1 Theorie

1.1 Wahrscheinlichkeitsraum

1.1.1 Definition

Ein Wahrscheinlichkeitsraum ist ein Tripel $(\Omega, \mathcal{A}, P)$ bestehend aus der Grundmenge Ω, einer σ-Algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ und einer Abbildung $P : \mathcal{A} \to [0, 1]$

$$(i)\ P(\Omega) = 1$$
$$(ii)\ P\left(\bigcup_i A_i\right) = \sum_i P(A_i),\ \text{mit } A_i \cap A_j = \emptyset \text{ für } i \neq j$$

Die Elemente von Ω werden elementare Ereignisse und die von $\mathcal{A}$ Ereignisse genannt. Mengen M mit $P(M) = 0$ werden Nullmengen genannt. Die Abbildung P wird Wahrscheinlichkeitsmaß genannt.

1.1.2 σ-Algebra

Es sei Ω eine Menge und $\mathcal{A} \subset \mathcal{P}(\Omega)$ ein System von Teilmengen (= Ereignissen). $\mathcal{A}$ heißt σ-Algebra (Sigma-Algebra) falls gilt:

$$(i)\ \Omega \in \mathcal{A}$$
$$(ii)\ A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$$
$$(iii)\ A_i \in \mathcal{A} \Rightarrow \bigcup_i A_i \in \mathcal{A}$$

$(A^c = \Omega \setminus A)$

1.1.3 Diskreter Wahrscheinlichkeitsraum

Ein diskreter Wahrscheinlichkeitsraum ist ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, bei dem die Grundmenge Ω abzählbar ist und die Menge der Ereignisse $\mathcal{A} := \mathcal{P}(\Omega)$ der Potenzmenge entspricht.

1.1.4 Laplace Experiment

Ein Laplace-Experiment ist ein Zufallsexperiment bei dem der Ereignisraum Ω endlich viele Elemente und ein Ereignis $A \subseteq \Omega$ die Wahrscheinlichkeit $P(A) = \frac{\#A}{\#\Omega}$ hat.

1.2 Bedingte Wahrscheinlichkeit

Für $A, B \in \mathcal{A}$ und $P(B) > 0$ heißt

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

die bedingte Wahrscheinlichkeit (von A unter B).

1.3 Variation und Kombination

- $\#Var_k^n(\Omega, m.W.) = n^k$
- $\#Var_k^n(\Omega, o.W.) = n_k = \frac{n!}{(n-k)!}$
- $\#Kom_k^n(\Omega, o.W.) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
- $\#Kom_k^n(\Omega, m.W.) = \binom{n+k-1}{k}$

1.4 Spamfilter / Satz von Bayes

1.4.1 Satz der totalen Wahrscheinlichkeit

Für eine Zerlegung $\Omega = \bigcup_{j=1}^{n} B_j,$ mit $B_i \cap B_k = \emptyset$ für $i \neq k$, gilt

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j) \cdot P(B_j)$$

1.4.2 Satz von Bayes

Für $A, B \in \mathcal{A}$ mit $P(B) > 0$ gilt

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

1.4.3 Stochastische Unabhängigkeit

Zwei Ereignisse A, B heißen stochastisch unabhängig, falls

$$P(A \cap B) = P(A) \cdot P(B)$$

gilt. Gleichbedeutend damit ist $P(A|B) = P(A)$ und $P(B|A) = P(B)$.

1.4.4 Naiver Bayes'scher Spam Filter

Gegeben ist eine E-Mail E. Wir möchten anhand des Vorkommens bestimmter Wörter $A_1, \ldots A_n$ in der Mail entscheiden, ob es sich um eine erwünschte Mail H oder eine unerwünschte Mail S handelt.

Aus einer Datenbank kann man das Vorkommen dieser Wörter in allen E-Mails zählen und damit empirisch die Wahrscheinlichkeiten $P(A_i|S)$ und $P(A_i|H)$ des Vorkommens dieser Wörter in Spam und Ham Mails ermitteln. Wir gehen davon aus, dass es sich bei der Mail prinzipiell mit Wahrscheinlichkeit $P(E = S) = P(E = H) = \frac{1}{2}$ um eine erwünschte Mail H oder eine unerwünschte Mail S handeln kann.

