Anhang: Informationstheorie

Wir wollen uns des abstrakten Begriffes der „Information" ermächtigen, ihn qualitativ und quantitativ beschreibbar machen. Das ist nicht nur von philosophischem Interesse sondern erweist sich in vielen Gebieten als nützlich, unter anderem: Informatik, Nachrichtentechnik, Kryptologie, Neurologie, Biologie, Suche nach Plagiaten.

Versuchen wir, die „Informationsmenge" als eine reelle Zahl $f \in \mathbb{R}$ zu definieren. Ein (zu) einfaches Modell: Es gibt f die Menge an Information, die nötig ist, um eine „Frage" zu beantworten. Gesetzt, eine Frage kann in ja/nein-Fragen zerlegt werden, dann suchen wir die Anzahl solcher Fragen, die nötig ist, um eine Antwort zu finden, und versehen diese Anzahl mit der Einheit Bit.

Bsp. Eine Schulklasse bestehe aus 64 Schülern (Lehrermangel!).
Frage: „Welcher Schüler ist Tim?"

Strategie: Teile die Klasse in zwei Gruppen A_1 und A_2 und frage „Ist Tim in A_1?"; teile die Gruppe, die Tim enthält erneut in zwei Gruppen, u.s.w., bis nur noch zwei Schüler übrig sind.

Die Anzahl benötigter ja/nein-Fragen beträgt $\log_2(64) = 6$.
Das heißt: Wir assoziieren die Existenz von Tim in einer Gruppe aus 64 Schülern mit einer Informationsmenge von $f = 6$ bit.

Dem Beispiel folgend ordnen wir einer N-elementigen (diskreten) Menge $\mathcal{A} = \{a_1, a_2, \dots, a_N\}$ den Informationsgehalt

$$f(\mathcal{A}) = \log_2(N) \text{ bit}$$

zu. Da im Allgemeinen $N \neq 2^n$, werden auch nicht-ganze Bits zugelassen.

Achtung: Das ist nicht die mathematische Definition (s. unten)! Wir behelfen uns mit dieser Zuordnung nur vorläufig.

Die Informationsmenge erhöht sich, wenn mehrere Mengen in Kombination betrachtet werden, z.B.:

$\mathcal{A}_1 = \{\text{rot, gelb, grün, blau}\} \longrightarrow f(\mathcal{A}_1) = 2 \text{ bit}$

$\mathcal{A}_2 = \{\text{Hund, Katze, Maus}\} \longrightarrow f(\mathcal{A}_2) = \log_2(3) \text{ bit}$

Produktmenge: $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$

$= \{$roter Hund, rote Katze, rote Maus, gelber Hund, gelbe Katze, ...$\}$

$\longrightarrow f(\mathcal{A}) = \log_2(4 \cdot 3) \text{ bit} = 2 \text{ bit} + \log_2(3) \text{ bit}$

Offenbar gilt ein Additivitätsgesetz,

$$f(\mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_n) = f(\mathcal{A}_1) + f(\mathcal{A}_2) + \dots + f(\mathcal{A}_n).$$

Bem. Ließe man nur ganzzahlige Bits zu, würde in jedem Summanden aufgerundet werden. Subadditivität:

$$f(\mathcal{A}_1 \times \dots \times \mathcal{A}_n) \leq f(\mathcal{A}_1) + \dots + f(\mathcal{A}_n).$$

Im Additivitätsgesetz tragen alle Mengen gemäß der Anzahl ihrer Elemente zur Gesamtinformation bei, d.h. die $f(A_i)$ gelten alle als gleich „wichtig". Das wird im Allgemeinen nicht so sein.

Bsp. Es soll ein englisches Wort aus 5 Buchstaben erraten (identifiziert) werden. Offenbar würden wir der Aussage „Es kommt ein ‚e' darin vor" einen geringeren Informationsgehalt zuordnen als der Aussage „Es kommt ein ‚j' darin vor", da ‚e' der häufigste Buchstabe des englischen Alphabets ist, ‚j' hingegen der seltenste. Letztere Aussage schränkt also die Menge möglicher Worte stärker ein als erstere.

