Wissenserzeugung nach dem Prinzip der maximalen Entropie

Hans Braun, Reiner Marchthaler
Hochschule Esslingen

7. Februar 2023

Zusammenfassung

Das Prinzip der maximalen Entropie ist ein Verfahren der künstlichen Intelligenz, mit dem fehlendes stochastisches Wissen generiert werden kann. Dadurch ist die Methode für alle Aufgabenstellungen anwendbar, in denen temporär oder dauerhaft nur unvollständiges Wissen vorliegt. Das Prinzip fügt zu vorhandenem lückenhaften Wissen so viel, wie möglich Unsicherheit hinzu und minimiert dadurch nicht gerechtfertigte sichere Annahmen. Anhand eines einfachen Beispiels mit zwei booleschen Zufallsvariablen wird die Überwachungseinrichtung einer Produktionsanlage modelliert. Dabei liegt über die Güte der Überwachung nur unvollständiges Wissen vor. Aus den gegebenen Informationen werden nun die Berechnungsschritte hin bis zu einer vollständigen Wahrscheinlichkeitsverteilung demonstriert. Diese Verteilung repräsentiert das vollständige Wissen aller Zusammenhänge des Modells. Die so gewonnene Wahrscheinlichkeitsverteilung wird abschließend zur Bewertung der Güte der Überwachungsanlage genutzt und ermöglicht dabei statistische Aussagen, welche mit dem ursprünglich gegebenen Wissen nicht möglich waren.

1 Einleitung

Das Prinzip der maximalen Entropie ist eine Methode zur Ermittlung unbekannter Wahrscheinlichkeiten. Damit kann aus stochastischen Modellen über Zusammenhänge zwischen Zufallsvariablen, deren Parameter zunächst nur teilweise bekannt sind ein vollständiges Modell erzeugt werden. In dem vorliegenden Artikel werden die Grundlagen der Methode kurz beschrieben und deren Anwendung anhand eines einfachen Beispiels erklärt.

Zusammenhänge zwischen Zufallsvariablen können durch ihre Wahrscheinlichkeitsverteilung beschrieben werden. Dabei enthält diese Verteilung Angaben über die Wahrscheinlichkeiten aller möglichen Kombinationen aller Zufallsvariablen. Häufig sind allerdings nicht alle Quantitäten dieser Kombinationen bekannt. Die Methode der maximalen Entropie ermittelt dann das

fehlende Wissen über einzelne Zusammenhänge derart, dass der Wahrscheinlichkeitsverteilung so wenig, wie möglich vermeintliches Wissen hinzugefügt wird. Dadurch entsteht eine vollständige Wahrscheinlichkeitsverteilung, welche beliebige Aussagen über die einzelnen Zufallsvariablen und deren Zusammenhänge ermöglicht.

Das Prinzip der maximalen Entropie kommt in unterschiedlichsten Domänen zur Anwendung, wie z. B. Anfragebewertung im Anlagengeschäft, technische Diagnosen, medizinische Diagnosen [1], Mustererkennung, Information Retrieval oder der Sicherheitsbewertung von Fahrsituationen im Straßenverkehr [2].

Der vorliegende Beitrag, gibt eine Einführung in das Prinzip der maximalen Entropie, um dem Leser ein Basiswissen für den Einsatz der Methode zu vermitteln. Zunächst werden die Grundlagen des Prinzips vorgestellt. Zusätzlich wird der Lagrange-Ansatz als ein Verfahren zur Berechnung der maximalen Entropie präsentiert. Zur Vertiefung dieser beiden Themen finden sich in dem Abschnitt eine Reihe von Literaturhinweisen. Im darauffolgenden Abschnitt werden die eingeführten Grundlagen auf ein Einfachstmodell angewendet und in einem weiteren Abschnitt in einem Bεrechnungsbeispiel umgesetzt.

2 Das Prinzip der maximalen Entropie

Grundlage des Prinzips der maximalen Entropie ist die Informationstheorie, welche wesentlich auf Arbeiten von C laude E. Shannon [3] zurückgeht. Die Informationstheorie umfasst die Quantifizierung, Übertragung und Speicherung von Informationen. Die Inhalte dieses Abschnitts sind zum überwiegenden Teil Auszüge aus einer Dissertationsschrift [2]. Dort finden sich vertiefende Informationen und weitere Quellenangaben.

Ein einfaches Modell der Nachrichtenübertragung besteht aus einem Sender (Quelle), einer Übertragungsstrecke und einem Empfänger (Senke) [4]. Dabei ist die Bedeutung einer Nachricht für den Empfänger von seinem Informationsgewinn gegenüber seinem bisherigen Wissensstand abhängig. So erhöht beispielsweise eine dem Empfänger bereits bekannte Information das Wissen des Empfängers nicht. Shannons Informationstheorie stellt ein quantitatives Maß zu dieser qualitativen Aussage bereit und verknüpft den Informationsgehalt einer Nachricht mit dem Kehrwert seiner Auftrittswahrscheinlichkeit. Der Informationsgehalt einer Nachricht kann anschaulich auch als ein Überraschungs- oder Unsicherheitsmaß [4] interpretiert werden, weswegen Nachrichten auch als Zufallsvariablen betrachtet werden können.

Definition

Gegeben seien eine diskrete Zufallsvariable $V = \{v_1, v_2, \dots, v_K\}$ mit den zugehörigen Wahrscheinlichkeiten $P = \{P(v_1), P(v_2), \dots, P(v_K)\}$[1]. Dann beträgt der Informationsgehalt $\boldsymbol{\mathcal{I}}(v_k)$ eines Zeichens v_k

$$\boldsymbol{\mathcal{I}}(v_k) = log_2\left(\frac{1}{P(v_k)}\right)[Bits], \qquad k = 1, \dots, K, \forall v_k \in V. \tag{2.1}$$

■

[1] $P(v_k)$ steht als Kurzschreibweise für $P(V = v_k)$.

