Informationstheorie SS 2009
Institut für Neuroinformatik
Prof. G. Palm

Grundbegriffe und Notation

1. Wahrscheinlichkeitsraum:

Ein *Wahrscheinlichkeitsraum* ist immer gegeben durch das Tripel (Ω, Σ, p).

Eine Abbildung $X : \Omega \to \mathbb{R}$ heisst *Zufallsvariable* (ZV) falls für alle $r \in \mathbb{R}$ die Aussage $[X \leq r] \in \Sigma$ ist.

Für eine ZV X bedeutet z.B. $[X \leq r] = \{\omega \in \Omega : X(\omega) \leq r\}$, oder
$[X = r] = \{\omega \in \Omega : X(\omega) = r\}$, oder $[X \in A] = \{\omega \in \Omega : X(\omega) \in A\}$.

2. Beschreibungen:

Eine Abbildung $d : \Omega \to \Sigma$ heißt *Beschreibung* falls $\omega \in d(\omega),\ \forall \omega \in \Omega$. Außerdem muss $p(d(\omega)) > 0,\ \forall \omega \in \Omega$ gelten.

Für eine Beschreibung d definieren wir die *Komplettierung* (oder *Vervollständigung*) $\tilde{d}$ durch $\tilde{d}(\omega) := [d = d(\omega)] = \{\omega' \in \Omega : d(\omega') = d(\omega)\}$.

Eine Beschreibung d mit $d = \tilde{d}$ nennen wir *komplett* (oder *vollständig*).

Für eine ZV X definieren wir mit $\tilde{X} : \Omega \to \Sigma; \omega \mapsto [X = X(\omega)]$ die *Beschreibung durch* X.

3. Neuigkeit:

Für eine Aussage A definieren wir die *Neuigkeit* oder *Information* als
$N(A) = I(A) := -\log_2 p(A)$.

Die *Neuigkeit einer Beschreibung* d ist eine ZV $N_d : \Omega \to \mathbb{R}_+; \omega \mapsto -\log_2 p(d(\omega)) \geq 0$. Wir schreiben $N(d) := E(N_d)$.

Die *bedingte Neuigkeit* $N_{c|d}$ einer Beschreibung c gegeben eine weitere Beschreibung d ist eine Zufallsvariable $N_{c|d} : \Omega \to \mathbb{R}_+$ mit $N_{c|d}(\omega) = -\log_2 p(c(\omega)|d(\omega)) \geq 0$. Wir schreiben wieder $N(c|d) := E(N_{c|d})$.

4. Information:

Wir definieren die *Information einer Beschreibung* d als $I(d) := N(\tilde{d})$, und die *Information einer ZV* X als $I(X) := N(\tilde{X})$.

Sätze Informationstheorie

1. Eigenschaften von *Beschreibungen*:

Für Beschreibungen c und d und deren Komplettierungen $\tilde{c}$ und $\tilde{d}$ gilt:

$$\begin{aligned} \tilde{c} &\subseteq c && (1) \\ \tilde{c} \cap \tilde{d} &\subseteq \widetilde{c \cap d} && (2) \end{aligned}$$

Für ZVn X und Y bzw. für *komplette* Beschreibungen c, d gilt:

$$\begin{aligned} \tilde{c} \cap \tilde{d} &= \widetilde{c \cap d} && (3) \\ \tilde{X} \cap \tilde{Y} &= \widetilde{(X, Y)} && (4) \\ \tilde{X} &= \widetilde{X^{\leq}} = \widetilde{X^{\geq}} && (5) \end{aligned}$$

2. Eigenschaften von *Neuigkeit*:

Für Beschreibungen c und d gilt:

