Internationales Mathe-Turnier 2021

Vorbereitungsmaterial

7. September 2021

Inhaltsverzeichnis

1 Einleitung

Vielen Menschen begegnen täglich Angebotssysteme. Immer, wenn man auf Youtube ein paar interessante Videos angeschaut hat, empfiehlt das Programm weitere Videos. Oder wenn man auf Netflix seine Lieblingsserie geguckt hat, stellt das Programm weitere Serien vor, die man auch anschauen sollte. Auch andere Firmen wie Spotify benutzen solche Systeme. All das geschieht mit einem einzigen Ziel: Man soll länger auf der Seite verweilen und dadurch den Gewinn der Programmbetreiber*innen steigern.

Wieviel Geld damit zu verdienen ist, wurde etwa im Jahr 2006 deutlich. Netflix lobte einen Betrag von 1 Million Dollar an diejenige Person aus, die das Angebotssystem von Netflix um mindestens 10% verbessert. Im Kapitel 7 werden wir darüber mehr erfahren.

Aber wie arbeitet ein solches Angebotssystem eigentlich? Wir werden im Vorbereitungsmaterial darüber Auskunft geben. Zuerst behandeln wir den Satz von Bayes. Dabei werden wir versuchen, Vorhersagen zu machen, die auf vergangenen Daten basieren. Ebenso werden Vorhersagen aus k-Nachbarschaften erarbeitet.
In Kapitel 5 werden wir mit Hilfe von Formeln eine Vorhersage darüber versuchen, welche Bewertung jemand zum Beispiel für einen Film abgeben wird.

Für das Vorbereitungsmaterial haben wir verschiedene Quellen benutzt. Da ist zuerst das Buch *Recommender systems* von Charu C. Aggarwal[1], wobei wir vornehmlich auf dessen Kapitel 3 über die Bayes-Methode zurückgegriffen haben. Aus dem Buch *Social Media Mining* von Reza Zafarani cs.[2] haben wir Anregungen zum Kapitel über gemeinsame Filterung entnommen. Für die k-Clustermittelung haben wir den Essay *Computationeel denken bij vwo wiskunde* A[3] herangezogen. Schließlich haben wir für das letzte Kapitel über Netflix auf die offizielle Seite für den Netflix-Preis [4] zugegriffen.

[1] Recommended Systems, ISBN 978-3-319-29657-9, DOI 10.1007/978-3-319-29659-3
[2] Gratis einsehbar über die Website: https://dmml.asu.edu/smm
[3] https://essay.utwente.nl/79806/1/OvO%20Jelle%20Neeft.pdf
[4] Site Netflix Prize: https://www.netflixprize.com/index.html

2 Grundbegriffe

2.1 Das Summenzeichen

In der Mathematik benutzt man ein spezielles Zeichen zur Angabe von Summen, nämlich '$\sum$'. Es erlaubt, fortgesetzte Additionen kompakt und übersichtlich zu notieren. Das kann dann zum Beispiel so aussehen:

$$\sum_{i=s}^{n} x_i = x_s + x_{s+1} + ... + x_{n-1} + x_n$$

Hier ist jedes der x_i eine gegebene Zahl. Die Summe startet bei s, und deswegen heißt s die Untergrenze der Summation. Die Summation läuft bis einschließlich n, der Obergrenze der Summation. Die Zahl i durchläuft all diese Werte; i heißt Summationsindex. Die Summation beginnt also mit dem Summanden x_i, der zu $i = s$ gehört; jeder folgende zu addierende Wert gehört zu einem um 1 höheren Index, bis der Index den Wert $i = n$ erreicht; dann ist die Summe fertig.

Beispiel 2.1 Angenommen wir haben $x_i = i^2$, und die Summe läuft von 1 bis einschließlich 3. Dann schreiben wir das wie folgt:

$$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$$

Der Index i ist eine sogenannte Dummy-Variable. Das bedeutet, dass wir stattdessen irgendeinen anderen Variablennamen wie k oder j benutzen dürfen. Diesen kann man sich passend aussuchen.

