigen Fragen erraten.



Claude Shannon
(1916–2001)

Begründer der Informationstheorie

Shannons Ratespiel (1951)



Claude Shannon
(1916–2001)

Begründer der
Informationstheorie

Spiel:

- Ein Computer wählt ein zufälliges deutsches Wort (ohne Sonderzeichen) aus allen Büchern einer Bibliothek aus.
- Wir wollen das Wort erraten, und zwar Buchstabe für Buchstabe von links nach rechts mit Ja/Nein-Fragen.
- Ziel: Das Wort mit im Schnitt möglichst wenigen Fragen erraten.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

5 Minuten Spiel- & Diskussionszeit → Ähnlichkeiten/Unterschiede zu Zahlenraten?

Shannons Ratespiel: Beobachtungen

1. **Obergrenze wie bei naiver Strategie beim Zahlenraten:**
   Pro Buchstabe maximal $\lceil \log_2(26) \rceil = 5$ Fragen (Möglichkeiten halbieren).

Shannons Ratespiel: Beobachtungen

1. **Obergrenze wie bei naiver Strategie beim Zahlenraten:**
   Pro Buchstabe maximal $\lceil \log_2(26) \rceil = 5$ Fragen (Möglichkeiten halbieren).

2. **Vorherige Buchstaben helfen:**
   Nach "M A T H" ist der nächste Buchstabe stark eingeschränkt.
   Wir brauchen in der Regel weniger Fragen, je mehr Kontext wir haben!

Shannons Ratespiel: Beobachtungen

1. **Obergrenze wie bei naiver Strategie beim Zahlenraten:**
   Pro Buchstabe maximal $\lceil \log_2(26) \rceil = 5$ Fragen (Möglichkeiten halbieren).

2. **Vorherige Buchstaben helfen:**
   Nach "M A T H" ist der nächste Buchstabe stark eingeschränkt.
   Wir brauchen in der Regel weniger Fragen, je mehr Kontext wir haben!

3. **Nicht alle Buchstaben sind gleich häufig:**
   Seltene Buchstaben erraten braucht länger, ist aber überraschender/informativer.
   Q kommt fast nie vor $\rightarrow$ sehr überraschend/sehr viel neue Information.
   E kommt eher oft vor $\rightarrow$ weniger überraschend/weniger neue Information.
   Um gut zu raten (50-50 Ja-Nein) müssen wir Häufigkeiten einschätzen können.


1
2
3
4
Intuition
Zahlenraten
und
Shannons Ratespiel
Information
ohne Kontext
Entropie
Information
mit Kontext
Bedingte Entropie
Informations-
diskrepanz
Relative Entropie
und
Lernziele in ML

Formalisierung von Information einzelner Ereignisse

Wie viel Information liefert ein Ereignis, welches mit Wahrscheinlichkeit p eintritt?

Wünschenswerte Eigenschaften:

- Je kleiner p, desto größer die Überraschung bzw. die gelieferte Information
- Sicheres Ereignis ($p = 1$): 0 Information
- Ja/Nein-Frage welche Möglichkeiten halbiert ($p = 0.5$): 1 Bit Information
- Ereignis Kombination zweier Ereignisse, die nichts miteinander zu tun haben: Information addiert sich

Formalisierung von Information einzelner Ereignisse

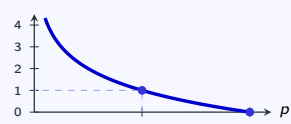
Wie viel Information liefert ein Ereignis, welches mit Wahrscheinlichkeit p eintritt?

Wünschenswerte Eigenschaften:

- Je kleiner p, desto größer die Überraschung bzw. die gelieferte Information
- Sicheres Ereignis ($p = 1$): 0 Information
- Ja/Nein-Frage welche Möglichkeiten halbiert ($p = 0.5$): 1 Bit Information
- Ereignis Kombination zweier Ereignisse, die nichts miteinander zu tun haben: Information addiert sich

Die einzige Funktion, die das leistet:

$$\text{Information}(p) = \log_2\left(\frac{1}{p}\right)$$



Formalisierung von Information: Entropie

Zufallsexperiment X: Ausgänge $x_1, \ldots, x_n$ mit Wahrscheinlichkeiten $p(x_1), \ldots, p(x_n)$.

