# Internationales Mathe-Turnier 2021
# Aufgaben „Sum of Us“

24. September 2021

Liebe Teilnehmer*innen,

dies sind die Aufgaben für den zweiten Teil des Internationalen Mathe-Turniers 2021, die „Sum of Us“-Runde. Wie ihr wisst, steht das diesjährige Turnier unter dem Motto *Angebotssysteme*. Bevor ihr beginnt, empfehlen wir euch die Anweisungen auf der nächsten Seite sorgfältig durchzulesen.

Die Aufgaben wurden von Roy Hendricks und Kiki van den Bercken, Studierende der Mathematik an der Radboud Universiteit Nijmegen, entwickelt, und die endgültige Version wurde in Zusammenarbeit mit Maarten Solleveld und Joeri Van der Veken erstellt.

Wir wünschen euch viel Glück - aber vor allem viel Spaß - beim Mathe-Turnier!

Die Organisatoren,

Stefan Hartmann und Rainer Kaenders (Universität Bonn)
Maarten Solleveld (Radboud Universiteit Nijmegen)
Joeri Van der Veken (KU Leuven)

# Einleitung

## Lösungen

- Schreib deine Antworten auf die bereitgestellten Antwortbögen.
- Runde die Endergebnisse immer auf die zweite Dezimalstelle.
- Wenn die Antwort ein Wahrscheinlichkeitsverhältnis ist, gebe dies wie folgt an:
$$1 : x_1 : \ldots : x_n.$$
- Bei Anwendung der Bayes-Methode verwende die Laplace-Glättung mit Parameter $\alpha = 1$ (aber immer nur dann, wenn diese nötig ist).

## Punkteverteilung

**Aufgabe 1: 80**
**Aufgabe 2: 155**
**Aufgabe 3: 155**
**Aufgabe 4: 110**
Die maximal zu erreichende Punktzahl für diese Runde beträgt 500 Punkte.

## Zugelassene Materialien

- ein nicht-grafikfähiger Taschenrechner
- das bereitgestellte Vorbereitungsmaterial
- Ausarbeitungen zu den Übungen des Vorbereitungsmaterials

# 1 Ham of Spam?

Ein Spam-Filter ist im Grunde nur ein Empfehlungssystem, das euch sagt, welche E-Mails ihr lesen solltet und welche nicht. *"Bag of words"-Filter* basieren auf der Bayes-Methode und funktionieren - in ihrer einfachsten Form - wie folgt: Der*die Benutzer*in gibt bei einigen Nachrichten an, ob es sich um *Spam* oder *Ham* (engl. Schinken, d. h. kein Spam) handelt. Das System merkt sich, welche Wörter in diesen Ham- und Spam-Nachrichten vorkommen. Der Name dieses Filtertyps kommt also daher, dass Nachrichten lediglich als Wortansammlungen betrachtet werden.

Wenn dann eine neue Nachricht mit $n$ Wörtern eintrifft, schätzt das System das Verhältnis zwischen der bedingten Wahrscheinlichkeit, dass es sich bei der Nachricht um Ham handelt, wenn sie die $n$ Wörter enthält, und der bedingten Wahrscheinlichkeit, dass es sich bei der Nachricht um Spam handelt, wenn sie die $n$ Wörter enthält. Liegt dieses Verhältnis unter $5 : 95$, so wird die Nachricht als Spam eingestuft, andernfalls als Ham.

Angenommen, ein*e Benutzer*in hat die folgenden vier E-Mails als Ham markiert:

1. `Your order will be delivered today.`
2. `Congratulations! Happy Birthday!`
3. `Could you go to the grocery store today? I will cook.`
4. `See you soon!`

und die folgenden drei als Spam:

1. `Click here to renew your password.`
2. `You won the lottery. Congratulations!`
3. `Click here to order Viagra from our online store.`

(a) Schätze auf Grundlage dieser Daten das Wahrscheinlichkeitsverhältnis „die Nachricht ist Ham“ : „die Nachricht ist Spam“ für die folgende E-Mail:

`Congratulations, click here to see your birthday card.`

Verwende die Bayes-Methode. Wird das System die Nachricht als Ham oder Spam kategorisieren?

(b) Laut einer aktuellen Studie sind 85% des gesamten E-Mail-Verkehrs Spam. Verwende dies, um die Schätzung des Wahrscheinlichkeitsverhältnisses aus (a) zu verbessern. Wird die Nachricht nun als Ham oder Spam kategorisiert?