Ordnen wir jedem A_i eine Wichtung / Wahrscheinlichkeit $w_i(A_i)$ zu mit $\sum_{i=1}^{n} w_i(A_i) = 1$, dann gilt:

$$f(A) = w_1 \log_2\left(\frac{1}{w_1}\right) + w_2 \log_2\left(\frac{1}{w_2}\right) + \dots + w_n \log_2\left(\frac{1}{w_n}\right).$$

Das ist die berühmte <u>Shannon-Formel</u>. Sie geht zurück auf Pauli und von Neumann.

→ Sind alle Wahrscheinlichkeiten gleich, gilt $w_i = \frac{1}{n}$.

→ Gibt es nur ein Element mit Wahrscheinlichkeit 1, dann gilt: $f(A) = \log_2(1) = 0$. Ein Ereignis, das mit Sicherheit eintritt, liefert keine Information.

Ein wenig erinnert die Information an die Definition der Boltzmann-Entropie, $S \sim \ln(\text{Anzahl Realisierungsmöglichkeiten})$. Tatsächlich sind beide Größen fast identisch:

Ω: mikrokanonische Zustandssumme $\longrightarrow$ Wahrscheinlichkeit $w = \frac{1}{\Omega}$

$$\left.\begin{aligned} \text{Shannon:} \quad & f(\Omega) = \log_2(\Omega) \\ \text{Boltzmann:} \quad & S = k_B \ln(\Omega) \end{aligned}\right\} \quad S = k_B \ln(2)\, f$$

Diese Identifizierung klappt auch für die (groß-)kanonische Zustandssumme.

Das heißt: Die Informationsmenge einer Zustandssumme ist gleich der fehlenden Information über die Mikrozustände, wenn nur der Makrozustand bekannt ist.

Wir können die <u>Shannon-Entropie</u> als das allgemeinere Konzept auffassen und die Boltzmann-Entropie als einen Spezialfall dessen (wenn lediglich der Makrozustand gegeben ist). Damit löst sich auch der Maxwell'sche Dämon auf, da er die Mikrozustände kennt.

In der Informationstheorie gibt die Shannon-Entropie die optimale Kompressionsrate an, d.h. die theoretischen Obergrenzen für die Verschlüsselung von Information.

Eine anständige Definition der Information

Nach der bisherigen, sehr umgangssprachlichen Beschreibung der Informationsmenge mag man sich nach einer klareren Sprechweise sehnen. Daher folgt nun eine mathematisch rigorose Einführung auf Grundlage einfacher Stochastik.

Die Ausführungen folgen größtenteils Mittelbach, Wolf, Jorswieck: „Skript zur Lehrveranstaltung Informationstheorie", TU Dresden, 2016.

Ein realer/gedachter Vorgang, dessen Ausgang nicht (notwendigerweise) vorhersagbar ist, heißt Zufallsexperiment. Alle möglichen Ausgänge eines Zufallsexperimentes bilden die Grundmenge Ω. Die Elemente $\omega \in \Omega$ heißen Elementarereignisse. Teilmengen[1] von Ω heißen Ereignisse.

Def. Eine Funktion P, die jedem Ereignis $A \subseteq \Omega$ eine Zahl $P(A) \in [0,1]$ zuordnet, heißt Wahrscheinlichkeitsmaß, wenn gilt:

- Nichtnegativität, $P(A) \geq 0$;
- Normierung, $P(\Omega) = 1$;
- σ-Additivität, $P(\bigcup_k A_k) = \sum_k P(A_k)$ für abzählbare, unvereinbare Ereignisse $A_1, A_2, \ldots$.

Der Funktionswert $P(A)$ heißt Wahrscheinlichkeit des Ereignisses A.

1: Genauer: Elemente der sogenannten σ-Algebra von Ω.

Def. Eine messbare[2] Funktion $X: \Omega \to \mathbb{R}$ heißt Zufallsgröße.

Eine messbare Funktion $X = (X_1, \dots, X_n): \Omega \to \mathbb{R}^n$ heißt Zufallsvektor.

Es seien nun in $x = (x_1, \dots, x_n)$ mögliche Werte der Funktion X zusammengefasst. Dann gibt $p_X(x) := P(X = x)$ die Wahrscheinlichkeit an, dass $X_1 = x_1$, $X_2 = x_2$, u.s.w. und heißt Wahrscheinlichkeitsfunktion.