Glg. (2.1) besagt u. a., dass der Informationsgehalt eines sicheren Ereignisses mit $P(v_k) = 1$ gleich Null ist, da der Wissensstand des Empfängers hierdurch nicht erhöht wird. Des Weiteren erkennt man, dass der Informationsgehalt einer sehr unwahrscheinlichen Nachricht, wie z. B. „ich habe eine Sechs im Lotto“ sehr hoch ist.

Aufbauend auf obiger Definition des Informationsgehalts eines Elementarereignisses v_k einer Zufallsvariablen V kann der gewichtete Informationsgehalt aller Elementarereignisse dieser Zufallsvariablen ermittelt werden und wird als deren Entropie $\mathcal{H}(P)$ bezeichnet.

Definition

Gegeben seien eine diskrete Zufallsvariable $V = \{v_1, v_2, \dots, v_K\}$ mit den zugehörigen Wahrscheinlichkeiten $P = \{P(v_1), P(v_2), \dots, P(v_K)\}$ und mit $\boldsymbol{\mathcal{I}}(v_k)$ deren jeweiliger Informationsgehalt. Dann beträgt die Entropie $\mathcal{H}$ der Verteilung P [5]:

$$\mathcal{H}(P) = \sum_k [P(v_k) \cdot \boldsymbol{\mathcal{I}}(v_k)], \qquad k = 1, \dots, K. \tag{2.2}$$

■

Die Entropie $\mathcal{H}(P)$ ist ein Maß für die Unsicherheit der Wahrscheinllichkeitsverteilung einer oder mehrerer Zufallsvariablen[2]. Sind alle möglichen Elementarereignisse v_k gleich wahrscheinlich, so spricht man von einer Gleichverteilung. Diese besitzt unter allen möglichen Verteilungen die größtmögliche Entropie, da hierbei die höchste Unsicherheit über den Ausgang eines Experiments besteht. Besitzt ein Elementarereignis v_k die Wahrscheinlichkeit Eins, so gilt $\mathcal{H}(P) = 0$, da es keinerlei Unsicherheit über den Ausgang eines Experiments gibt.

Ein bekanntes Beispiel der Entropiefunktion ist die sogenannte Shannon-Funktion gemäß Abbildung 1, welche die Entropie $H(P)$ einer binären Zufallsvariablen $V = \{v_1, v_2\}$ mit $P = \{P(v_1), 1 - P(v_1)\}$ beschreibt. Man erkennt, dass bei sicherem $P(v_1) = 0$ oder $P(v_2) = 1 - P(v_1) = 0$ die Entropie H gleich Null ist, da in diesen Fällen keine Unsicherheit herrscht. Maximale Entropie liegt bei $P(v_1) = P(v_2) = 0{,}5$ vor, da hier beide möglichen Ereignisse v_1 und v_2 gleich wahrscheinlich sind. Der Maximalwert $\mathcal{H}(P) = 1$ ergibt sich hierbei aus der Wahl des Logarithmus Dualis log_2.

Definition

Die Menge aller möglichen Kombinationen der möglichen Ereignisse (oder auch Ausprägungen genannt) aller I Zufallsvariablen $\mathrm{V} = \{V_1, V_2, \dots . V_I\}$ wird als deren Ereignisraum $\mathcal{P}(V)$ bezeichnet.

Ordnet man jedem der Z Elemente von $\mathcal{P}(V)$ eine Wahrscheinlichkeit $P_z,\ z \in 1, \dots, Z$ zu, so entsteht der zugehörige Wahrscheinlichkeitsraum $(P, \mathcal{P}, V)$.

■

Der Wahrscheinlichkeitsraum $(P, \mathcal{P}, V)$ enthält die Wahrscheinlichkeitsverteilung P und beschreibt die stochastischen Zusammenhänge aller Zufallsvariablen. Häufig sind allerdings nicht alle Elemente von P bekannt und bei großen Wahrscheinlichkeitsräumen ist die experimentelle

[2] Der Begriff Entropie wird auch in der Thermodynamik verwendet. Dabei ist die Entropie S ein Maß für die Irreversibilität eines Prozesses, wobei in der Regel nicht die Entropie S selbst, sondern deren Änderung ΔS betrachtet wird.

Bestimmung einer Vielzahl fehlender Elemente nahezu unmöglich. Aus mathematischer Sicht besteht dann eine Situation, in der nicht genügend Bestimmungsgleichungen zur Verfügung stehen und daher die Lösungsmenge größer Eins ist. Gesucht ist somit eine Strategie, den bestmöglichen Repräsentanten P^* unter einer Vielzahl zulässiger Wahrscheinlichkeirsverteilungen P zu finden.

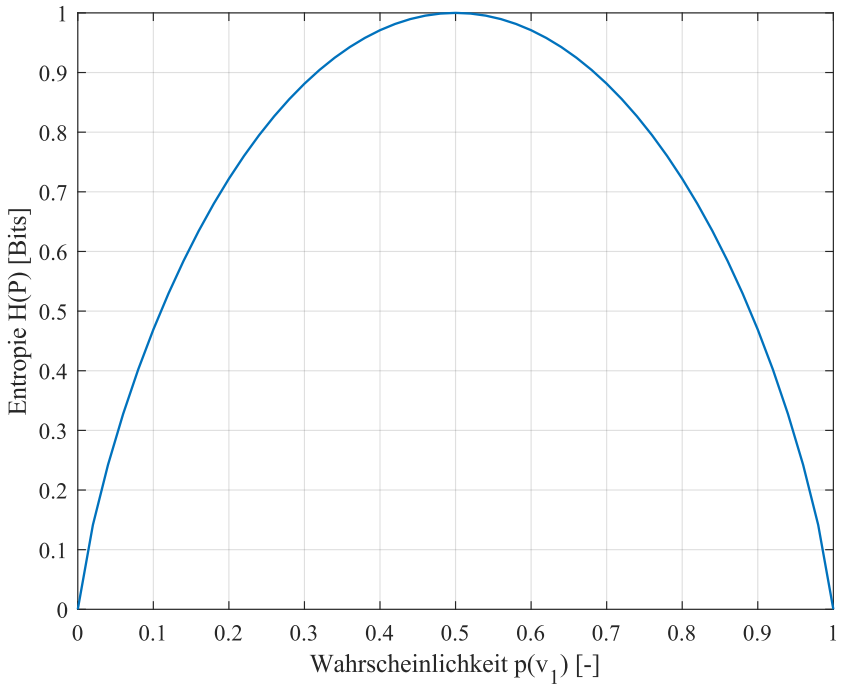

Abbildung 1: Shannon-Funktion einer binären Zufallsvariablen V.

