Mustererkennung: Wahrscheinlichkeitstheorie

Definitionen (axiomatisch)

Wahrscheinlichkeitsraum – (Ω, σ, P), mit

Ω – Die Grundmenge, die Menge der elementaren Ereignisse,

σ – σ-Algebra,

P – Wahrscheinlichkeitsmaß.

σ-Algebra über die Grundmenge Ω ist das System von Teilmengen von Ω, d.h. $\sigma \subseteq \mathcal{P}(\Omega)$ ($\mathcal{P}$ ist die Potenzmenge) mit:

$$\begin{aligned} \Omega \in \sigma & \\ A \in \sigma \quad & \Rightarrow \quad \Omega/A \in \sigma \\ A_i \in \sigma, i = 1, \ldots, n \quad & \Rightarrow \quad \bigcup_{i=1}^{n} A_i \in \sigma \end{aligned}$$

σ ist bezüglich Kompliment-Bildung und abzählbarer Vereinigung abgeschlossen. Daraus folgt: $\emptyset \in \sigma$, abzählbarer Schnitt $\in \sigma$ (über De Morganschen Gesetz).

Beispiele:

- $\sigma = \{\emptyset, \Omega\}$ (kleinste) und $\sigma = \mathcal{P}(\Omega)$ (größte) σ-Algebren über Ω
- Die minimale σ-Algebra über Ω, die $A \subset \Omega$ enthält ist $\{\emptyset, A, \Omega/A, \Omega\}$
- Ω – diskret und endlich, $\sigma = 2^{\Omega}$
- $\Omega = \mathbb{R}$, die σ-Algebra der Borelschen Teilmengen (enthält u.a. alle Intervalle).

Definitionen (axiomatisch)

Wahrscheinlichkeitsmaß $P : \sigma \rightarrow [0, 1]$ ist ein „Maß“ (Π) mit Normierung:

$$P(\Omega) = 1$$

σ-Additivität: sei $A_i \in \sigma$ paarweise disjunkt, d.h. $A_i \cap A_{i'} = \emptyset$. Dann

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Achtung! Es gibt Mengen, für die kein „Maß“ existiert.
Beispiele: Menge irrationaler Zahlen, Funktionsräume $\mathbb{R}^\infty$ (?) etc.
Banach-Tarski-Paradoxon

Definitionen (menschlich) – praktisch relevante Spezialfälle:

- Menge der elementaren Ereignisse Ω ist „gutartig“, z.B. $\mathbb{R}^n$, endlich usw., $\sigma = \mathcal{P}(\Omega)$
- Ein (zusammengesetztes) Ereignis $A \subseteq \Omega$ (Vereinigung der elementaren Ereignisse)
- Wahrscheinlichkeit

$$P(A) = \sum_{\omega \in A} P(\omega)$$

Statistische Größen

Allgemein: eine statistische Größe ξ ist eine Abbildung $\xi : \Omega \to E$ (E ist der Beobachtungsraum) mit bestimmten Eigenschaften.

Spezialfall – eine reellwertige statistische Größe ξ für einen Wahrscheinlichkeitsraum (Ω, σ, P) ist eine Abbildung $\xi : \Omega \to \mathbb{R}$ wenn

$$\{\omega : \xi(\omega) \leq r\} \in \sigma \quad \forall r \in \mathbb{R}$$

(immer erfüllt, wenn $\sigma = \mathcal{P}(\Omega)$).

Verteilungsfunktion einer statistischen Größe ξ ist

$$F_\xi(r) = P(\{\omega : \xi(\omega) \leq r\})$$

Wahrscheinlichkeitsverteilung einer diskreten statistischen Größe $\xi : \Omega \to \mathbb{Z}$ ist

$$p_\xi(r) = P(\{\omega : \xi(\omega) = r\})$$

Wahrscheinlichkeitsdichte einer kontinuierlichen statistischen Größe $\xi : \Omega \to \mathbb{R}$ ist

$$p_\xi(r) = \frac{\partial F_\xi(r)}{\partial r}$$

Statistische Größen

Bemerkung: Elementare Ereignisse sind keine Zahlen!!!
Im Gegensatz dazu kann man statistische Größen addieren, multiplizieren usw.

Mittelwert einer statistischen Größe:

$$\mathbb{E}_P(\xi) = \sum_{\omega \in \Omega} P(\omega) \cdot \xi(\omega) = \sum_r \sum_{\omega : \xi(\omega) = r} P(\omega) \cdot r = \sum_r p_\xi(r) \cdot r$$

Beispiel: Augenzahl eines Würfels.

Grundmenge (6 Facetten) $\Omega = \{a, b, c, d, e, f\}$
Wahrscheinlichkeitsmaß $P(\{a\}) = 1/6$, $P(\{c, f\}) = 1/3$ ...
Statistische Größe (Augenzahl) $\xi(a) = 1$, $\xi(b) = 2$ ... $\xi(f) = 6$.
Verteilungsfunktion $F_\xi(3) = 1/2$, $F_\xi(4.5) = 2/3$ ...
Wahrscheinlichkeitsverteilung $p_\xi(1) = p_\xi(2) = \ldots = p_\xi(6) = 1/6$
Erwartungswert $\mathbb{E}_P(\xi) = 3.5$

Eine andere statistische Größe (Augenzahl^2) $\xi'(a) = 1$, $\xi'(b) = 4$ ... $\xi'(f) = 36$
Erwartungswert $\mathbb{E}_P(\xi') = 15\frac{1}{6}$

Statistische Größen

Beispiel: Augenzahlen der zwei (unabhängigen) Würfel.