Def. Die Entropie $f(X)$ eines Zufallsvektors X mit Wahrscheinlichkeitsfunktion p_X ist definiert als[3]

$$f(X) := -\sum_x p_X(x) \log_2 p_X(x) .$$

Demnach ist die „Informationsmenge" nicht (wie oben behelfsmäßig notiert) der Grundmenge zugeordnet, sondern Funktionen darauf.

Exakter gibt $-\log_2 p_X(x)$ die Information, die damit assoziiert ist, dass der Zufallsvektor X den Wert x annimmt. In diesem Sinne ist die Entropie die gewichtete Summe der einzelnen Informationen, sprich der mittlere Informationsgehalt einer Zufallsgröße.

Wie zuvor: Je geringer die Wahrscheinlichkeit eines Ereignisses, desto größer die Information, wenn es eintritt.

2: Hier nicht erklärt, siehe Maßtheorie.

3: Beachte, dass das Minus nicht neu ist: $\log_2\left(\frac{1}{p}\right) = \underbrace{\log_2(1)}_{0} - \log_2(p)$.

Bsp. Kehren wir zurück zum Eingangsbeispiel der 64-köpfigen Schulklasse.

Zufallsexperiment: Ziehe einen aus 64 Schülern.

$\Omega = \{\text{Schüler 1}, \text{Schüler 2}, \ldots, \text{Schüler 64}\}$

Anstelle eines Namens ordnen wir jedem Schüler einen Vektor zu:

$$\begin{aligned} X(\text{Schüler 1}) &= (1,0,0,\ldots,0) \\ X(\text{Schüler 2}) &= (0,1,0,\ldots,0) \\ &\vdots \\ X(\text{Schüler 64}) &= (0,0,0,\ldots,1). \end{aligned}$$

Die Wahrscheinlichkeit gezogen zu werden ist für jeden Schüler gleich,

$$p_X(1,0,\ldots,0) = p_X(0,1,\ldots,0) = \ldots = p_X(0,0,\ldots,1) = \frac{1}{64}.$$

Dass jedem Schüler ein Name/Vektor zugeordnet ist, ist also mit einer Informationsentropie von

$$I(X) = -\sum_x p_X(x) \log_2 p_X(x) = -64 \cdot \frac{1}{64} \log_2 \frac{1}{64} = 6 \text{ bit}$$

verbunden.

Die Menge aller Werte $x \in \mathbb{R}^n$, die von dem Zufallsvektor X angenommen werden können, bezeichnet man als <u>Alphabet</u>. Ein Alphabet kann je nach Anwendung willkürlich festgelegt werden; im Beispiel spielt es für das Ergebnis keine Rolle, dass wir den Schülern 64-dimensionale Einheitsvektoren zugeordnet haben.

Abschließend definieren wir noch, was bei gleichzeitiger Betrachtung zweier Zufallsgrößen geschieht.

Def. Es seien A_1 und A_2 Ereignisse und $P(A_2) > 0$. Dann heißt

$$P(A_1 | A_2) := \frac{P(A_1 \cap A_2)}{P(A_2)}$$

die bedingte Wahrscheinlichkeit von A_1 unter der Bedingung A_2.

Seien $X = (X_1, \dots, X_m)$ und $Y = (Y_1, \dots, Y_n)$ Zufallsvektoren und sei $p_{X,Y}$ die Wahrscheinlichkeitsfunktion von (X, Y) und p_X die von X. Dann heißt

$$p_{Y|X}(y|x) := \frac{p_{X,Y}(x,y)}{p_X(x)}$$

die bedingte Wahrscheinlichkeitsfunktion von Y unter der Bedingung $X = x$.

Def. Es seien X, Y, $p_{Y|X}$ und p_X wie oben. Die bedingte Entropie von Y unter der Bedingung $X = x$ ist definiert als

$$f(Y|X=x) := -\sum_Y p_{Y|X}(y|x) \log_2 p_{Y|X}(y|x).$$

Es heißt

$$f(Y|X) := \sum_x p_X(x) f(Y|X=x)$$

die bedingte Entropie von Y unter der Bedingung X.