Zulässige Lösungen des Problems sind alle Verteilungen P, welche die verfügbaren Nebenbedingungen, also das vorhandene Wissen sowie die Normierungsbedingung[3] erfüllen. Sei R die Menge aller gegebenen Nebenbedingungen und P eine Wahrscheinlichkeitsverteilung, welche alle R erfüllt, so nennt man P den epistemischen Zustand [6] von R und schreibt $P \vDash R$.

Der amerikanische Physiker Edwin Thompson Jaynes formulierte in „Probability Theory: The Logic of Science“ [7] ein Postulat für die Eigenschaften des bestmöglichen Repräsentanten P^* aus einer Menge P von Verteilungen, welche die Nebenbedingungen R erfüllen. Jaynes forderte, dass eine zunächst unbekannte Wahrscheinlichkeitsverteilung P so zu wählen sei, dass diese maximale Entropie besitzt. Dadurch ist sichergestellt, dass P^* größtmögliche Unsicherheit repräsentiert, was im Umkehrschluss bedeutet, dass so wenig, wie möglich vermeintlich sicheres Wissen hinzugefügt wird. Dieser Grundsatz wird das Prinzip der maximalen Entropie genannt. Mit diesem Postulat transformierte Jaynes die Erzeugung fehlenden Wissens in ein Optimierungsproblem.

Definition

Gegeben seien ein Ereignisraum $\mathcal{P}(V)$ mit zugehörigem Wahrscheinlichkeitsraum $(P, \mathcal{P}, V)$ sowie eine Menge $\mathcal{R}$ von J Nebenbedingungen in Form eines linearen Gleichungssystems.

[3] Die Normierungsbedingung besagt, dass die Summe aller Wahrscheinlichkeiten einer Verteilung gleich Eins ist.

Dann ist die gesuchte optimale Verteilung $P^* \in P$ so zu wählen, dass die Entropie $\mathcal{H}(P)$ maximal ist und P^* alle Nebenbedingungen $\mathcal{R}$ erfüllt:

$$P^* = \arg\max_P\big(\mathcal{H}(P)\big), \qquad P \vDash \mathcal{R},\ P^* \in P. \tag{2.3}$$

■

Nebenbedingungen als Wissen über Zusammenhänge zwischen Zufallsvariablen sind häufig in der Form von bedingten Wahrscheinlichkeiten (auch Konditionale genannt) gegeben. Dabei wird die Wahrscheinlichkeit x_j eines Ereignisses B_j unter der Bedingung, dass ein Ereignis A_j vorliegt formal wie folgt beschrieben:

$$\mathcal{R}_j = \big\{P\big(B_j|\ A_j\big) = x_j\big\}, \qquad 0 \le x_j \le 1 \tag{2.4}$$

Zur Berechnung der Wahrscheinlichkeitsverteilung P^* mit maximaler Entropie müssen die Nebenbedingungen $\mathcal{R} = \{\mathcal{R}_j\}$ als algebraische Gleichungen vorliegen. Zur Wandlung von Konditionalen der Form (2.4) in algebraische Gleichungen dient folgende Formel[4]:

$$P\big(B_j, A_j\big) \cdot \big(1 - x_j\big) - P\big(\bar{B}_j, A_j\big) \cdot x_j = 0, \qquad 0 \le x_j \le 1 \tag{2.5}$$

Die Maximierung der Entropie $\mathcal{H}$ einer Verteilung P stellt ein Optimierungsproblem dar. Die zu optimierende Zielfunktion $\mathcal{H}(P)$ ist wegen des enthaltenen Logarithmus Dualis nichtlinear und mit den Regeln R liegen typischerweise auch Nebenbedingungen vor. Man spricht dann von einem nichtlinearen Optimierungsproblem mit Nebenbedingungen. Da die Entropiefunktion $\mathcal{H}(P)$, wie exemplarisch in Abbildung 1 zu erkennen konkav ist [8], stellt das Optimum der Funktion stets ein globales Optimum dar, was die Ermittlung von P^* erheblich vereinfacht.

Eine mögliche Hilfsfunktion ist die sogenannte Lagrange-Funktion $\mathcal{L}$ [9], bei der jede der J Nebenbedingungen[5] mit einem Faktor λ_j multipliziert und zur Zielfunktion addiert wird. Formal wird diese Funktion beschrieben durch

$$\mathcal{L} = -\sum_z P_z \cdot log_2(P_z) + \sum_j \lambda_j \cdot R_j, \qquad z = 1, \dots, Z, j = 1, \dots, J. \tag{2.6}$$

$\mathcal{L}$ ist eine Funktion in den Veränderlichen P_z und λ_j, deren analytische Optimierung analog zur Extremwertbestimmung von Funktionen in einer Veränderlichen erfolgen kann. Dabei wird anstatt der ersten Ableitung der Funktion deren Gradient gebildet und die Untersuchung der zweiten Ableitung durch die Bestimmung der Definitheit der Hesse-Matrix ersetzt. Bei Bedarf können diese elementaren Methoden in Grundlagenwerken, wie z. B. [10] repetiert werden. Die Bestimmungsgleichungen des Gradienten von $\mathcal{L}$ lauten:

$$\nabla_{P_z,\lambda_j}\mathcal{L} = \underline{0}. \tag{2.7}$$

Bzw.