- Positivität:

$$N(c) \geq 0 \quad \text{und} \quad \mathrm{N(d|c)} \geq 0 \tag{6}$$

- Monotonie:

$$c \subseteq d \;\Rightarrow\; N(c) \geq N(d) \tag{7}$$

- Additivität:

$$\begin{aligned} N(c \cap d) &= N(c) + N(d|c) && (8) \\ N(c \cap d) &\leq N(c) + N(d) \quad \text{, falls c und d komplett} && (9) \\ N(c \cap d) &= N(c) + N(d) \quad \text{, falls c und d unabhängig} && (10) \end{aligned}$$

3. Eigenschaften von *Information*:

Für Beschreibungen b, c und d und deren Komplettierungen $\tilde{b}$, $\tilde{c}$ und $\tilde{d}$, sowie für ZVn X, Y und Z gilt:

- Bezug zur Neuigkeit:

$$\begin{aligned} I(c) &:= N(\tilde{c}) \;\geq\; N(c) && (11) \\ I(X) &:= N(\tilde{X}) && (12) \\ I(c|d) &:= N(\tilde{c}|\tilde{d}) && (13) \\ I(X|Y) &:= N(\tilde{X}|\tilde{Y}) && (14) \end{aligned}$$

- Monotonie:

$$\begin{aligned} \tilde{c} \subseteq \tilde{d} &\;\Rightarrow\; I(c) \geq I(d) && (15) \\ \tilde{c} \subseteq \tilde{d} &\;\Rightarrow\; I(b|c) \leq I(b|d) \quad \text{und} \quad \mathrm{I(c|b)} \geq \mathrm{I(d|b)} && (16) \\ X \succeq Y &\;\Rightarrow\; I(X) \geq I(Y) && (17) \\ X \succeq Y &\;\Rightarrow\; I(Z|X) \leq I(Z|Y) \quad \text{und} \quad \mathrm{I(X|Z)} \geq \mathrm{I(Y|Z)} && (18) \end{aligned}$$

($X \succeq Y$ bedeutet $\tilde{X} \subseteq \tilde{Y}$.)

- Additivität:

$$
\begin{aligned}
I(c \cap d) &= I(c) + I(d|c) \ \leq I(c) + I(d) && (19)\\
I(c \cap d) &= I(c) + I(d) \quad \text{, falls c und d unabhängig} && (20)\\
I(X,Y) &= I(X) + I(Y|X) \ \leq I(X) + I(Y) && (21)\\
I(X,Y) &= I(X) + I(Y) \quad \text{, falls X und Y unabhängig} && (22)
\end{aligned}
$$

4. *Subjektive Information* **und** *Information Gain*

Für Wahrscheinlichkeitsraum (Ω, Σ, p) nimmt man neben der tatsächlichen Verteilung p die subjektive Verteilung q an: Dann kann man die *subjektive Neuigkeit* einer Beschreibung d als

$$N_q(d) := N_{pq} := E_p(-\log_2(q(d(\omega)))) \tag{23}$$

definieren. Ähnlich kann man für eine Beschreibung d bzw. eine ZV X die *subjektive Information* definieren:

$$
\begin{aligned}
I_q(d) := I_{pq}(d) := E_p(-\log_2(q(\tilde{d}(\omega)))) && (24)\\
I_q(X) := I_{pq}(X) := E_p(-\log_2(q(\tilde{X}(\omega)))) && (25)
\end{aligned}
$$

Der *Information Gain* einer ZV X definiert man als

$$G_{p,q}(X) := I_{pq}(X) - I_{pp}(X) = I_q(X) - I(X) \tag{26}$$

Für $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ und $p_i = p(\omega_i)$ und $q_i = q(\omega_i)$ gilt

$$
\begin{aligned}
G_{p,q}(X) &= G_{p,q} := \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \geq 0, && (27)\\
D_{p,q} &:= G_{p,q} + G_{q,p} \quad (\text{Kullback} - \text{Leibler} - \text{Distanz}). && (28)
\end{aligned}
$$