Wir machen zudem die (naive) Annahme, dass das Vorkommen der Wörter stochastisch unabhängig ist, also

$$P(A_1 \cap \cdots \cap A_n|S) = P(A_1|S) \cdot P(A_2|S) \cdots P(A_n|S)$$
$$P(A_1 \cap \cdots \cap A_n|H) = P(A_1|H) \cdot P(A_2|H) \cdots P(A_n|H)$$

gilt.

Mit der Formel von Bayes und der totalen Wahrscheinlichkeit können wir somit berechnen:

$P(E = S|A_1 \cap \cdots \cap A_n)$ `(-> "E =" wird im folgenden weggelassen)`

$$= \frac{P(A_1 \cap \cdots \cap A_n|S) \cdot P(S)}{P(A_1 \cap \cdots \cap A_n)} \quad \texttt{(-> Satz von Bayes)}$$

$$= \frac{P(A_1|S) \cdots P(A_n|S) \cdot P(S)}{P(A_1 \cap \cdots \cap A_n|H) \cdot P(H) + P(A_1 \cap \cdots \cap A_n|S) \cdot P(S)} \quad \begin{matrix}\texttt{(-> Stoch. Unabhängigkeit)} \\ \hline \texttt{(-> Satz der t. Wahrscheinlichkeit)}\end{matrix}$$

$$= \frac{P(A_1|S) \cdots P(A_n|S)}{P(A_1 \cap \cdots \cap A_n|H) + P(A_1 \cap \cdots \cap A_n|S)} \quad \texttt{(-> kürzen da P(H) = P(S))}$$

$$= \frac{P(A_1|S) \cdots P(A_n|S)}{P(A_1|H) \cdots P(A_n|H) + P(A_1|S) \cdots P(A_n|S)} \quad \begin{matrix}\\ \hline \texttt{(-> Stoch. Unabhängigkeit)}\end{matrix}$$

Bemerkung: $P(E = H|A_1 \cap \cdots \cap A_n) = 1 - P(E = S|A_1 \cap \cdots \cap A_n)$

Schlussfolgerung:
Die Wahrscheinlichkeiten $P(A_i|S)$ und $P(A_i|H)$ sind durch Empirie (Datenbank) bekannt. Daher lässt sich so nun die Wahrscheinlichkeit berechnen, ob es sich bei einer E-Mail, welche die Wörter A_i enthält, um eine Spam E-Mail handelt.

1.5 Zufallsvariablen

1.5.1 Allgemeine Zufallsvariable

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(\Omega', \mathcal{A}')$ ein Messraum. Eine Zufallsvariable ist eine Abbildung

$$X : \Omega \to \Omega'$$

so dass für alle Ereignisse $A' \in \mathcal{A}'$

$$X^{-1}(A') \in \mathcal{A}$$

ein Ereignis in $\mathcal{A}$ ist. Urbilder von Ereignissen sind also Ereignisse.

1.5.2 Messraum

Ein Messraum ist ein Paar $(\Omega, \mathcal{A})$ bestehend aus einer Menge Ω und einer Sigma-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$.

1.5.3 Reelle Zufallsvariable

Unter einer reellen Zufallsvariable verstehen wir eine Zufallsvariable

$$\begin{aligned} X &: \Omega \to \mathbb{R}^n \\ X(\omega) &:= \Big(X_1(\omega), \cdots, X_n(\omega) \Big), \end{aligned}$$

wobei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum ist und $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ der $\mathbb{R}^n$ zusammen mit der Borel'schen Sigma-Algebra ist.

1.6 Erwartungswert

1.6.1 Definition

Für eine reelle, integrierbare Zufallsvariable X ist der Erwartungswert definiert durch

$$\mathbb{E}(X) := \int_{\Omega} X \, dP \, .$$

Ist $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine eindimensionale reelle Zufallsvariable, so ist

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega)$$

1.6.2 Eigenschaften

Sind $X, Y : \Omega \to \mathbb{R}^n$ reelle, integrierbare Zufallsvariablen und $a, b \in \mathbb{R}$ konstant, so gilt:

$$\begin{aligned}
&\mathbb{E}(a \cdot X \pm b \cdot Y) = a \cdot \mathbb{E}(X) \pm b \cdot \mathbb{E}(Y) \\
&\forall x \in \Omega : X(x) \leq Y(x) \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y) \\
&X, Y \text{ stoch. unabhängig} \Rightarrow \mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y) \\
&\mathbb{E}(1_A) = P(A)
\end{aligned}$$