[4] Die Herleitung der Formel findet sich neben weiteren vertiefenden Informationen zum Prinzip der maximalen Entropie in [2].

[5] Die Nebenbedingungen müssen hierbei als Nullstellen von Funktionen definiert sein. Dadurch dürfen sie jederzeit zu einer Zielfunktion addiert werden.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial P_1} &= 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda_1} = 0 \\ &\vdots \qquad\qquad \vdots \\ \frac{\partial \mathcal{L}}{\partial P_Z} &= 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda_J} = 0. \end{aligned} \tag{2.8}$$

Glg. (2.8) beschreibt ein System nichtlinearer algebraischer Gleichungen, welches nur in einfachen Fällen (kleiner Ereignisraum, wenige Nebenbedingungen) analytisch lösbar ist.

In den folgenden Abschnitten wird ein sehr einfaches Beispiel zum Prinzip der maximalen Entropie zunächst in symbolischer Form und danach als Zahlenbeispiel präsentiert.

3 Ein Einfachstmodell

Das Einfachstmodell besteht lediglich aus zwei booleschen Zufallsvariablen V_1 und V_2 und dient der Beschreibung des Rechenwegs der Methode der maximalen Entropie.

Grundlage des Modells ist der Modus Ponens, welcher eine Inferenzregel der klassischen Aussagenlogik darstellt [1]. Dabei wird von einer Aussage V_1 und einer Regel $V_1 \rightarrow V_2$ auf eine Aussage V_2 geschlossen.

Beispiel[6]

Bei der Getränkeproduktion wird die Befüllung von Flaschen automatisch überwacht. Dabei werden defekte Flaschen aussortiert.

Aussage V_1: „Flasche ist defekt“

Regel: $V_1 \rightarrow V_2$: „Wenn die Flasche defekt ist, dann wird sie aussortiert“

Schluss V_2: „Die Flasche wird aussortiert“

■

Die Übertragung dieser Inferenzregel in die Wahrscheinlichkeitsrechnung führt zur Wahrscheinlichkeit $P(V_1) = p_A$ der Aussage V_1 sowie zu $P(V_2|V_1) = p_R$, welche die Wahrscheinlichkeit der Regel $V_1 \rightarrow V_2$ repräsentiert. Gesucht ist dann die Wahrscheinlichkeit $P(V_2)$ der Aussage V_2. Die Berechnung von $P(V_2)$ mit der Methode der maximalen Entropie ist Gegenstand des Einfachstbeispiels.

3.1 Berechnungsmethodik

Die beiden Zufallsvariablen V_1 und V_2 können bei gleichzeitiger Betrachtung folgende vier Zustände einnehmen und bilden zusammen den Ereignisraum $\mathcal{P}(V)$:

[6] Dieses Beispiel wird im Kapitel 4 ausführlich diskutiert.

$$\mathcal{P}(V) = \{(V_1, V_2), (\bar{V}_1, V_2), (V_1, \bar{V}_2), (\bar{V}_1, \bar{V}_2)\}. \tag{3.1}$$

Ordnet man jedem Element von $\mathcal{P}(V)$ eine Wahrscheinlichkeit $P(\mathcal{P}(V))$ zu, so wird diese Zuordnung Wahrscheinlichkeitsfunktion oder W-Funktion genannt. Die W-Funktion bildet somit den Ereignisraum $\mathcal{P}(V)$ in den Wahrscheinlichkeitsraum $(P, \mathcal{P}, V)$ ab, welcher die Wahrscheinlichkeiten aller möglichen Zustände der Produktmenge von V_1 und V_2 beinhaltet. Dabei gilt:

$$P: \mathcal{P}(V) \rightarrow [0,1]. \tag{3.2}$$

Die Wahrscheinlichkeitsverteilung P des Einfachstmodells besteht aus folgenden vier Elementen P_z:

$$\begin{aligned} P_1 &= P(V1, V2) \\ P_2 &= P(\overline{V1}, V2) \\ P_3 &= P(V1, \overline{V2}) \\ P_4 &= P(\overline{V1}, \overline{V2}) \end{aligned} \tag{3.3}$$

3.1.1 Vorhandenes Wissen

Das vorhandene Wissen aus dem probabilistischen Modus Ponens muss als System algebraischer Gleichungen der Variablen P_z gemäß ((2.3) formuliert werden.

Dabei gelten die Aussage V_1 mit der Wahrscheinlichkeit $P(V_1) = p_A$ und die Regel $V_1 \rightarrow V_2$ mit der Wahrscheinlichkeit $P(V_2|V_1) = p_R$:

$$P(V_1) = P(V_1, V_2) + P(V_1, \bar{V}_2) = P_1 + P_2 = p_A, \qquad 0 \leq p_A \leq 1 \tag{3.4}$$

$$P(V_1 \rightarrow V_2) = P(V_2|V_1) = p_R, \qquad 0 \leq p_R \leq 1 \tag{3.5}$$

Die Wandlung einer bedingten Wahrscheinlichkeit $P(V_2|V_1) = p_R$ in Glg. (3.5) erfolgt nach folgender Formel [2]:

$$P(V_2|V_1) = p_R \leftrightarrow P(V_2, V_1) \cdot (1 - p_R) - P(\bar{V}_2, V_1) \cdot p_R = 0 \tag{3.6}$$

Zusätzlich zu diesen Regeln muss auch stets die Normierungsbedingung der Wahrscheinlichkeitsrechnung gelten:

$$P_1 + P_2 + P_3 + P_4 = 1 \tag{3.7}$$

Mit den Formeln (3.4), (3.6) und (3.7) stehen drei Gleichungen für die vier unbekannten Wahrscheinlichkeiten $P_1, \ldots, P_4$ zur Verfügung, wodurch das Gleichungssystem unterbestimmt ist. Zur Lösung des Problems wird das Prinzip der maximalen Entropie [7] angewandt Das Prinzip überführt die Aufgabe der algebraischen Lösung des Gleichungssystems in ein Optimierungsproblem, dessen Lösung in Abschnitt 3.1.2 beschrieben wird.