# 2 Musik

Bei der Registrierung auf einer auf Pop-, Rock- und klassische Musik spezialisierten Plattform für das Streaming von Musik muss jede*r Nutzer*in die folgenden drei Multiple-Choice-Fragen beantworten:

| | |
|---|---|
| Geschlecht: | o M<br>o W<br>o D |
| Welche Musik haben Sie gehört, als Sie jung waren? | o Pop<br>o Rock<br>o Klassik<br>o andere |
| Was ist Ihr höchster Bildungsabschluss? | o Grundschule<br>o Gymnasial-/Realschul/-Hauptschulabschluss o.ä.<br>o Universitäts- bzw. Hochschulabschluss |

Die Administrator*innen der Plattform kennen den Musikgeschmack von 13 Nutzer*innen, indem sie speichern, welche Lieder diese streamen:

| Nutzer | Geschlecht | Jugendmusik | Abschluss | Musikgeschmack |
|---|---|---|---|---|
| 1 | W | Klassik | hoch | Klassik |
| 2 | M | Rock | niedrig | Rock |
| 3 | W | Rock | mittel | Rock |
| 4 | M | Pop | niedrig | Pop |
| 5 | W | Klassik | hoch | Klassik |
| 6 | W | Pop | mittel | Pop |
| 7 | M | Rock | hoch | Rock |
| 8 | W | Pop | niedrig | Pop |
| 9 | M | Klassik | mittel | Klassik |
| 10 | M | Rock | hoch | Klassik |
| 11 | M | Pop | niedrig | Rock |
| 12 | M | Rock | hoch | Pop |
| 13 | W | Klassik | mittel | Pop |

Zwei neue Nutzer*innen melden sich an: Josef hat einen gymnasialen Schulabschluss und ist der Sohn begeisterter Rocker*innen und Lisa ist eine Frau,

deren Eltern immer klassische Musik spielten, als sie jung war. Lisa hat lediglich einen niedrigen Bildungsabschluss, sie musste die Hauptschule abbrechen. Die Administrator*innen der Plattform wollen wissen, welche Musikrichtung sie Josef und Lisa am ehesten empfehlen sollten.

(a) Mache eine (mathematisch begründete) Vorhersage zum Musikgeschmack von Josef mit der Bayes-Methode. Schätze dazu folgendes Wahrscheinlichkeitsverhältnis:

'Josef gefällt Pop' : 'Josef gefällt Rock' : 'Josef gefällt klassische Musik'.

(b) Wenn 'Geschlecht', 'Jugendmusik' und 'Bildung' mit 2, 3 bzw. 1 gewichtet werden, was ist das kleinste $k > 0$, für das Josefs $k$ nächste Nachbar*innen nicht den gleichen Abstand haben wie sein*e $(k+1)$-te*r Nachbar*in? Wer sind diese $k$-Nachbar*innen? Was ist der voraussichtliche Musikgeschmack?

(c) Mache eine (mathematisch begründete) Vorhersage zum Musikgeschmack von Lisa mit der Bayes-Methode. Schätze dazu folgendes Wahrscheinlichkeitsverhältnis:

'Lisa gefällt Pop'. : 'Lisa gefällt Rock'. : 'Lisa gefällt klassische Musik'.

(d) Wenn 'Geschlecht', 'Jugendmusik' und 'Bildung' mit 1, 2 bzw. 1 gewichtet werden, was ist das kleinste $k > 0$, für das Lisas $k$ nächste Nachbar*innen nicht den gleichen Abstand haben wie ihr*e $(k+1)$-te*r Nachbar*in? Wer sind diese $k$-Nachbar*innen? Was ist der voraussichtliche Musikgeschmack?

# 3 No Time to Die

Ali, Bas, Cas, Dim und Eva haben vier James-Bond-Filme mit mindestens 1 Stern und maximal 5 Sternen bewertet. Je mehr Sterne sie einem Film gaben, desto besser fanden sie ihn. Allerdings hatte Dim zum Zeitpunkt der Bewertung "Casino Royale" noch nicht gesehen.

| | Dr. No | GoldenEye | Casino Royale | Spectre |
|---|---|---|---|---|
| Ali | ** | * * * | * * ** | * * * * * |
| Bas | * * * * * | * * ** | * * * * * | ** |
| Cas | ** | ** | * * * * * | * * ** |
| Dim | * * ** | ** | | * * * * * |
| Eva | ** | * * ** | * * * | * * * |

(a) Sage die Anzahl der Sterne voraus, die Dim für "Casino Royale" geben wird, indem du eine *personenbasierte* gemeinsame Filterung mit Umgebungsgröße 3 verwendest. Wie immer wird die Antwort auf die zweite Dezimalstelle gerundet. Welche Umgebung hast du dafür verwendet?