In der Summe muss es nicht der Fall sein, das jeder dort auftretende Summand den Index überhaupt enthält:

Beispiel 2.2 Angenommen dass wir wollen die Summe der $k + 2$ berechnen, in der k von 2 bis 5 läuft. Dann erhalten wir Folgendes:

$$\sum_{k=2}^{5}(k + 2) = (2 + 2) + (3 + 2) + (4 + 2) + (5 + 2) = 4 + 5 + 6 + 7 = 22$$

Aufgabe 1 Berechne die folgenden Summen:

a) $\sum_{i=16}^{20} \frac{i}{4}$

b) $\sum_{k=1}^{4} \frac{3k}{2}$

c) $\sum_{i=2}^{5}(6i + i^3)$

2.2 Matrizen

Eine Matrix (Mehrzahl: Matrizen) ist ein rechteckiges Schema von Zahlen und hat viel mit einer Tabelle gemeinsam. Der Vorteil einer Matrix ist, dass man in ihr einfacher rechnen kann (das Rechnen mit mehreren Matrizen zugleich gehört nicht zum Vorbereitungsmaterial). Hier folgt nun ein Beispiel für eine Matrix:

$$A = \begin{bmatrix} 3 & 4 & 1 \\ 6 & 8 & 2 \end{bmatrix}$$

In unseren Unterlagen benutzen wir stets rechteckige Klammern. In anderen Skripten werden auch runde Klammern verwendet.

Statt der Matrix den Namen A zu geben, können wir auch die Zeilen und Spalten benennen. In der folgenden Matrix geht es um Karen, Kirsten und Katharina, die Spiele ausgetragen haben (nicht unbedingt gegeneinander).

	Gewonnen	Verloren	Unentschieden
Karen	6	4	2
Kirsten	7	4	1
Katharina	4	5	3

2.2.1 Matrixwerte ablesen

Aus einer Matrix können wir die eingetragenen Werte ablesen.

Definition 2.3 Wenn wir einen Eintrag aus einer Matrix ablesen wollen, notieren wir das mit dem Buchstaben r und dann den Angaben, welchen Wert wir ablesen wollen: $r_{Reihe,Spalte}$. Wenn wir im obenstehenden Beispiel etwa ablesen wollen, wie viele Spiele Karin gewonnen hat, bezeichnen wir diesen Wert mit $r_{Karen,Gewonnen}$.

Beispiel 2.4 Wir wollen die Werte von $r_{Karen,Gewonnen}$ und $r_{Kirsten,Unentschieden}$ ablesen. Um den Wert von $r_{Karen,Gewonnen}$ zu ermitteln, schauen wir in der Zeile von Karen und in der Spalte von Gewonnen, welche Zahl dort steht. Dann sehen wir, dass

$$r_{Karen,Gewonnen} = 6 \qquad \text{und ebenso:} \qquad r_{Kirsten,Unentschieden} = 1.$$

2.2.2 Zeilen- und Spaltenmittel

Den Mittelwert (kurz das Mittel) einer Zeile oder einer Spalte einer Matrix kann viele Bedeutungen haben. Denke beispielsweise an deine Durchschnittsnote auf dem letzten Zeugnis oder den Durchschnittswert eines Tests. Den Mittelwert einer Matrixzeile oder -spalte rechnen wir wie folgt aus:

Definition 2.5 Angenommen wir wollen das Mittel der Spalte 'Gewonnen' ausrechnen; das notieren wir als $\overline{r}_{Gewonnen}$ und berechnen wir, indem wir alle Werte der

Spalte 'Gewonnen' addieren und die Summe durch die Anzahl der Werte (hier: 3) dividieren.

Beispiel 2.6 Um $\overline{r}_{Gewonnen}$ auszurechnen, berechnen wir das Mittel der ersten Spalte; für $\overline{r}_{Verloren}$ machen wir das Gleiche für die zweite Spalte.

$$\overline{r}_{Gewonnen} = \frac{6+7+4}{3} \approx 5,67$$
$$\overline{r}_{Verloren} = \frac{4+4+5}{3} \approx 4,33$$

So kann man auch den Mittelwert einer Zeile bestimmen, beispielsweise den von Katharina:

$$\overline{r}_{Katharina} = \frac{4+5+3}{3} = 4$$

Aufgabe 2 Rechne $\overline{r}_{Karen}$ und $\overline{r}_{Unentschieden}$ aus.