5. Eigenschaften von *Transinformation*

Es gilt

$$
\begin{aligned}
T(X,Y) &= \begin{cases} I(X) + I(Y) - I(X,Y) & \\ I(X) - I(X|Y) & \leq \quad I(X) \\ I(Y) - I(Y|X) & \leq \quad I(Y) \end{cases} && (29)\\
T(X,Y) &\geq 0 && (30)\\
T(X,Y) = 0 &\Leftrightarrow X \text{ und Y unabhängig.} && (31)
\end{aligned}
$$

Definitionen und Sätze zu Kanälen und Kanalkapazität

1. Einfacher Kanal und Übergangswahrscheinlichkeit $P : \Omega_1 \rightsquigarrow \Omega_2$:

- Definition: Seien $(\Omega_1, \Sigma_1, p_1)$, $(\Omega_2, \Sigma_2, p_2)$ zwei Wahrscheinlichkeitsräume.

  $$P : \Omega_1 \times \Sigma_2 \to [0;1]$$

  heißt *Übergangswahrscheinlichkeit* (Ü.-W.), falls

  $$P_\omega : \begin{cases} \Sigma_2 & \to & [0,1] \\ A & \mapsto & P(\omega, A) \end{cases}$$

  für jedes $\omega \in \Omega_1$ Wahrscheinlichkeit auf (Ω_2, Σ_2) ist, und falls

  $$P_A : \begin{cases} \Omega_1 & \to & [0,1] \\ \omega & \mapsto & P(\omega, A) \end{cases}$$

  für jedes $A \in \Sigma_2$ ZV auf (Ω_1, Σ_1) ist.

  Mit Hilfe einer Ü.-W. $P : \Omega_1 \rightsquigarrow \Omega_2$ und einer Wahrscheinlichkeit p auf Ω_1 kann man eine neue Wahrscheinlichkeit p' auf $\Omega_1 \times \Omega_2$ konstruieren: Für $A \in \Sigma_1$ und $B \in \Sigma_2$ ist

  $$p'(A \times B) \quad := \quad E_{\Omega_1}(1_A \cdot P_B). \tag{32}$$

  Die beiden Randverteilungen von p' sind

  $$p'_1(A) \quad = \quad p'(A \times \Omega_2) = E_{\Omega_1}(1_A \cdot P_{\Omega_2}) = E(1_A) = p(A) \quad \text{für A} \in \Sigma_1 \tag{33}$$
  $$p'_2(B) \quad = \quad p'(\Omega_2 \times A) = E_{\Omega_1}(1_{\Omega_1} \cdot P_B) = E(P_B) \quad \text{für B} \in \Sigma_2 \tag{34}$$

- Ein *einfacher Kanal* K kann mit einer Übergangswahrscheinlichkeit $K : A \rightsquigarrow B$: identifiziert werden. Seine Kapazität ist definiert als

  $$C(K) \quad := \quad \max_X \{T(X;Y) : X : \Omega \to A, K : X \rightsquigarrow Y\}, \tag{35}$$

  wobei X eine ZV mit Werten aus A ist und Y eine ZV mit Werten aus B ist die über K an X gekoppelt ist. Die Maximierung erfolgt über die Verteilung von X.

2. Blockcodierung und Informationsrate

Idee: Kodiere nicht nur ZV X, sondern X mehrfach wiederholt. Seien $(X_i)_{i \in \mathbb{N}_0}$ identisch verteilte (i.d.) diskrete ZVn, d.h. $p[X_i = x] = p[X_0 = x] \forall x \in W(X_0) \forall i \in \mathbb{N}$

- Die *Informationsrate* der *Informationsquelle* $\mathcal{X} = (X_i)_{i \in \mathbb{N}_0}$ ist definiert als

  $$I_r(\mathcal{X}) \quad := \quad \lim_{n \to \infty} \frac{1}{n} I(X_1, X_2, ..., X_n) \tag{36}$$
  $$\quad = \quad \lim_{n \to \infty} I(X_n | X_1, X_2, ..., X_{n-1}) \tag{37}$$