3.1.2 Optimierung

Die bzgl. der Z Elemente p_z von P zu maximierende Entropiefunktion [4] lautet:

$$\mathcal{H}(P) = -\sum_z P_z \cdot log_2\,(P_z), \qquad P_z \in \mathbb{R}, 0 \leq P_z \leq 1, z = 1, \dots, Z. \tag{3.8}$$

Gesucht ist die Wahrscheinlichkeitsverteilung P^* aus einer Menge von Verteilungen $\{P\}$, welche maximale Entropie besitzt und dabei die gegebenen Nebenbedingungen $\Re$ erfüllt. Formal wird die Berechnung von P^* wie folgt beschrieben:

$$P^* = arg\max_P\big(\mathcal{H}(P)\big), \quad P \vDash \mathcal{R}. \tag{3.9}$$

Dabei bedeutet $P \vDash \mathcal{R}$, dass von jeder zulässigen Verteilung P semantisch auf alle gegebenen Regeln $\mathcal{R}$ geschlossen werden kann.

Das Optimierungsproblem gemäß Glg. (3.9) ist aufgrund der Logarithmusfunktion in Glg. (3.8) nichtlinear und besitzt als Nebenbedingungen $\mathcal{R}$ die Gleichungen (3.4), (3.6) und (3.7). Die analytische Lösung des Optimierungsproblems kann gemäß Abschnitt 2 mit der Lagrange-Methode [11] erfolgen, bei der die Nebenbedingungen, nach Null aufgelöst, in die zu optimierende Zielfunktion integriert werden. Dabei wird jede Nebenbedingung mit einem Streckungsfaktor genannt Lagrange-Multiplikator λ_j multipliziert und das Optimierungsproblem mit Nebenbedingungen in einer einzigen zu optimierenden Hilfsfunktion, $\mathcal{L}$ zusammengefasst:

$$\mathcal{L} = -\sum_z P_z \cdot log_2 P_z + \sum_j \lambda_j \left(\underline{C}_j \underline{P},\right), \qquad z = 1, \dots, Z, j = 1, \dots, J \tag{3.10}$$

Dabei beschreiben $\underline{C}$ die Koeffizientenmatrix der Nebenbedingungen[7], $\underline{C}_j$ die Koeffizienten der j-ten Nebenbedingung und $\underline{P}$ den Spaltenvektor der gesuchten Wahrscheinlichkeiten. Somit stellt $\underline{C}_j \underline{P}$ das Gleichungssystem aller Nebenbedingungen dar.

Die Ermittlung des Maximums[8] der Lagrange-Funktion $\mathcal{L}$ erfolgt nach Abschnitt 2 durch Berechnung ihres Gradienten und Nullsetzen aller partiellen Ableitungen. Hieraus resultiert ein algebraisches Gleichungssystem, dessen Lösung die gesuchten Wahrscheinlichkeiten $p_1, p_{,2}, p_3$ und p_4 ergibt. Es gilt

$$\nabla \mathcal{L}_{P_z, \lambda_j} = \underline{0}, \tag{3.11}$$

was bedeutet, dass $\mathcal{L}$ partiell nach allen Variablen $P_z \in \underline{P}$ und λ_j abzuleiten ist. Daraus resultieren $Z + J$ algebraische Gleichungen[9], von denen jede gleich Null zu setzen ist. Aus dem algebraischen Gleichungssystem sind nun die, in diesem Fall nicht interessierenden drei Lagrange-Multiplikatoren λ_j zu eliminieren, so dass bei $Z = 4$ vier Gleichungen verbleiben, welche nach den gesuchten Wahrscheinlichkeiten P_1 bis P_4 aufzulösen sind.

[7] Hierbei wird von homogenen Gleichungen als Nebenbedingungen ausgegangen. Im Fall inhomogener Gleichungen können diese durch die Normierungsbedingung in homogene Gleichungen überführt werden.
[8] Da die Entropiefunktion konkav ist [8], stellt das gefundene Maximum stets ein globales Maximum dar.
[9] Im gewählten einfachen Beispiel des Modus Ponens betragen $Z = 4$ und $J = 3$.

4 Berechnungsbeispiel

Eine Getränkeabfüllanlage wird automatisch überwacht. Dabei sollen schadhafte Flaschen aussortiert werden: Allerdings werden nicht alle Schäden erkannt und teilweise auch fehlerfreie Flaschen als defekt gemeldet / beanstandet. Daher gibt es insgesamt vier mögliche Konstellationen:

- defekt ∧ beanstandet
- nicht defekt ∧ beanstandet
- defekt ∧ nicht beanstandet
- nicht defekt ∧ nicht beanstandet

Die Überwachungseinrichtung meldet bei jeder überprüften Flasche ein Kontrollergebnis, welches durch eine boolesche Variable V_1 repräsentiert wird:

$$V_1 = \{(\text{nicht beanstandet}), (\text{beanstandet})\}.$$

Zusätzlich kann der tatsächliche Zustand jeder Flasche durch eine weitere boolesche Variable V_2 beschrieben werden:

$$V_2 = \{(\text{defekt}), (\text{nicht defekt})\}$$

Zur besseren Lesbarkeit der nachfolgenden Berechnungen werden folgende Zuordnungen getroffen:

$V_1 = b$: beanstandet

$V_1 = \bar{b}$: nicht beanstandet

$V_2 = d$: defekt

$V_2 = \bar{d}$: nicht defekt

Aus statistischen Untersuchungen des Abfüllprozesses sind folgende Wahrscheinlichkeiten bekannt:

- 5% aller Flaschen werden beanstandet: $p(b) = 0{,}05$.
- 1% aller defekten Flaschen waren nicht beanstandet worden: $p(\bar{b}|d) = 0{,}01$.