3 Voraussagen mit der Bayes-Methode

Eine erste Methode, um Voraussagen zu treffen, verwendet die Bayes-Methode. Bevor wir sie studieren können, müssen wir erst die Aussage des Satzes von Bayes kennen, und dafür brauchen wir den Begriff der bedingten Wahrscheinlichkeit.

3.1 Bedingte Wahrscheinlichkeiten

Ehe wir uns bedingten Wahrscheinlichkeiten zuwenden, machen wir uns noch einmal klar, was Wahrscheinlichkeiten eigentlich sind. Angenommen wir haben einen Würfel und es wird gefragt, mit welcher Wahrscheinlichkeit man 6 Augen würfelt. Dann würden wir sagen, dass $P(6 \text{ Augen}) = 1/6$. Das P steht für das englische Wort probability (=Wahrscheinlichkeit). Wenn X das Ereignis bezeichnet, dass man eine 6 würfelt, kann man noch kürzer $P(X) = 1/6$ schreiben. So werden wir im folgenden Text Wahrscheinlichkeiten notieren.

Wir sind bisher davon ausgegangen, dass wir wissen wollen, mit welcher Wahrscheinlichkeit **ein** Ereignis auftritt. Es können aber auch mehrere Ereignisse $X_1, ..., X_n$ eine Rolle spielen und wir möchten wissen, mit welcher Wahrscheinlichkeit sie alle zugleich auftreten. Diese Wahrscheinlichkeit notieren wir als $P(X_1, X_2, ..., X_n)$.

Wir notieren die bedingte Wahrscheinlichkeit als $P(X|Y)$; dies ist die Wahrscheinlichkeit, dass Ereignis X auftritt unter der Voraussetzung, dass Y eingetreten ist. Wir sprechen von der Wahrscheinlichkeit für X unter der Bedingung Y.

Definition 3.1 Die bedingte Wahrscheinlichkeit $P(X|Y)$ ist die Wahrscheinlichkeit, dass X eintritt unter der Bedingung, dass das Ereignis Y bereits eingetreten ist. Die zugehörige Gleichung dafür ist:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Beispiel 3.2 Angenommen, vor uns liegt ein vollständiges Kartenspiel, also 52 Karten. Wir ziehen eine Karte und achten dabei auf folgende Ereignisse:
X: die gezogene Karte ist eine Herzkarte
Y: die gezogene Karte zeigt einen König.

Um nun $P(X|Y)$ zu berechnen, können wir die Definition der bedingten Wahrschienlichkeit nutzen. Also müssen wir $P(X,Y)$ und $P(Y)$ berechnen. Wir beginnen mit $P(Y)$, also der Wahrscheinlichkeit für einen König. Insgesamt haben wir 52 Karten, darunter 4 Könige, so dass 4 Karten den Forderungen von Y nachkommen; wir sprechen von 4 günstigen Ergebnissen für Y. Wir wissen, dass stets $P(Y) = \frac{\text{Anzahl günstiger Ergebnisse}}{\text{Anzahl möglicher Ergebnisse}}$; so ergibt sich $P(Y) = \frac{4}{52}$.

$P(X,Y)$ ist also gleich $P(X \text{ und } Y)$, also gleich der Wahrscheichlichkeit, dass die Ereignisse X und Y beide zugleich eintreten.

Im Beispiel bedeutet das, dass die Ereignisse 'eine Herz-Karte' und 'ein König' beide eintreten. Immer noch haben wir 52 Karten und nur 1 Karte, der Herz-König, erfüllt gleichzeitig beide Bedingungen. Das bedeutet $P(X,Y) = \frac{1}{52}$ und somit:

$$P(X|Y) = \frac{\frac{1}{52}}{\frac{4}{52}} = \frac{1}{4}$$

Wir sind bisher davon ausgegangen, dass X nur unter der Bedingung eines einzigen Ereignisses Y auftreten soll, aber das können wir auf die Voraussetzung mehrerer Ereignisse verallgemeinern:

$$P(X_n|X_1, X_2, \ldots, X_{n-1}) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_1, X_2, \ldots, X_{n-1})}$$

Aufgabe 3 Berechne mit den Bezeichnungen aus Beispiel 3.2 die Wahrscheinlichkeit $P(Y|X)$.