(Z.B. $x_1 =$ "Anfangsbuchstabe ist A" mit $p(x_1) \approx 0.1$, $x_2 =$ "Anfangsbuchstabe ist B" $\ldots$)

Entropie $H(X)$ ist der durchschnittliche Informationsgehalt

$$H(X) = \underbrace{p(x_1)}_{\text{Wie oft?}} \cdot \underbrace{\log_2\Big(\frac{1}{p(x_1)}\Big)}_{\text{Informationsgehalt}} + \cdots + p(x_n) \cdot \log_2\Big(\frac{1}{p(x_n)}\Big)$$

Formalisierung von Information: Entropie

Zufallsexperiment X: Ausgänge $x_1, \ldots, x_n$ mit Wahrscheinlichkeiten $p(x_1), \ldots, p(x_n)$.
(Z.B. x_1 = "Anfangsbuchstabe ist A" mit $p(x_1) \approx 0.1$, x_2 = "Anfangsbuchstabe ist B" ...)

Entropie $H(X)$ ist der durchschnittliche Informationsgehalt

$$H(X) = \underbrace{p(x_1)}_{\text{Wie oft?}} \cdot \underbrace{\log_2\Big(\frac{1}{p(x_1)}\Big)}_{\text{Informationsgehalt}} + \cdots + p(x_n) \cdot \log_2\Big(\frac{1}{p(x_n)}\Big)$$

Ja/Nein-Frage beim Zahlenraten: $0.5 \log_2(\frac{1}{0.5}) + 0.5 \log_2(\frac{1}{0.5}) = \log_2(2) = 1$ Bit

Formalisierung von Information: Entropie

Zufallsexperiment X: Ausgänge $x_1, \ldots, x_n$ mit Wahrscheinlichkeiten $p(x_1), \ldots, p(x_n)$.
(Z.B. x_1 = "Anfangsbuchstabe ist A" mit $p(x_1) \approx 0.1$, x_2 = "Anfangsbuchstabe ist B" ...)

Entropie $H(X)$ ist der durchschnittliche Informationsgehalt

$$H(X) = \underbrace{p(x_1)}_{\text{Wie oft?}} \cdot \underbrace{\log_2\Big(\frac{1}{p(x_1)}\Big)}_{\text{Informationsgehalt}} + \cdots + p(x_n) \cdot \log_2\Big(\frac{1}{p(x_n)}\Big)$$

Ja/Nein-Frage beim Zahlenraten: $0.5 \log_2(\frac{1}{0.5}) + 0.5 \log_2(\frac{1}{0.5}) = \log_2(2) = 1$ Bit

"Ist der Anfangsbuchstabe ein A?": $0.1 \cdot \log_2(\frac{1}{0.1}) + 0.9 \cdot \log_2(\frac{1}{0.9}) \approx 0.47$ Bit

Formalisierung von Information: Entropie

Zufallsexperiment X: Ausgänge $x_1, \ldots, x_n$ mit Wahrscheinlichkeiten $p(x_1), \ldots, p(x_n)$.
(Z.B. x_1 = "Anfangsbuchstabe ist A" mit $p(x_1) \approx 0.1$, x_2 = "Anfangsbuchstabe ist B" ...)

Entropie $H(X)$ ist der durchschnittliche Informationsgehalt

$$H(X) = \underbrace{p(x_1)}_{\text{Wie oft?}} \cdot \underbrace{\log_2\Big(\frac{1}{p(x_1)}\Big)}_{\text{Informationsgehalt}} + \cdots + p(x_n) \cdot \log_2\Big(\frac{1}{p(x_n)}\Big)$$

Ja/Nein-Frage beim Zahlenraten:	$0.5 \log_2(\frac{1}{0.5}) + 0.5 \log_2(\frac{1}{0.5}) = \log_2(2) = 1$ Bit
"Ist der Anfangsbuchstabe ein A?":	$0.1 \cdot \log_2(\frac{1}{0.1}) + 0.9 \cdot \log_2(\frac{1}{0.9}) \approx 0.47$ Bit
Zufälliger Anfangsbuchstabe:	$0.1 \cdot \log_2(\frac{1}{0.1}) + \ldots + 0.03 \cdot \log_2(\frac{1}{0.03}) \approx 4.2$ Bit

Formalisierung von Information: Entropie

Zufallsexperiment X: Ausgänge $x_1, \ldots, x_n$ mit Wahrscheinlichkeiten $p(x_1), \ldots, p(x_n)$.
(Z.B. x_1 = "Anfangsbuchstabe ist A" mit $p(x_1) \approx 0.1$, x_2 = "Anfangsbuchstabe ist B" ...)