1.7 Varianz

Für eine reelle Zufallsvariable X ist die Varianz definiert durch

$$\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

1.8 Verteilungen

1.8.1 Normalverteilung

Die Normalverteilung $N(\mu, \sigma^2)$ auf $\mathbb{R}$ ist definiert durch

$$\text{Dichte: } f(x) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$\text{Verteilung: } F(x) = N(\mu, \sigma^2)(-\infty, x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt$$

Erwartungswert und Varianz bei $X \sim N(\mu, \sigma^2)$:

$$\mathbb{E}(X) = \mu$$
$$\mathbb{V}(X) = \sigma^2$$

1.8.2 Verteilungsfunktion

Für eine reelle Zufallsvariable X heißt

$$F_X : \Omega \to [0, 1]$$
$$F_X(x) := P(X \leq x) := P_X((-\infty, x)) = P(X^{-1}(-\infty, x))$$

Verteilungsfunktion von X.

1.8.3 Gleichverteilung

Die Gleichverteilung $U(a, b)$ auf einem Intervall $(a, b) \subset \mathbb{R}$ ist definiert durch

$$\text{Dichte: } f(x) := \frac{1_{(a,b)}}{|b-a|}$$

$$\text{Verteilung: } F(x) = P_f((-\infty, x)) = \int_{-\infty}^{x} \frac{1_{(a,b)}}{|b-a|} dt$$

$$= \begin{cases} 0 \text{ für } x \leq a \\ \frac{x-a}{|b-a|} \text{ für } a \leq x \leq b \\ 1 \text{ für } x \geq b \end{cases}$$

Erwartungswert und Varianz bei $X \sim U(a, b)$:

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{3}\frac{b^3-a^3}{b-a} - \left(\frac{a+b}{2}\right)^2$$

$$= \frac{1}{12}(b-a)^2$$

1.8.4 Dichte

Sei $\Omega \subset \mathbb{R}^n$ und $(\Omega, \mathcal{A})$ ein Messraum, wobei alle $A \in \mathcal{A}$ Lebesgue-messbar sind. Eine Funktion $f : \Omega \to \mathbb{R}$ heißt Dichte, falls für ihr Lebesgue-Integral $\int_\Omega f d\mu = 1$ gilt.

1.9 Schwaches Gesetz der großen Zahlen

1.9.1 Definition

Seien $X_i : \Omega \to \mathbb{R}$ unabhängige, reelle Zufallsvariablen mit $\mathbb{E}(X_i) = \mu < \infty$ und $\mathbb{V}(X_i) = \sigma < \infty$, dann gilt

$$P\big(\big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\big| \geq \epsilon\big) \leq \frac{\sigma}{n \cdot \epsilon^2} \xrightarrow[n\to\infty]{} 0$$

(stochastische Konvergenz).

1.9.2 Bedeutung

Das schwache Gesetz der großen Zahlen besagt, dass das arithmetische Mittel einer großen Stichprobe einer Zufallsvariable mit einer beliebig kleinen Wahrscheinlichkeit dem Erwartungswert der Zufallsvariable entspricht.

Gegenteilige (äquivalente) Formulierung:
Die Wahrscheinlichkeit, dass die Differenz zwischen beobachteter relativer Häufigkeit und theoretischer Wahrscheinlichkeit kleiner ist als eine beliebig kleine positive Zahl ϵ, geht für eine unendlich große Stichprobe gegen 1.

1.10 Zentraler Grenzwertsatz

1.10.1 Definition

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X_n : \Omega \to \mathbb{R}$ eine folge stochastisch unabhängiger, identisch verteilter, reeller Zufallsvariablen mit $\mathbb{E}(X_n) = \mu$ und $\mathbb{V}(X_n) = \sigma^2$. Dann gilt für das arithmetische Mittel $S_n := \frac{1}{n}\sum_{i=1}^{n} X_i$

$$P_{\frac{\sqrt{n}}{\sigma}(S_n - \mu)} \to P_{N(0,1)}$$

wobei $P_{N(0,1)}$ das Wahrscheinlichkeits-Maß mit der Dichte $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ ist.

1.10.2 Bedeutung

Die Summe von n identisch verteilten, stochastisch unabhängigen Zufallsvariablen ist näherungsweise normalverteilt.

Beispiel Würfel:
Die Augensumme von $n \to \infty$ Würfeln ist normalverteilt, wenn alle Würfel von einander stochastisch unabhängig und gleichverteilt sind.