Diese beiden Erkenntnisse lassen aber einige Fragen offen, wie z. B.: Wie hoch ist die Wahrscheinlichkeit $P(d)$[10] für tatsächlich defekte Flaschen? Die Antwort lässt sich nicht aus den Ergebnissen der statistischen Untersuchungen ablesen, da dort beispielsweise keine Angaben gemacht werden über den Anteil $P(d, b)$ beanstandeter Flaschen, welche tatsächlich defekt sind. Für die gesuchte Wahrscheinlichkeit $P(d)$ gilt:

$$P(defekt) = P(d) = P(d, b) + P(d, \bar{b}) = ? \tag{4.1}$$

[10] $P(d)$ steht als Kurzschreibweise für $P(V_2 = d)$.

Genauso wenig, wie diese Fragestellung können auch andere Zusammenhänge nicht direkt aus dem gegebenen Wissen berechnet werden. Ganz offensichtlich reicht das vorhandene Wissen aus den zuvor beschriebenen statistischen Auswertungen nicht aus, um damit alle Elemente einer Wahrscheinlichkeitsverteilung der beiden Zufallsvariablen V_1 und V_2 zu berechnen. Daher soll das fehlende Wissen mit der Methode der maximalen Entropie ermittelt werden.

4.1 Wahrscheinlichkeitsverteilung

Die zugehörige Wahrscheinlichkeitsverteilung P kann als Matrix dargestellt werden und ihren Z Elementen zur weiteren Vereinfachung jeweils ein Variablennamen p_z zugewiesen werden[11]:

Tabelle 1: Elemente und Variablennamen der Wahrscheinlichkeitsverteilung P.

$P_1 = P(d, b)$	$P_3 = P(d, \bar{b})$
$P_2 = P(\bar{d}, b)$	$P_4 = P(\bar{d}, \bar{b})$

Liegen alle Elemente dieser Wahrscheinlichkeitsverteilung als Zahlenwerte vor, so können hieraus alle interessierenden Wahrscheinlichkeiten, wie z. B. $P(d)$, also die Wahrscheinlichkeit tatsächlich defekter Flaschen berechnet werden. Genauso lassen sich Angaben über die Genauigkeit der Überwachungseinrichtung machen.

4.2 Vorhandenes Wissen

Die Nebenbedingungen, also das vorhandene Wissen des Optimierungsproblems ergeben sich aus den statistischen Auswertungen der Überwachungsanlage sowie der Normierungsbedingung (4.4):

$$P(b) = 0{,}05 \tag{4.2}$$

$$P(\bar{b}|d) = 0{,}01 \tag{4.3}$$

$$P_1 + P_2 + P_3 + P_4 = 1 \tag{4.4}$$

Dabei stellen Glg. (4.2) eine Aussage mit der Wahrscheinlichkeit $p_A = P(b)$ und Glg. (4.3) eine Regel mit der Wahrscheinlichkeit$t$ $p_R = P(\overline{b|}\, d)$ dar.

Zunächst sind alle Nebenbedingungen in algebraische Gleichungen zu wandeln, als Funktionen der Variablen $p_1 \ldots p_4$ auszudrücken und nach Null aufzulösen. Dies bedeutet für Glg. (4.2):

$$P(b) = 0{,}05 \Leftrightarrow P_1 + P_2 - 0{,}05 = 0\,. \tag{4.5}$$

Zur Umformung des Konditionals in Glg. (4.3) wird Formel (3.6) angewendet:

$$P(\bar{b}|d) = 0{,}01 \Leftrightarrow 0{,}99 \cdot P_3 - 0{,}01 \cdot P_1 = 0. \tag{4.6}$$

Die Normierungsbedingung in Glg. (4.4) muss lediglich nach Null aufgelöst werden:

[11] Für die Konjunktion der Verteilung gilt das Kommutativgesetz $p(d, b) = p(b, d)$.

$$P_1 + P_2 + P_3 + P_4 - 1 = 0. \tag{4.7}$$

Mit den Gleichungen (4.5), (4.6) und (4.7) liegen nun alle drei Nebenbedingungen in Form von Nullsummen vor, welche damit in eine Lagrangefunktion gemäß Glg. (3.10) integriert werden können (siehe Abschnitt 4.3).

4.3 Optimierung der Entropie

Die Lagrangefunktion $\mathcal{L}$ des Optimierungsproblems gemäß Glg. (3.10) lautet:

$$\begin{aligned}\mathcal{L} = &-P_1 \cdot log_2 P_1 - P_2 \cdot log_2 P_2 - P_3 \cdot log_2 P_3 - P_4 \cdot log_2 P_4 \\ &+ \lambda_1(P_1 + P_2 - 0{,}05) + \lambda_2(-0{,}01 \cdot P_1 + 0{,}99 \cdot P_3) \\ &+ \lambda_3(P_1 + P_2 + P_3 + P_4 - 1), \qquad \forall P_z \in [0,1].\end{aligned} \tag{4.8}$$

Zur Ermittlung des Maximalwerts der Lagrangefunktion L wird der Gradient von L bezüglich aller P_z und λ_j berechnet und jede der $i + j$ partiellen Ableitungen gleich Null gesetzt:

$$\nabla \mathcal{L}_{P_z,\lambda_j} = \underline{0}. \tag{4.9}$$

Die Ableitung eines Terms $-P_z \cdot log_2 P_z$ nach P_z ergibt mit $-log(P_z)/log(2) - 1/log(2)$ einen „etwas unhandlichen“ Ausdruck. Zur Vereinfachung wird daher in Glg. (4.8) der Logarithmus Dualis log_2 durch den Logarithmus Naturalis ln ersetzt. Da sich beide Logarithmusfunktionen nur um einen konstanten Faktor $1/ln(2)$ unterscheiden, führt dieses zum selben Optimierungsergebnis.