Entropie $H(X)$ ist der durchschnittliche Informationsgehalt

$$H(X) = \underbrace{p(x_1)}_{\text{Wie oft?}} \cdot \underbrace{\log_2\Big(\frac{1}{p(x_1)}\Big)}_{\text{Informationsgehalt}} + \cdots + p(x_n) \cdot \log_2\Big(\frac{1}{p(x_n)}\Big)$$

Ja/Nein-Frage beim Zahlenraten:	$0.5 \log_2(\frac{1}{0.5}) + 0.5 \log_2(\frac{1}{0.5}) = \log_2(2) = 1$ Bit
"Ist der Anfangsbuchstabe ein A?":	$0.1 \cdot \log_2(\frac{1}{0.1}) + 0.9 \cdot \log_2(\frac{1}{0.9}) \approx 0.47$ Bit
Zufälliger Anfangsbuchstabe:	$0.1 \cdot \log_2(\frac{1}{0.1}) + \ldots + 0.03 \cdot \log_2(\frac{1}{0.03}) \approx 4.2$ Bit
Theorem: (Obergrenze Entropie)	Es gilt stets $H(X) \leq \log_2(n)$

Entropie Obergrenze $H(X) \leq \log_2(n)$

- $\log_2(n)$ ist die Entropie, wenn alle n Möglichkeiten gleich wahrscheinlich sind.
- $H(X) \leq \log_2(n)$ sagt: Die Entropie/Unsicherheit ist am größten, wenn alle Möglichkeiten gleich wahrscheinlich sind.

Entropie Obergrenze $H(X) \leq \log_2(n)$

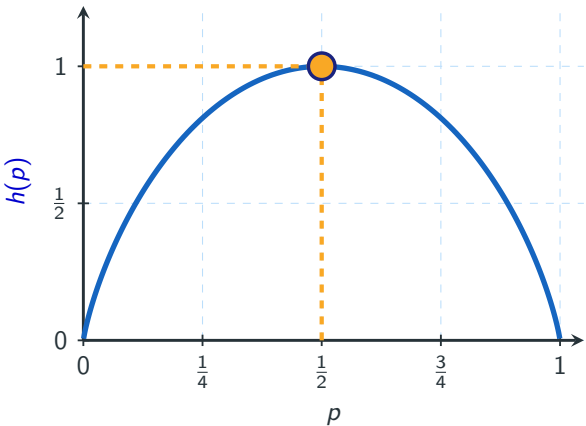
- $\log_2(n)$ ist die Entropie, wenn alle n Möglichkeiten gleich wahrscheinlich sind.
- $H(X) \leq \log_2(n)$ sagt: Die Entropie/Unsicherheit ist am größten, wenn alle Möglichkeiten gleich wahrscheinlich sind.

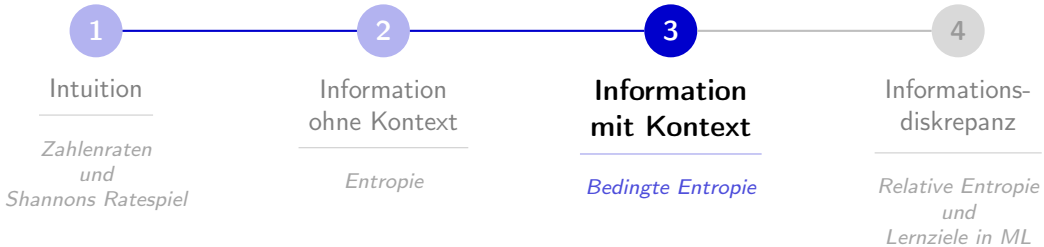
Für $n = 2$ zwei Möglichkeiten mit W'keiten p und $(1-p)$ ist Entropie

$$h(p) = p \log_2\left(\tfrac{1}{p}\right) + (1-p) \log_2\left(\tfrac{1}{1-p}\right)$$

Maximal für $p = \frac{1}{2}$.

$\rightarrow$ Maximale Info $1 = \log_2(2)$ durch 50-50 Ja/Nein Frage erzielt




1
Intuition
Zahlenraten
und
Shannons Ratespiel
2
Information
ohne Kontext
Entropie
3
Information
mit Kontext
Bedingte Entropie
4
Informations-
diskrepanz
Relative Entropie
und
Lernziele in ML

Information mit Kontext: Bedingte Entropie

Zufallsexperiment X mit Kontext Y.