(b) Sage die Anzahl der Sterne voraus, die Dim für "Casino Royale" geben wird, indem du eine *itembasierte* gemeinsame Filterung mit Umgebungsgröße 3 verwendest. Wie immer wird die Antwort auf die zweite Dezimalstelle gerundet. Welche Umgebung hast du dafür verwendet?

Nachdem er "Casino Royale" gesehen hat, gibt Dim dem Film 5 Sterne. Ali, Bas und Eva gehören zu den wenigen Glücklichen, die den lang erwarteten neuen Bond-Film "No Time to Die" bereits gesehen haben und ihn mit 4, 3 bzw. 3 Sternen bewerten.

(c) Verwende alle verfügbaren Daten, um die Bewertungen von Cas und Dim für "No Time to Die" mit Hilfe von itembasierter gemeinsamer Filterung mit Umgebungsgröße 3 vorherzusagen. Wie immer wird die Antwort auf die zweite Dezimalstelle gerundet. Würdest du eher Cas oder Dim den neuen Film empfehlen?

# 4 Binge-Watching

Bei einem Streaming-Dienst für digitales Fernsehen bewerteten 14 Nutzer*innen eine Reihe von Fernsehserien mit mindestens 1 und höchstens 5 Sternen. Die Serien wurden nach Genres gruppiert, und die Durchschnittsnote, die jede Person für jedes Genre vergeben hat, ist in der nachstehenden Tabelle zu finden.

| Nutzer | Action | Science fiction | Comedy | Drama |
|---|---|---|---|---|
| 1 | 1.9 | 4.4 | 3.7 | 4.0 |
| 2 | 4.3 | 2.4 | 4.0 | 2.7 |
| 3 | 2.2 | 4.6 | 3.8 | 4.8 |
| 4 | 2.3 | 4.0 | 3.4 | 2.9 |
| 5 | 4.8 | 4.3 | 3.0 | 4.4 |
| 6 | 4.0 | 4.7 | 3.5 | 2.4 |
| 7 | 1.2 | 4.6 | 4.7 | 3.1 |
| 8 | 1.6 | 2.3 | 3.2 | 2.2 |
| 9 | 4.7 | 3.8 | 2.9 | 4.1 |
| 10 | 1.1 | 3.9 | 3.2 | 1.4 |
| 11 | 2.4 | 4.2 | 2.7 | 1.8 |
| 12 | 4.1 | 2.6 | 3.9 | 2.6 |
| 13 | 4.5 | 4.1 | 2.4 | 3.9 |
| 14 | 1.9 | 3.9 | 2.7 | 4.7 |

Die Nutzer*innen werden durch Punkte in der Ebene repräsentiert, wobei die erste Koordinate ihre durchschnittliche Punktzahl für Action-Serien und die zweite Koordinate ihre durchschnittliche Punktzahl für Drama-Serien angibt. Sie werden dann in fünf Cluster unterteilt, die wir A, B, C, D und E nennen. Die Cluster A, B und C enthalten jeweils drei Punkte und ihre Clusterzentren sind $(2.0; 4.5)$, $(4.7; 4.1)$ bzw. $(4.1; 2.6)$, wobei auf das erste Komma gerundet wurde.

(a) Gib für jedes der Cluster A, B und C an, welche Personen dazu gehören.

(b) Ein 15. Nutzer hat Action-Serien durchschnittlich 4.0 Sterne und Drama-Serien durchschnittlich 4.5 Sterne gegeben. Zu welchem Cluster sollte er hinzugefügt werden? Würdest du ihm eher eine Science-Fiction-Serie oder eine Comedy-Serie empfehlen?

(c) Eine 16. Nutzerin hat Action-Serien durchschnittlich 1.3 Sterne und Drama-Serien durchschnittlich 4.3 Sterne gegeben. Welchem Cluster

sollte sie hinzugefügt werden? Würdest du ihr eher eine Science-Fiction-Serie oder eine Comedy-Serie empfehlen?

(d) Ein 17. Nutzer hat ebenfalls genügend Bewertungen abgegeben, um in die Liste aufgenommen zu werden. Er wird auf der Grundlage seiner Bewertungen zu Cluster A hinzugefügt, wodurch sich das Clusterzentrum auf $(1.8; 4.4)$ verschiebt. Wie hoch sind seine durchschnittlichen Beurteilungen für Action- und Dramaserien?