Anwendungen

1. Redundanz der Sprache

Die Informationsentropie ist dann am größten, wenn man Gleichverteilungen betrachtet. Sind die Wahrscheinlichkeiten der Zufallsgrößen nicht gleich, so ergibt sich eine geringere Entropie und man bezeichnet die Differenz als Redundanz, $R = I_{max} - I$.

In der deutschen Sprache werden nicht alle 26 Buchstaben gleich häufig benutzt; man erhält („experimentell") eine Information (pro Zeichen) von

$$I = 4{,}063 \text{ bit}.$$

Bei Gleichverteilung würde die Information pro Zeichen $I_{max} = \log_2(26) = 4{,}700$ bit betragen.

$\longrightarrow$ Redundanz pro Zeichen $R = 0{,}637$ bit

$\longrightarrow \frac{R \cdot 26}{I} = 4{,}08 \Rightarrow$ Es gibt 4 redundante Zeichen im deutschen Alphabet.

Bem.
- Hier wird nur die reine Kodierung betrachtet, nicht die Zusammenfassung häufiger Kombinationen („ch", „sch", „ie") oder ähnlich klingender Laute.
- Nachrichtentechnisch gesehen erhöhen Redundanzen die Übertragungssicherheit.

2. Informationsgehalt medizinischer Tests

In folgender klinischer Studie wurden 7 Patienten 4 verschiedenen Tests unterzogen. Davon unabhängig wurde festgestellt, welche der Patienten eine Therapie benötigen.

X_1 Alter über 60	X_2 Raucher	X_3 systolischer Blutdruck >140	X_4 diastolischer Blutdruck >90	X Therapie nötig
nein	ja	nein	nein	nein
nein	ja	ja	ja	ja
ja	nein	ja	nein	nein
ja	ja	ja	ja	ja
ja	nein	nein	nein	nein
ja	ja	nein	nein	nein
nein	ja	ja	nein	nein

Wir wollen herausfinden, welcher der Tests am aussagekräftigsten ist.

Die Verteilung der Zufallsvariable X führt auf eine Entropie

$$f(X) = \frac{2}{7}\log_2\left(\frac{7}{2}\right) + \frac{5}{7}\log_2\left(\frac{7}{5}\right) = 0{,}863 \text{ bit}.$$

Da die Ereignisse der X_i nicht stochastisch unabhängig von denen in X sind, muss hier jeweils die bedingte Entropie berechnet werden; genauer gesagt: die bedingte Entropie von X unter der Bedingung X_i.

Betrachten wir beispielsweise das Alter (X_1):

3 von 7 sind unter 60:

→ $p_{X_1}(\text{nein}) = \frac{3}{7}$

→ davon: 2 von 3 keine Therapie ⇒ $p_{X|X_1}(\text{nein}|\text{nein}) = \frac{2}{3}$

1 von 3 Therapie ⇒ $p_{X|X_1}(\text{ja}|\text{nein}) = \frac{1}{3}$

4 von 7 sind über 60:

→ $p_{X_1}(\text{ja}) = \frac{4}{7}$

→ davon: 3 von 4 keine Therapie ⇒ $p_{X|X_1}(\text{nein}|\text{ja}) = \frac{3}{4}$

1 von 4 Therapie ⇒ $p_{X|X_1}(\text{ja}|\text{ja}) = \frac{1}{4}$

alles zusammen:

$$f(X|X_1) = \frac{3}{7}\left(\frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3\right) + \frac{4}{7}\left(\frac{3}{4}\log_2\frac{4}{3} + \frac{1}{4}\log_2 4\right)$$

$$= \frac{3}{7}\log_2 3 - \frac{2}{7}\log_2 2 + \frac{4}{7}\log_2 4 - \frac{3}{7}\log_2 3$$

$$= \frac{8}{7} - \frac{2}{7} = \frac{6}{7} = \underline{0{,}857 \text{ bit}}$$

Damit ergibt sich für das Alter eine Redundanz von

$$R(X_1) = f(X) - f(X|X_1) = \underline{\underline{0{,}006 \text{ bit}}}.$$

Am Ende wird der Test mit der größten Redundanz der statistisch verlässlichste sein.