(Z.B. X zufälliger zweiter Buchstabe des Wortes im Kontext des ersten Buchstaben Y)

- Für jeden realisierten Kontext $Y = y$ ändern sich Wahrscheinlichkeiten für X.
  (Z.B. Für $y =$ "1.Buchstabe ist M" ist p("2. Buchstabe ist A") ≈ 0.26)
- Für Kontext $Y = y$ ist $H(X \mid Y = y)$ Entropie von X mit angepassten W'keiten.

Information mit Kontext: Bedingte Entropie

Zufallsexperiment X mit Kontext Y.

(Z.B. X zufälliger zweiter Buchstabe des Wortes im Kontext des ersten Buchstaben Y)

- Für jeden realisierten Kontext $Y = y$ ändern sich Wahrscheinlichkeiten für X.
  (Z.B. Für $y =$ "1.Buchstabe ist M" ist p("2. Buchstabe ist A") ≈ 0.26)
- Für Kontext $Y = y$ ist $H(X \mid Y = y)$ Entropie von X mit angepassten W'keiten.
- Bedingte Entropie $H(X \mid Y)$ ist gewichtetes Mittel von $H(X \mid Y = y)$ über alle y.

$$H(X \mid Y) = p(y_1) \cdot H(X \mid Y = y_1) + \ldots + p(y_n) \cdot H(X \mid Y = y_n)$$

Information mit Kontext: Bedingte Entropie

Zufallsexperiment X mit Kontext Y.

(Z.B. X zufälliger zweiter Buchstabe des Wortes im Kontext des ersten Buchstaben Y)

- Für jeden realisierten Kontext $Y = y$ ändern sich Wahrscheinlichkeiten für X.
  (Z.B. Für $y =$ "1.Buchstabe ist M" ist p("2. Buchstabe ist A") ≈ 0.26)
- Für Kontext $Y = y$ ist $H(X \mid Y = y)$ Entropie von X mit angepassten W'keiten.
- Bedingte Entropie $H(X \mid Y)$ ist gewichtetes Mittel von $H(X \mid Y = y)$ über alle y.

$$H(X \mid Y) = p(y_1) \cdot H(X \mid Y = y_1) + \ldots + p(y_n) \cdot H(X \mid Y = y_n)$$

Theorem (Im Durchschnitt hilft Kontext immer, Unsicherheit zu reduzieren)

Es gilt stets $H(X \mid Y) \leq H(X)$.


1
Intuition
Zahlenraten
und
Shannons Ratespiel
2
Information
ohne Kontext
Entropie
3
Information
mit Kontext
Bedingte Entropie
4
Informations-
diskrepanz
Relative Entropie
und
Lernziele in ML

Informationsdiskrepanz: Zurück zum Zahlenraten

- Um gute Fragen zu stellen, *müssen wir die Wahrscheinlichkeiten kennen*
- Oft kennt man diese nicht, dann herrscht *Informationsdiskrepanz*

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1
Wahrscheinlichkeiten P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$
Wir denken Q	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Informationsdiskrepanz: Zurück zum Zahlenraten

- Um gute Fragen zu stellen, *müssen wir die Wahrscheinlichkeiten kennen*
- Oft kennt man diese nicht, dann herrscht *Informationsdiskrepanz*

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1
Wahrscheinlichkeiten P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$
Wir denken Q	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

- Optimale Strategie unter Annahme von Q (Möglichkeiten halbieren) $\rightarrow$ 3 Fragen
- Optimale Strategie unter Annahme von P $\rightarrow$ Im Schnitt nur 2 Fragen

Informationsdiskrepanz: Zurück zum Zahlenraten

- Um gute Fragen zu stellen, *müssen wir die Wahrscheinlichkeiten kennen*
- Oft kennt man diese nicht, dann herrscht *Informationsdiskrepanz*

Zahl	1	2	3	4	5	6	7	8
Anzahl Zettel	64	32	16	8	4	2	1	1
Wahrscheinlichkeiten P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$
Wir denken Q	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

- Optimale Strategie unter Annahme von Q (Möglichkeiten halbieren) $\rightarrow$ 3 Fragen
- Optimale Strategie unter Annahme von P $\rightarrow$ Im Schnitt nur 2 Fragen
- Informationsdiskrepanz zwischen P und Q: 3 Fragen - 2 Fragen $=$ 1 Frage $\widehat{=}$ 1 bit.
- Diese Informationsdiskrepanz heißt **relative Entropie** zwischen P und Q.