- Die *Transinformationsrate* zweier Prozesse $\mathcal{X} = (X_i)_{i \in \mathbb{N}_0}$ und $\mathcal{Y} = (Y_i)_{i \in \mathbb{N}_0}$ ist definiert als

  $$T_r(\mathcal{X}; \mathcal{Y}) \quad := \quad \lim_{n \to \infty} \frac{1}{n} T((X_1, X_2, ..., X_n); (Y_1, Y_2, ..., Y_n)) \tag{38}$$
  $$\quad = \quad I_r(\mathcal{X}) + I_r(\mathcal{Y}) - I_r(\mathcal{X}, \mathcal{Y}) \tag{39}$$

- Ein n-Block-Code für $(X_i)_{i\in\mathbb{N}_0}$ ist definiert als ein Code C für den ZVe $(X_1,...,X_n)$.
- Die "Länge pro Zeichen" L_r des n-Blockcodes C ist defineiert als $L_r = L_c/n$.
- Satz: Sei $\mathcal{X} = (X_i)_i$ eine Informationsquelle. Für genügend großes n gibt es einen n-Blockcode mit
$$L_r \approx I_r(\mathcal{X}).$$

3. Kanal und Kapazität

- Einfacher Kanal (ohne Gedächtnis) ist gegeben durch eine Übergangswahrscheinlichkeit
$$K : A \rightsquigarrow B$$
- Kanal mit einfachem Gedächtnis gegeben durch
$$K : A \times A \rightsquigarrow B$$
- allgemeiner Kanal (beliebig langes Gedächtnis)
$$\mathcal{K} : A^{\mathbb{N}} \rightsquigarrow B^{\mathbb{N}}$$
Für $\mathcal{X} = (X_t)_{t\in\mathbb{N}}$ Prozess auf A und $\mathcal{Y} = (Y_t)_{t\in\mathbb{N}}$ durch $\mathcal{K}$ an $\mathcal{X}$ gekoppelter Prozess auf B schreibt man $\mathcal{K} : \mathcal{X} \rightsquigarrow \mathcal{Y}$, d.h.
$$p\left(\bigcap_{i=1}^{n}[X_i = a_i \wedge Y_i = b_i]\right) = p\left(\bigcap_{i=1}^{n}[X_i = a_i]\right) \cdot \mathcal{K}_{a_1...a_n}\left(\bigcap_{i=1}^{n}[Y_i = b_i]\right). \quad (40)$$
Für einfaches Gedächtnis gilt
$$\mathcal{K}_{a_0a_1...a_n}\left(\bigcap_{i=1}^{n}[Y_i = b_i]\right) = \prod_{i=1}^{n} K_{a_{i-1}a_i}[Y_i = b_i], \quad (41)$$
und ohne Gedächtnis gilt
$$\mathcal{K}_{a_1...a_n}\left(\bigcap_{i=1}^{n}[Y_i = b_i]\right) = \prod_{i=1}^{n} K_{a_i}[Y_i = b_i], \quad (42)$$
- Kanalkapazität
$$C(\mathcal{K}) := \sup_{\mathcal{X}} \{T_r(\mathcal{X};\mathcal{Y}) : \mathcal{X} \text{ stationär}, \mathcal{K} : \mathcal{X} \rightsquigarrow \mathcal{Y}\}. \quad (43)$$
Satz: Für Kanal ohne Gedächtnis gilt $C(\mathcal{K}) = C(K)$.