Grundmenge (6 Facetten×6 Facetten) $\Omega = \{a, b, c, d, e, f\} \times \{a, b, c, d, e, f\}$
Wahrscheinlichkeitsmaß $P(\{ab\}) = 1/36$, $P(\{cd, fa\}) = 1/18$ ...
Zwei statistischen Größen:
Augenzahl des ersten Würfels $\xi_1(ab) = 1$, $\xi_1(ac) = 1$ ... $\xi_1(ef) = 5$ ...
Augenzahl des zweiten Würfels $\xi_2(ab) = 2$, $\xi_2(ac) = 3$ ... $\xi_2(ef) = 6$ ...

Wahrscheinlichkeitsverteilungen:
$p_{\xi_1}(1) = p_{\xi_1}(2) = \ldots = p_{\xi_1}(6) = 1/6$
$p_{\xi_2}(1) = p_{\xi_2}(2) = \ldots = p_{\xi_2}(6) = 1/6$

Die neue statistische Größe (summe der Augenzahlen) $\xi = \xi_1 + \xi_2$
Die Wahrscheinlichkeitsverteilung p_ξ ist nicht mehr gleichwahrscheinlich – :-)
$p_\xi \sim (1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1)$

Erwartungswert $\mathbb{E}_P(\xi) = 7$

Generell für Erwartungswerte:

$$\mathbb{E}_P(\xi_1 + \xi_2) = \sum_{\omega \in \Omega} P(\omega) \cdot \Big(\xi_1(\omega) + \xi_2(\omega)\Big) = \mathbb{E}_P(\xi_1) + \mathbb{E}_P(\xi_2)$$

Statistische Größen

Mehrdimensionale (reellwertige) statistische Größen – $\xi : \Omega \to \mathbb{R}^n$
Äquivalent (Einfachheit halber $n = 2$): $\xi = (\xi_1, \xi_2)$ mit $\xi_1 : \Omega \to \mathbb{R}$ und $\xi_2 : \Omega \to \mathbb{R}$

Verteilungsfunktion:

$$F_\xi(r, s) = P(\{\omega : \xi_1(\omega) \leq r\} \cap \{\omega : \xi_2(\omega) \leq s\})$$

Verbundwahrscheinlichkeitsverteilung (diskret):

$$p_{\xi=(\xi_1,\xi_2)}(r, s) = P(\{\omega : \xi_1(\omega) = r\} \cap \{\omega : \xi_2(\omega) = s\})$$

Verbundwahrscheinlichkeitsdichte (kontinuierlich):

$$p_{\xi=(\xi_1,\xi_2)}(r, s) = \frac{\partial F_\xi(r, s)}{\partial r \partial s}$$

Weitere Definitionen

Zwei Ereignisse $A \subset \Omega$ und $B \subset \Omega$ sind **unabhängig**, wenn

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

Interessant: Ereignisse A und $\bar{B} = \Omega / B$ sind unabhängig, falls A und B unabhängig sind.

Bedingte Wahrscheinlichkeit:

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

Eine andere Definition der Unabhängigkeit: A und B sind unabhängig, wenn

$$P(A|B) = P(A) \;\; \text{bzw.} \quad P(B|A) = P(B)$$

(äquivalent, wenn $P(B) \neq 0$ bzw. $P(A) \neq 0$).

Zwei statistische Größen sind unabhängig, wenn

$$F_{\xi=(\xi_1,\xi_2)}(r,s) = F_{\xi_1}(r) \cdot F_{\xi_2}(s) \quad \forall r,s$$

Bayessche Formel:

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

Weitere Definitionen (für statistische Größen)

Kürzel: $p(x,y) \equiv p_\xi(x,y)$

Marginale Wahrscheinlichkeit:

$$p(x) = \sum_y p(x,y)$$

Bedingte Wahrscheinlichkeit:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

Merke: $\sum_x p(x|y) = 1$.

Unabhängige Wahrscheinlichkeitsverteilungen:

$$p(x,y) = p(x) \cdot p(y)$$

oder (analog zu Ereignissen)

$$p(x|y) = p(x)$$

Interessant: In das Intervall $[0,1]$ werden zufällig und nacheinander 3 Punkte geworfen. Wie groß ist die Wahrscheinlichkeit, dass die Koordinate des dritten Punktes größer ist, als die der beiden ersten?

Erkennung

Das Modell:
Gegeben sei zwei statistische Größen.
Typischerweise ist eine davon diskret (d.h. $k \in K$) und heißt „Klasse“.
Die andere ist allgemein (sehr oft kontinuierlich, d.h. $x \in \mathbb{R}^n$) und heißt „Beobachtung“.

„Gegeben“ sei die Verbundwahrscheinlichkeitsverteilung $p(x, k)$.
Da die k diskret ist, wird oft $p(x, k)$ durch $p(x, k) = p(k) \cdot p(x|k)$ spezifiziert.

Die Erkennungsaufgabe:
man beobachtet x, man sage etwas über k
– „welche Klasse hat die Beobachtung x verursacht“.

Schwierigkeiten (Fragen):

- Wie ist k anhand der x zu wählen?
- Die Wahrscheinlichkeitsverteilung ist oft nicht explizit angegeben
  $\rightarrow$ dafür sorgen, dass das eine Wahrscheinlichkeitsverteilung ist.
- Die Menge K ist manchmal sehr groß (erinnere an Hopfield-Netze).

Das Lernen:
Das Wahrscheinlichkeitsmodell ist oft bis auf unbekannte freie Parameter definiert.
Wie sind diese zu wählen?