Aufgabe 4 Bei der Befragung von 240 Studierenden geben 113 an, dass sie Mathematik studieren und 89, dass sie ein naturwissenschaftliches Fach studieren. 37 Studierende studieren beide Fächer, also 76 Studierende nur Mathematik. Wir definieren folgende Ereignisse:
W: Der*die Studierende studiert Mathematik
N: Der*die Studierende studiert ein naturwissenschaftliches Fach
Berechne $P(W|N)$ und $P(N|W)$

3.2 Der Satz von Bayes

$P(X,Y)$ ist die Wahrscheinlichkeit, dass X und Y beide eintreten. Aus $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ und $P(Y|X) = \frac{P(X,Y)}{P(X)}$ folgt $P(X|Y) \cdot P(Y) = P(X,Y) = P(Y|X) \cdot P(X)$

Definition 3.3 Der Satz von Bayes gibt uns die folgende Gleichung:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

Dieser Satz ist sehr nützlich, um bedingte Wahrscheinlichkeiten auszurechnen. Man dreht hier nämlich die Rolle der Bedingung und des Ereignisses um. Das sorgt dafür, dass man bestimmte bedingte Wahrscheinlichkeiten auch dann berechnen kann, wenn bestimmte Größen nicht gegeben sind.

Beispiel 3.4 Angenommen wir betrachten die folgenden Ereignisse:
X: Ein*e Patient*in hat Krebs.
Y: Ein*e Patient*in ist ein*e Raucher*in

Wir wollen nun die Wahrscheinlichkeit berechnen, dass ein*e Patient*in der Uniklinik Krebs hat, wenn wir nur wissen, dass er*sie Raucher*in ist. Wenn wir das direkt aus der Definition der bedingten Wahrscheinlichkeit berechnen wollen, erscheint das sehr schwierig, aber mit dem Satz von Bayes wird die Antwort zugänglich. Dann müssen wir nämlich nur andere Wahrscheinlichkeiten kennen, die wir aus dem Datenbestand der Uniklinik erfahren könnten.
P(X) können wir erfragen, indem wir fragen, mit welcher Wahrscheinlichkeit ein*e Patient*in der Uniklink Krebs hatte. Diese wird beispielsweise als 0,075 angegeben.
P(Y) können wir auch ermitteln, denn das ist die Wahrscheinlichkeit, dass ein*e Patient*in der Uniklinik Raucher*in war. Die Angabe dieser Wahrscheinlichkeit ergibt 0,25.
$P(Y|X)$ ist die Wahrscheinlichkeit, dass ein*e Krebspatient*in Raucher*in war; auch diese können wir ermitteln: 0,5.

Dann gilt:

$$P(X|Y) = \frac{0,075 \cdot 0,5}{0,25} = 0,15$$

Das ergibt die Aussage, dass sich die Wahrscheinlichkeit, an Krebs zu erkranken, für Raucher*innen verdoppelt.

Wir sehen hier auch, warum der Satz von Bayes so nutzbringend ist - und was letztlich der Grund ist, weswegen wir ihn häufig benutzen werden.

Aufgabe 5 In der Werner-Beinhart-Schule fährt jedes Jahr eine Schüler*innengruppe während der Weihnachtsferien zum Wintersport. In diesem Jahr hatte die Schule 1000 Schüler*innen, von denen 20% an der Tour teilgenommen haben; 1% dieser Wintersportler*innen kommen mit einem gebrochenen Bein von der Reise zurück. Im Moment gibt es auf der Schule insgesamt 4 Schüler*innen, die sich ein Bein gebrochen haben.
Berechne die Wahrscheinlichkeit, dass ein*e Schüler*in mit einem gebrochenen Bein an der Wintersporttour teilgenommen hat!

Der Satz von Bayes ist auch bekannt unter der Formulierung

$$P(X|Y) \propto P(Y|X) \cdot P(X).$$

Dabei bedeutet '$\propto$' 'ist proportional zu'; wenn sich also $P(X|Y)$ verdoppelt, dann wird $P(Y|X) \cdot P(X)$ ebenfalls zweimal so groß.

3.3 Vorhersagen mit Hilfe des Satzes von Bayes

Wir könnten uns für die Ereignisse $X_1, \ldots, X_n$ zum Beispiel lauter Filme vorstellen. Diese Filme lassen wir durch verschiedene Teilnehmer*innen $1, \ldots, s$ beurteilen; die Teilnehmer*innen können die Filme nur mit 0 oder 1 beurteilen.
Von einer neu hinzugekommenen Person $s+1$ wissen wir nur ihre Beurteilungen der Filme $X_1, \ldots, X_{n-1}$ und wollen die Beurteilung des Films X_n vorhersagen. Für jede der möglichen Beurteilungen 0 und 1 bestimmen wir die bedingte Wahrscheinlichkeit.