Damit ergibt sich aus Glg. (4.8) folgendes System nichtlinearer Gleichungen:

$$\frac{\partial \mathcal{L}}{\partial P_1} = -ln(P_1) - 1 + \lambda_1 - 0{,}01 \cdot \lambda_2 + \lambda_3 = 0 \tag{4.10}$$

$$\frac{\partial \mathcal{L}}{\partial P_2} = -ln(P_2) - 1 + \lambda_1 + \lambda_3 = 0 \tag{4.11}$$

$$\frac{\partial \mathcal{L}}{\partial P_3} = -ln(P_3) - 1 + 0{,}99 \cdot \lambda_2 + \lambda_3 = 0 \tag{4.12}$$

$$\frac{\partial \mathcal{L}}{\partial P_4} = -ln(P_4) - 1 + \lambda_3 = 0 \tag{4.13}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = P_1 + P_2 - 0{,}05 = 0 \tag{4.14}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} = -0{,}01 \cdot P_1 + 0{,}99 \cdot P_3 = 0 \tag{4.15}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_3} = P_1 + P_2 + P_3 + P_4 - 1 = 0 \tag{4.16}$$

Die sieben Gleichungen (4.10) bis (4.16) beschreiben ein nichtlineares Gleichungssystem, dessen symbolische Lösung nichttrivial ist. Daher wird ein nummerisches Verfahren angewandt, welches beispielsweise in *Matlab*© implementiert ist. Zur Anwendung kommt der Levenberg-Marquardt-Algorithmus [12], welcher auf der Methode der kleinsten Quadrate basiert und in *Matlab* in der Funktion *fsolve* verfügbar ist.

4.3.1 Matlab-Code

Im Matlab-Code in Tabelle 2 werden die sieben Gleichungen (4.10) bis (4.16) in der Funktion $F = pd\_lagr(p)$[12] deklariert. Zur Vereinfachung der Schreibweise wurden die drei Lagrange-Multiplikatoren eben falls als Variablen in P beschrieben: $\lambda_1 == P(5)$, $\lambda_2 = P(6)$ und $\lambda_3 = P(7)$.

Tabelle 2: Matlab-Code zur ´Deklaration der partiellen Ableitungen der Lagrangefunktion.

```
function F = pd_lagr(p)

F(1) = -log(p(1))-1+p(5)-0.01*p(6)+p(7);
F(2) = -log(p(2))-1+p(5)+p(7);
F(3) = -log(p(3))-1+0.99*p(6)+p(7);
F(4) = -log(p(4))-1+p(7);
F(5) = p(1)+p(2)-0.05;
F(6) = -0.01*p(1)+0.99*p(3);
F(7) = p(1)+p(2)+p(3)+p(4)-1;
```

Tabelle 3 zeigt den Matlab-Code zur Optimierung. Dabei wird $@pd\_lagr$ als Function Handle in fun deklariert. Danach werden der Startvektor $\underline{P_0}$ der Optimierung festgelegt, der Levenberg-Marquard-Algorithmus ausgewählt und Optionen [13] zur Steuerung des Solvers definiert. Abschließend wird als eine erste Plausibilisierung des Optimierungsergebnisses die Summe s der Variablen P_1 bis P_4 berechnet. Diese muss gemäß der Normierungsbedingung stets gleich Eins sein.

[12] pd_{lagr} steht für Partial Derivates of Lagrangian.

Tabelle 3_ Matlab-Code zur Optimierung.

```
fun = @pd_lagr;

P0 = [0.01,0.01,0.01,0.01,0.01,0.01,0.01];
options = optimoptions('fsolve','Algorithm','levenberg-mar-
quardt');
options.MaxFunctionEvaluations = 1.400000e+05;
options.MaxIterations = 4.000000e+03;
P = fsolve(fun,P0,options)

s = P(1)+P(2)+P(3)+P(4)% Check der Normierungsbed.: s=1
```

4.4 Optimierungsergebnisse

Tabelle 4 zeigt die Wahrscheinlichkeitsverteilung P als Ergebnis des Optimierungsprozesses. Die Genauigkeit der Resultate hängt dabei von der gewählten Konvergenzschwelle und dem Maximalwert an Iterationen des Algorithmus ab.

Tabelle 4: Wahrscheinlichkeitsverteilung P als Ergebnis der Optimierung.

$P_1 = P(d,b) = 0{,}0260$	$P_3 = P(d,\bar{b}) = 0{,}0003$
$P_2 = P(\bar{d},b) = 0{,}0240$	$P_4 = P(\bar{d},\bar{b}) = 0{,}9497$

Zunächst ist die Wahrscheinlichkeitsverteilung P als Ergebnis der Optimierung gemäß Tabelle 4 zu überprüfen. Hierzu werden die Wahrscheinlichkeiten der Nebenbedingungen (4.5), (4.6) und (4.7) aus der Tabelle 4 berechnet. Da die Gleichungen (4.5) bis (4.7) nach Null aufgelöst waren, sollten die Resultate dieser Plausibilisierung ebenfalls gleich Null sein oder sehr kleine Werte ergeben.

Tabelle 5 zeigt, dass die vorgegebenen Nebenbedingungen bei der Optimierung korrekt berücksichtigt wurden. Als Ergebnis dieser Betrachtungen kann die ermittelte Wahrscheinlichkeitsverteilung P nun für Anfragen nach Wahrscheinlichkeiten genutzt werden.

Aus $P_3 = P(d,\bar{b}) = 0{,}0003$ liest man unmittelbar ab, dass 0,3 Promille aller Flaschen einen Defekt haben, der nicht erkannt wurde und $P_2 = P(\bar{d},b) = 0{,}0240$ besagt, dass 2,4 Prozent aller Flaschen ohne Defekt beanstandet wurden. Man erkennt, dass die Überwachung so eingestellt wurde, dass möglichst keine Defekte übersehen werden und nimmt dafür einen höheren Anteil an Fehlbeanstandungen in Kauf.