1.11 Schätzer

1.11.1 Ausgangslage

Angenommen man findet einen Apparat, der zufällig Zahlen in einem Intervall $[0, \rho]$ ausgibt. Anhand von Beobachtungen der Zahlen möchte man ρ schätzen. Wir machen die Annahme, dass alle Zahlen in dem Intervall gleich wahrscheinlich auftreten und nehmen n Stichproben $X_1, \cdots, X_n$. Einen Schätzer für ρ bezeichnen wir mit T_n.

1.11.2 Maximalwert-Schätzer

Eine einfache und einleuchtende Idee ist es, ρ durch die größte beobachtete Zahl zu schätzen, also $T_n^{max} := \max(X_1, \cdots, X_n)$.
Dieser Schätzer konvergiert für $n \to \infty$ gegen ρ, also

$$P(|T_n^{max} - \rho| \geq \epsilon) \underset{n\to\infty}{\longrightarrow} 0$$

Der Erwartungswert dieses Schätzers ist

$$\mathbb{E}(T_n^{max}) = \frac{n}{n+1}\rho \underset{n\to\infty}{\longrightarrow} \rho$$

1.11.3 Erwartungswert-Schätzer

Da das Auftreten der Zahlen gleich wahrscheinlich ist, ist der Erwartungswert des Zufallsexperiments $\rho/2$. Unter Berufung auf das schwache Gesetz der großen Zahlen erscheint der Schätzer $T_n^E := 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ sehr plausibel.
Dieser Schätzer konvergiert für $n \to \infty$ gegen ρ, also

$$P(|T_n^E - \rho| \geq \epsilon) \underset{n\to\infty}{\longrightarrow} 0$$

Der Erwartungswert dieses Schätzers ist

$$\mathbb{E}(T_n^E) = \rho$$

1.11.4 Bedeutung

Beide Schätzer sind asymptotisch exakt und konvergieren gegen ρ (= Erwartungstreu). Der Erwartungswert-Schätzer ist aber durchschnittlich (Erwartungswert) früher exakt ρ als der Maximalwert-Schätzer mit seinem Bias $\frac{n}{n+1}$. Die Varianz von beiden Schätzern geht für $n \to \infty$ gegen 0, beide werden also immer aussagekräftiger mit zunehmender Stichprobenanzahl.
Kein Schätzer ist allgemein perfekt, den aktuell passenden Schätzer kann man nach Kriterien wie Konvergenzgeschwindigkeit zum gesuchten Wert, Konvergenzgeschwindigkeit des Erwartungswertes und der Konvergenzgeschwindigkeit der Varianz auswählen.

1.12 Informationstheorie / Entropie

1.12.1 Ausgangslage

Gegeben eine reelle, diskrete Zufallsvariable $X : \mathcal{X} \to \mathbb{R}$ mit endlichem Grundraum $\#\mathcal{X} \geq 2$ und Verteilung Q.
Mit $L : \mathcal{X} \to \mathbb{N}$ bezeichnen wir die Anzahl an Fragen $L(x)$, die benötigt werden, um bei einer gewählten Strategie den Wert von $X = x$ zu erraten.
Wir suchen eine Strategie, so dass die mittlere Anzahl an Fragen (Erwartungswert) $\mathbb{E}(L(X)) := \sum_{x \in \mathcal{X}} L(x)Q(x)$ möglichst klein ist.

1.12.2 Wörter und Wortmenge

Gegeben sei eine endliche Menge $\mathcal{A}$ mit $\#A \geq 2$, genannt Alphabet. Ein Wort der Länge k ist gegeben durch ein Tupel $w = b_1 b_2 \cdots b_k$ mit Buchstaben $b_i \in \mathcal{A}$.
Die Menge aller Wörter bezeichnen wir mit

$$\mathcal{W}(\mathcal{A}) := \{b_1 \cdots b_k \mid i \in \mathbb{N},\ b_i \in \mathcal{A}\}$$

Mit $l(w) := k$ für $w = b_1 \cdots b_k$ bezeichnen wir die Länge des Wortes.

1.12.3 Kode und Präfix

Ein $\mathcal{A}$-Kode für die Menge $\mathcal{X}$ mit Alphabet $\mathcal{A}$ ist eine injektive Abbildung (1-zu-1)

$$\kappa : \mathcal{X} \to \mathcal{W}(\mathcal{A})$$

die jedem Wort $x \in \mathcal{X}$ eindeutig ein Kodewort $\kappa(x)$ zuordnet.