Beispiel 3.5 Von drei Teilnehmer*innen kennen wir ihre Beurteilung von drei verschiedenen Filmen; sie konnten die Filme nur mit 0 oder 1 beurteilen. Von einer vierten Person kennen wir die Beurteilung von Film 1 und Film 2 (siehe Tabelle 1). Wir berechnen nun die bedingten Wahrscheinlichkeiten für die möglichen Beurteilungen von Film 3.

Teilnehmer*in	Film 1	Film 2	Film 3
1	1	1	0
2	0	0	1
3	1	0	0
4	1	1	?

Tabelle 1: Beurteilung von Filmen.

Wir erhalten dann folgende Rechnung:

$$\begin{aligned} &P(\text{Film } 3 = 0 | \text{Film } 1 = 1, \text{Film } 2 = 1) \\ &= \frac{P(\text{Film } 3 = 0) \cdot P(\text{Film } 1 = \text{Film } 2 = 1 | \text{Film } 3 = 0)}{P(\text{Film } 1 = 1, \text{Film } 2 = 1)} \\ &\approx \frac{P(\text{Film } 3 = 0) \cdot P(\text{Film } 1 = 1 | \text{Film } 3 = 0) \cdot P(\text{Film } 2 = 1 | \text{Film } 3 = 0)}{P(\text{Film } 1 = 1, \text{Film } 2 = 1)} \end{aligned}$$

Oder:

$$
\begin{aligned}
&P(\text{Film } 3 = 1 | \text{Film } 1 = 1, \text{Film } 2 = 1) \\
&= \frac{P(\text{Film } 3 = 1) \cdot P(\text{Film } 1 = \text{Film } 2 = 1 | \text{Film } 3 = 1)}{P(\text{Film } 1 = 1, \text{Film } 2 = 1)} \\
&\approx \frac{P(\text{Film } 3 = 1) \cdot P(\text{Film } 1 = 1 | \text{Film } 3 = 1) \cdot P(\text{Film } 2 = 1 | \text{Film } 3 = 1)}{P(\text{Film } 1 = 1, \text{Film } 2 = 1)}
\end{aligned}
$$

In der letzten Zeile haben wir jeweils eine Näherung für eine bedingte Wahrscheinlichkeit genutzt, die wir auch zukünftig verwenden werden: Wir berechnen die bedingte gemeinsame Wahrscheinlichkeit von zwei Ereignissen als Produkt der bedingten Wahrscheinlichkeiten der beiden einzelnen Ereignisse. Nur durch diese Näherung können wir zu einer Vorhersage kommen; eine direkte Berechnung von bedingten gemeinsamen Wahrscheinlichkeiten ist oft nicht möglich.

Unser Ziel ist es, herauszufinden, welche der Beurteilungen 0 oder 1 für den Film 3 durch die vierte Person wahrscheinlicher ist. Wir sehen, dass in beiden Termen die Nenner gleich sind; sie spielen keine Rolle bei der Vorhersage, welche Beurteilung wahrscheinlicher ist. Von jetzt an werden wir beim Erstellen von Vorhersagen darum die Nenner außer Betracht lassen.

Definition 3.6 Wenn wir eine Angebotsempfehlung mit der Bayes-Methode unterbreiten, dann benutzen wir die Proportionalitätsformulierung des Satzes von Bayes:

$$P(X|Y) \propto P(Y|X) \cdot P(X)$$

Ausgehend von mehreren Ereignissen $Y_1, ..., Y_n$ erhalten wir die Proportionalitätsformulierung

$$P(X|Y_1, Y_2, \ldots, Y_n) \propto P(Y_1|X) \cdot P(Y_2|X) \cdots P(Y_n|X) \cdot P(X)$$

Wir verwenden im Folgenden die proportionale Formulierung des Satzes von Bayes; also werden die Produkte auf der rechten Seite keine tatsächlichen Wahrscheinlichkeiten liefern, sondern nur relative Angaben. Wir können also nicht errechnen, wie groß die Wahrscheinlichkeiten tatsächlich sind, dass bestimmte Beurteilungen auftreten werden. Wir können aber wohl deren Verhältnisse bestimmen. Aus diesem Verhältnis können wir dann festlegen, ob es plausibler ist, für den Film 3 die Beurteilung 0 oder die Beurteilung 1 zu erwarten - und das wird dann unsere Vorhersage.