Tabelle 5: Vergleich von vorgegebenen und berechneten Wahrscheinlichkeiten.

Gegebene Nebenbedingung	Wahrscheinlichkeit aus Tabelle 4	Ergebnis
$P(b) = 0{,}05 \Leftrightarrow P_1 + P_2 - 0{,}05 = 0$. (4.5)	$P_1 + P_2 - 0{,}05 = 0{,}0260 + 0{,}0240 - 0{,}05 = 0.$	✓
$P(\bar{b}\|d) = 0{,}01 \Leftrightarrow 0{,}99 \cdot P_3 - 0{,}01 \cdot P_1 = 0.$ (4.6)	$0{,}99 \cdot P_3 - 0{,}01 \cdot P_1 = 0{,}99 \cdot 0{,}0003 - 0{,}01 \cdot 0{,}0260 = -0{,}000252278$	✓ [13]
$P_1 + P_2 + P_3 + P_4 - 1 = 0.$ (4.7)	$P_1 + P_2 + P_3 + P_4 - 1 = 0{,}0260 + 0{,}0240 + 0{,}0003 + 0{,}9497 = 0$	✓

In der Einleitung dieses Abschnitts 4 wurde die Frage nach der Wahrscheinlichkeit $p(d)$ für tatsächlich defekte Flaschen gestellt. Die gesuchte Wahrscheinlichkeit $p(d)$ ergibt sich als Summe der Wahrscheinlichkeiten von defekten beanstandeten und defekten nicht beanstandeten Flaschen. Somit ergibt sich aus Tabelle 4:

$$P(d) = P(d, b) + P(d, \bar{b}) = 0{,}0260 + 0{,}0003 = 0{,}0263 \tag{4.17}$$

Auf diese Weise können auch alle anderen a priori Wahrscheinlichkeiten[14] aus der vollständigen Wahrscheinlichkeitsverteilung P gemäß Tabelle 4 berechnet werden.

Zur Ermittlung von Konditionalen aus Tabelle 4 wird die Definition der bedingten Wahrscheinlichkeit verwendet:

$$PP(B_j|\, A_j) = \frac{P(A_j, B_j)}{P(A_j)} \tag{4.18}$$

So beträgt beispielsweise der Anteil nicht defekter Flaschen unter den nicht beanstandeten Flaschen

$$P(\bar{d}|\bar{b}) = \frac{P(\bar{d}, \bar{b})}{P(\bar{b})} = \frac{0{,}9497}{0{,}9500} = 0{,}9997. \tag{4.19}$$

Das einfache Berechnungsbeispiel soll die grundlegende Vorgehensweise zur Anwendung des Prinzips der maximalen Entropie beschreiben. Das ganze Potenzial des Prinzips zeigt sich aber in zunehmendem Maße erst mit steigender Modellkomplexität und einer größeren Anzahl an Nebenbedingungen.

Die durch das Prinzip der maximalen Entropie berechneten Wahrscheinlichkeiten repräsentieren die aus informationstheoretischer Sicht bestmöglichen Annahmen, da sie dem vorhandenen Wissen so wenig wie möglich nicht gegebenes Wissen hinzufügen. Dadurch ermittelt das

[13] Die (in diesem Beispiel geringe) Abweichung vom Sollergebnis Null hängt von den gewählten Optimierungsoptionen, wie z. B. maximale Anzahl der Iterationen des Algorithmus ab.

[14] A priori Wahrscheinlichkeiten sind naheliegende oder augenscheinliche Wahrscheinlichkeiten, welche noch nicht durch Experimente a posteori abgesichert sind.

Prinzip der maximalen Entropie die bestmögliche Wahrscheinlichkeitsverteilung unter der Prämisse dieser bestmöglichen Annahmen.

5 Literaturangaben

[1] Ertel, W.: Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung. Computational Intelligence. Wiesbaden: Springer Vieweg 2016

[2] Braun, H.: Sicherheitsbewertung von Fahrsituationen mit der Methode der minimalen relativen Entropie, University Library Hagen 2022

[3] Shannon, C. E.: A Mathematical Theory of Communication. The Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656 (1948)

[4] Rohling,, H.: Einführung in die Informations- und Codierungstheorie. Springer 1995

[5] Meyer, C.-H.: Korrektes Schließen bei unvollständiger Information. Anwendung des Prinzips der maximalen Entropie in einem probabilistischen Expertensystem. Zugl.: Hagen, Fernuniv., Diss., 1997. Europäische Hochschulschriften Reihe 41, Informatik, Bd. 29. Frankfurt am Main: Lang 1998

[6] Rödder, W., Reucher, E. u. Kulmann, F.: Features of the Expert-System-Shell SPIRIT. Logic Journal of IGPL 14 (2006) 3, S. 483–500

[7] Bretthorst, G. L. u. Jaynes, E. T. (Hrsg.): Probability theory. The logic of science. Cambridge: Cambridge Univ. Press 2013

[8] Radhakrishna, C.: CONVEXITY PROPERTIES OF ENTROPY FUNCTIONS AND ANALYSIS OF DIVERSITY. Pittsburgh: IMS Lecture Notes 1984

[9] Förster, m. O.: Analysis 2. Differentialrechnung im IRn, gewöhnliche Differentialgleichungen. Grundkurs Mathematik. Wiesbaden: Vieweg+Teubner 2008

[10] Papula, L.: Mathematik für Ingenieure und Naturwissenschaftler. Wiesbaden, Heidelberg: Springer Vieweg 2018

[11] Forster, O.: Analysis. Vieweg-Studium : Grundkurs Mathematik. Wiesbaden: Vieweg 2008

[12] Marquardt, D. W.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Society for Industrial and Applied Mathematics 11 (1963) 2, S. 431–441

[13] MathWorks: fsolve. Solve system of nonlinear equations, 2022. https://de.mathworks.com/help/optim/ug/fsolve.html, abgerufen am: 07.12.2022