Informationsdiskrepanz: Relative Entropie

- Perfektes Raten benötigt Kenntnis der W'keiten $P = (p(x_1), \ldots, p(x_n))$.
- Ein Spieler schätzt diese fälschlicherweise als $Q = (q(x_1), \ldots, q(x_n))$ ein.
- *Informationsdiskrepanz*: **relative Entropie** $R(P, Q)$ zwischen P und Q.

$$R(P, Q) = p(x_1) \cdot \log_2(\tfrac{p(x_1)}{q(x_1)}) + \ldots + p(x_n) \cdot \log_2(\tfrac{p(x_n)}{q(x_n)})$$

Hierbei ist $\log_2(\frac{p(x_1)}{q(x_1)}) = \log_2(\frac{1}{q(x_1)}) - \log_2(\frac{1}{p(x_1)})$ der Unterschied zwischen der falschen Überraschung und der korrekten Überraschung beim Eintreffen des Ereignisses x_1.

Informationsdiskrepanz: Relative Entropie

- Perfektes Raten benötigt Kenntnis der W'keiten $P = (p(x_1), \ldots, p(x_n))$.
- Ein Spieler schätzt diese fälschlicherweise als $Q = (q(x_1), \ldots, q(x_n))$ ein.
- *Informationsdiskrepanz*: **relative Entropie** $R(P, Q)$ zwischen P und Q.

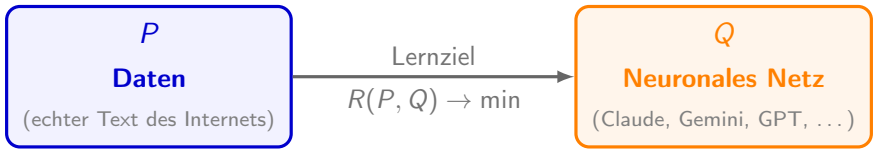
$$R(P, Q) = p(x_1) \cdot \log_2(\tfrac{p(x_1)}{q(x_1)}) + \ldots + p(x_n) \cdot \log_2(\tfrac{p(x_n)}{q(x_n)})$$

Hierbei ist $\log_2(\frac{p(x_1)}{q(x_1)}) = \log_2(\frac{1}{q(x_1)}) - \log_2(\frac{1}{p(x_1)})$ der Unterschied zwischen der falschen Überraschung und der korrekten Überraschung beim Eintreffen des Ereignisses x_1.

Theorem (Bis auf Rundung)

Spielt man ein Ratespiel anhand falscher Wahrscheinlichkeiten Q statt P, so braucht man im Schnitt $R(P, Q)$ viele zusätzliche Fragen.

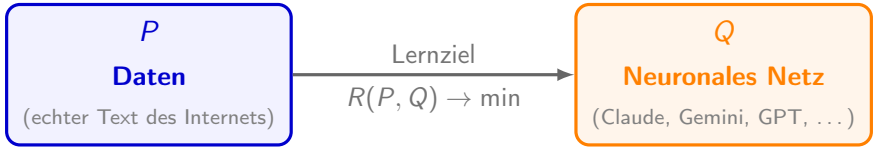
Wie Maschinen sprechen lernten: Ratespiele und Sprachmodelle



Training von Sprachmodellen $\widehat{=}$ Shannons Ratespiel auf dem Internet:

Das Modell Q versucht, Wort für Wort den nächsten Text vorherzusagen.
Je kleiner $R(P, Q)$, desto besser und desto mehr wird es belohnt.

Wie Maschinen sprechen lernten: Ratespiele und Sprachmodelle



Training von Sprachmodellen $\widehat{=}$ Shannons Ratespiel auf dem Internet:
Das Modell Q versucht, Wort für Wort den nächsten Text vorherzusagen.
Je kleiner $R(P, Q)$, desto besser und desto mehr wird es belohnt.

- Sprachmodelle zerlegen Wörter nicht in Buchstaben, sondern *Token*. Z.B. Mat-hemat-ik
- Relative Entropie minimieren $\widehat{=}$ Training mit *cross-entropy-loss* $\widehat{=}$ möglichst gut raten
- Effizient Kontext einbringen durch *Transformer*-Architektur von neuronalen Netzen
- Shannons Ratespiel "nur" der erste (und größte) Teil des Trainings, das sog. *Pretraining*