

Mathematik 3: Zusatzübung - Wahrscheinlichkeit bei KI (LLM)

1. Aufgabe:

Ein Large Language Model (LLM) sagt für den Satzanfang "Die Klausur ist..." folgende Wahrscheinlichkeiten für das nächste Token voraus:

Token	Wahrscheinlichkeit
"schwer"	0,35
"einfach"	0,15
"morgen"	0,20
"vorbei"	0,10
"lang"	0,08
"interessant"	0,07
sonstige Tokens	0,05

- Prüfen Sie rechnerisch, ob es sich um eine gültige Wahrscheinlichkeitsverteilung handelt.
- Mit welcher Wahrscheinlichkeit wählt das Modell **nicht** das Token "schwer"?
- Das Modell generiert (unabhängig voneinander) drei Sätze mit obiger Verteilung. Wie groß ist die Wahrscheinlichkeit, dass **alle drei** Male das Token "schwer" gewählt wird? Wie groß ist die Wahrscheinlichkeit, dass **mindestens einmal** das Token "schwer" gewählt wird?
- Wie groß ist die Wahrscheinlichkeit, dass das gewählte Token "schwer" **oder** "einfach" ist?

Lösung:

a) Bedingung: $\sum_i p_i = 1$ und $p_i \geq 0 \forall i$.

$$0,35 + 0,15 + 0,20 + 0,10 + 0,08 + 0,07 + 0,05 = \mathbf{1,00}$$

(Grundlagen

Wahrscheinlichkeit)

$\Rightarrow$ es handelt sich um eine **gültige Verteilung**.

b) Gegenereignis: $P(\overline{A}) = 1 - P(A)$

$$P(\text{nicht "schwer"}) = 1 - 0,35 = \mathbf{0,65} \text{ (65 \%)}$$

(Gegenwahrscheinlichkeit)

c) Multiplikationssatz (unabhängige Ereignisse):

$$P(3 \times \text{"schwer"}) = 0,35^3 = \mathbf{0,0429} \text{ (4,29 \%)} \quad (\text{Unabhängigkeit})$$

Über das Gegenereignis (kein einziges Mal "schwer"):

$$P(\text{mind. } 1 \times \text{"schwer"}) = 1 - (1 - 0,35)^3 = 1 - 0,65^3 = \mathbf{0,725} \text{ (72,5 \%)} \quad (\text{Gegenereignis + Unabhängigkeit})$$

d) Da sich beide Ereignisse gegenseitig ausschließen (es wird nur ein Token gewählt), gilt der Additionssatz:

$$P(\text{"schwer" oder "einfach"}) = 0,35 + 0,15 = \mathbf{0,50} \text{ (50 \%)}$$

(Additionssatz,

disjunkte Ereignisse)

2. Aufgabe:

Ein LLM wird auf Faktentreue geprüft. Aus Erfahrung ist bekannt:

- Bei 10 % aller Antworten "halluziniert" das Modell (erfindet falsche Fakten): $P(H) = 0,10$.
- Ein nachgeschalteter Faktencheck-Algorithmus erkennt eine Halluzination mit Wahrscheinlichkeit 0,90 korrekt als "auffällig" ($P(A | H) = 0,90$).
- Bei einer korrekten (nicht-halluzinierten) Antwort schlägt der Faktencheck trotzdem fälschlich mit Wahrscheinlichkeit 0,08 Alarm ($P(A | \overline{H}) = 0,08$).

Der Faktencheck schlägt bei einer Antwort Alarm (Ereignis A). Wie groß ist die Wahrscheinlichkeit, dass die Antwort tatsächlich eine Halluzination ist ($P(H | A)$)?

Lösung:

Satz von Bayes:

$$P(H | A) = \frac{P(A | H) \cdot P(H)}{P(A | H) \cdot P(H) + P(A | \overline{H}) \cdot P(\overline{H})}$$
$$P(H | A) = \frac{0,90 \cdot 0,10}{0,90 \cdot 0,10 + 0,08 \cdot 0,90} = \frac{0,09}{0,09 + 0,072} = \frac{0,09}{0,162} = \mathbf{0,556} \text{ (55,6 \%)} \quad (\text{Satz von Bayes})$$

Obwohl der Faktencheck bei tatsächlichen Halluzinationen sehr zuverlässig ist (90%), ist bei "Alarm" nur gut jede zweite Antwort tatsächlich falsch – Folge der geringen Basisrate (nur 10 % aller Antworten halluzinieren).

Lösung der Aufgaben mit R

Aufgabe 1a) – Summe der Token-Wahrscheinlichkeiten:

```
> sum(c(0.35, 0.15, 0.20, 0.10, 0.08, 0.07, 0.05))  
[1] 1
```

Aufgabe 1b) – Gegenwahrscheinlichkeit zu "schwer":

```
> 1 - 0.35  
[1] 0.65
```

Aufgabe 1c) – Drei Mal unabhängig "schwer":

```
> 0.35^3  
[1] 0.042875
```

Mindestens einmal "schwer" bei 3 Versuchen:

```
> 1 - (1 - 0.35)^3  
[1] 0.725375
```

Aufgabe 1d) – "schwer" oder "einfach" (disjunkt, Additionssatz):

```
> 0.35 + 0.15  
[1] 0.5
```

Aufgabe 2 – Satz von Bayes, $P(H | A)$:

```
> P_H <- 0.10; P_A_gegeben_H <- 0.90; P_A_gegeben_nichtH <- 0.08  
> zaehler <- P_A_gegeben_H * P_H  
> nenner <- zaehler + P_A_gegeben_nichtH * (1 - P_H)  
> zaehler / nenner  
[1] 0.5555556
```