4. Satz von Shannon:

Sei $K : A \rightsquigarrow B$ Kanal ohne Gedächtnis mit Kapazität c. Sei $(U_t)_{t\in\mathbb{N}}$ stationärer Prozess auf D mit Informationsrate $r < c$. Sei $\epsilon > 0$. Dann gibt es ein $n \in \mathbb{N}$ und Abbildungen $Q : D^n \to A^n$, $R : B^n \to D^n$, so daß
$$p[R \circ K \circ Q(U_1,...,U_n) \neq (U_1,...,U_n)] < \epsilon$$
Genau heißt das für $\mathcal{K} : (X_t)_{t\in\mathbb{N}} \rightsquigarrow (Y_t)_{t\in\mathbb{N}}$:
$$\sum_{d\in D^n} p[(U_1,...,U_n) = d] \cdot p[R(Y_1,...,Y_n) \neq d|(X_1,...,X_n) = Q(d)] < \epsilon.$$

Etwas mehr über stochastische Prozesse

1. Grundbegriffe:

Eine Familie $(X_t)_{t\in T}$ von ZVen mit $W(X_t) = W$ heißt *stochastischer Prozess* auf W mit Zeit T.
Typischerweise $T = \mathbb{N}, \mathbb{N}_0, \mathbb{Z}, \mathbb{R}, \mathbb{R}_+$ und W endlich oder $\subseteq \mathbb{R}$.

Wie geht man damit um?

- Man macht sich keine Sorgen um die Konstruktion der vielen ZVen (das geht).
- Man rechnet jeweils nur mit der gemeinsamen Verteilung von endlich vielen ZVen.

2. Eigenschaften von stochastischen Prozessen:

1. Identische Verteilung (i.d.):
$$\forall s,t : p[X_t \in A] = p[X_{t+s} \in A] \tag{44}$$
2. Stationarität:
$$\forall s, t_1, ..., t_n : p[X_{t_1} \in A_1, ..., X_{t_n} \in A_n] = p[X_{t_1+s} \in A_1, ..., X_{t_n+s} \in A_n] \tag{45}$$
3. Unabhängigkeit:
$$\forall s, t_1, ..., t_n : p[X_{t_1} \in A_1, ..., X_{t_n} \in A_n] = \prod_{i=1}^{n} p[X_{t_i} \in A_i] \tag{46}$$
4. Markov Eigenschaft:
$$\forall t \geq t_1 \geq t_2 \geq ... \geq t_n : p[X_t \in A|(X_{t_1}, ..., X_{t_n})] = p[X_t \in A|X_{t_1}] \tag{47}$$
5. homogen Markov:
$$\forall t, t_1, s : p[X_t \in A|X_{t_1}] = p[X_{t+s} \in A|X_{t_1+s}] \tag{48}$$

3. Sätze darüber:

- (45) $\Rightarrow$ (44)
- (46) $\Rightarrow$ (47)
- (46) + (44) $\Rightarrow$ (45) + (47) + (48)
- (47) + (45) $\Rightarrow$ (48)
- (47) + (48) + (44) $\Rightarrow$ (45)

4. typische Forderungen:

i.i.d. Prozess, d.h. (44) + (46)
$\Rightarrow$ stationärer Markov-Prozess, (45) + (47)
$\Rightarrow$ stationärer Prozess, (45)
$\Rightarrow$ i.d.-Prozess, (44)

5. Markov-Ketten:

Definition: Eine *Markov-Kette* ist ein homogener Markov-Prozess auf endlicher Menge $W = \{w_1, ..., w_n\}$ mit Zeit $\mathbb{N}_0$.
Die Markov-Kette $(X_t)_{t \in \mathbb{N}_0}$ ist bestimmt durch

- Startverteilung p mit $p_i = p[X_0 = w_i]$ und
- Übergangsmatrix P mit $P_{ij} = p[X_1 = w_i | X_0 = w_j]$

Daraus lassen sich die Verteilungen zu beliebigen Zeiten ableiten:

$$\begin{aligned} p_i^1 := p[X_1 = w_i] &= \sum_j p[X_0 = w_j, X_1 = w_i] = \sum_j p[X_0 = w_j] \cdot p[X_1 = w_i | X_0 = w_j] \\ &= \sum_j P_{ij} p_j = (Pp)_i \end{aligned}$$

Also $p^1 = Pp$, $p^2 = Pp^1 = P^2 p$, $p^3 = P^3 p$, u.s.w.