Ein Wort $v = a_1 \cdots a_j$ heißt Präfix des Wortes $w = b_1 \cdots b_k$, wenn $j \leq k$ und $v = b_1 \cdots b_j$ ist. Das Wort w heisst Fortsetzung von v. Ein Kode

$$\kappa : \mathcal{X} \to \mathcal{W}(\mathcal{A})$$

heißt präfixfrei, wenn kein Kodewort $\kappa(x)$ Präfix eines anderen Kodewortes $\kappa(y)$ ist.

1.12.4 Zusammenhang mit der Fragestrategie

Jede Fragestrategie liefert einen präfixfreien Kode mit Alphabet $\mathcal{A} := \{\mathrm{j}, \mathrm{n}\}$.
Das Maß für Informationsgehalt (= Entropie) der Quelle X ist das Minimum von $\mathbb{E}(l(\kappa(X)))$ über alle präfixfreien Kodes κ für X.

1.12.5 Kraftsche Ungleichung

Sei κ ein präfixfreier $\mathcal{A}$-Kode für $\mathcal{X}$ mit $\#\mathcal{A} = d$. Dann ist

$$\sum_{x \in \mathcal{X}} d^{-l(\kappa(x))} \leq 1$$

Für $x \in \mathcal{X}$ sei $L : \mathcal{X} \to \mathbb{N}$ eine Abbildung, so dass $\sum_{x \in \mathcal{X}} d^{-L(x)} \leq 1$ gilt. Dann gibt es einen präfixfreien $\mathcal{A}$-Kode für $\mathcal{X}$ mit

$$l(\kappa(x)) = L(x)\,.$$

1.12.6 Entropie

Die Entropie $H_d(X)$ der Ordnung d der Zufallsvariable X ist definiert als das Minimum von $\sum_{x\in\mathcal{X}} L(x)Q(x)$ über alle Abbildungen $L : \mathcal{X} \to \mathbb{N}$ mit $\sum_{x\in\mathcal{X}} d^{-L(x)} \leq 1$.

Die Entropie der Zufallsvariable X ist definiert durch

$$H(X) := -\sum_{x\in\mathcal{X}} Q(x)\log(Q(x))$$

Es gilt

$$\frac{H(X)}{\log(d)} \leq H_d(X) \leq \frac{H(X)}{\log(d)} + 1\,.$$

Berechne für $x \in \mathcal{X}$ die Funktion $L(x) := \lceil -\log_d(Q(x))\rceil$. Damit existiert ein präfixfreier $\mathcal{A}$-Kode κ mit $l(\kappa(x)) = L(x)$. Dann gilt $\mathbb{E}(l(\kappa(X))) < H_d(X) + 1$.

1.12.7 Bedeutung

Eine Fragestrategie ist äquivalent zu einem präfixfreien Kode. Die durchschnittliche Anzahl an notwendigen Fragen zur Bestimmung des Ergebnisses einer Zufallsvariable X ist somit ebenfalls äquivalent zur durchschnittlichen Wortlänge des zugehörigen präfixfreien Kodes.
Die Entropie ist ein Maß für den Informationsgehalt einer Zufallsvariable X. Je häufiger ein Ergebnis auftritt, desto geringer ist sein Informationswert.
Die Entropie beschreibt die durchschnittliche Wortlänge eines optimalen, präfixfreien Kodes für X. Dabei ist ein Kode genau dann optimal, wenn die durchschnittliche Wortlänge eines Kodewortes, im Vergleich mit allen möglichen präfixfreien Kodes für X, minimal ist.
Die minimale durchschnittliche Kodewortlänge (Entropie) hängt auch von den zur Verfügung stehenden Anzahl an Buchstaben d ab (mehr Antwortmöglichkeiten $\rightarrow$ weniger Fragen notwendig).
Das die minimale durchschnittliche Kodewortlänge tatsächlich der Entropie $H_d(X)$ entspricht, wird durch die kraftsche Ungleichung impliziert.
Die Entropie $H_d(X)$ kann zwar nicht direkt berechnet, aber durch die Entropie-Ungleichung unter Verwendung von $H(X)$ zwischen zwei natürliche Zahlen eingeschätzt werden.