Beispiel 3.7 Wir schauen nochmal auf Tabelle 1 aus Beispiel 3.5. Mit der Bayes-Methode berechnen wir die relative Wahrscheinlichkeit, dass Film 3 von Person 4 die Beurteilung 0 bzw. 1 bekommt.

Für die Beurteilung 0 für Film 3 berechnen wir die benötigten Wahrscheinlichkeiten erst getrennt aus:

$$P(\text{Film } 3 = 0) = \frac{\text{Anzahl der Teilnehmer*innen, die Film 3 mit 0 beurteilt haben}}{\text{Anzahl der Teilnehmer*innen, die Film 3 beurteilt haben}} = \frac{2}{3}$$

$$P(\text{Film } 1 = 1 | \text{Film } 3 = 0) = \frac{2}{2}$$

$$P(\text{Film } 2 = 1 | \text{Film } 3 = 0) = \frac{1}{2}$$

Dies übernehmen wir nun in den Satz von Bayes und erhalten ein relatives Maß für die Wahrscheinlichkeit, dass Film 3 die Bewertung 0 erhält:

$$P(\text{Film } 3 = 0 | \text{Film } 1 = 1, \text{Film } 2 = 1) \propto \frac{2}{3} \cdot \frac{2}{2} \cdot \frac{1}{2} = \frac{4}{12} \approx 0,333$$

Für die Beurteilung des Films 3 durch die 1 erhalten wir analog:

$$P(\text{Film } 3 = 1) = \frac{1}{3}$$

$$P(\text{Film } 1 = 1 | \text{Film } 3 = 1) = \frac{0}{1}$$

$$P(\text{Film } 2 = 1 | \text{Film } 3 = 1) = \frac{0}{1}$$

Dies übernehmen wir nun wieder in den Satz von Bayes und erhalten ein relatives Maß für die Wahrscheinlichkeit, dass Film 3 die Bewertung 1 erhält:

$$P(\text{Film } 3 = 1 | \text{Film } 1 = 1, \text{Film } 2 = 1) \propto \frac{1}{3} \cdot \frac{0}{1} \cdot \frac{0}{1} = 0$$

Wir schließen, dass Film 3 von der Person 4 wahrscheinlich mit 0 beurteilt werden wird.

Aufgabe 6 Betrachte Tabelle 2. Berechne mit Hilfe der Bayes-Methode eine Voraussage, ob Teilnehmer*in 3 dem Lied 1 die Beurteilung -1 oder 1 geben wird.

Teilnehmer	Lied 1	Lied 2	Lied 3	Lied 4	Lied 5	Lied 6
1	1	-1	1	-1	1	-1
2	1	1	?	-1	-1	-1
3	?	1	1	-1	-1	?
4	-1	-1	-1	1	1	1
5	-1	?	-1	1	1	1

Tabelle 2: Beurteilung von Liedern.

Aufgabe 7 Sag aus der Tabelle 2 mit der Bayes-Methode voraus, ob Teilnehmer*in 5 dem Lied 2 die Beurteilung -1 oder 1 geben wird.

Tag	Vorhersage	Temperatur	Luftfeuchtigkeit	Wind	Wird Tennis gespielt?
1	sonnig	warm	hoch	schwach	Nein
2	sonnig	warm	hoch	stark	Nein
3	bewölkt	warm	hoch	schwach	Ja
4	Regen	mittel	hoch	schwach	Ja
5	Regen	kalt	normal	schwach	Ja
6	Regen	kalt	normal	stark	Nein
7	bewölkt	kalt	normal	stark	Ja
8	sonnig	mittel	hoch	schwach	Nein
9	sonnig	kalt	normal	schwach	Ja
10	Regen	mittel	normal	schwach	Ja
11	sonnig	mittel	normal	stark	Ja
12	bewölkt	mittel	hoch	stark	Ja
13	bewölkt	warm	normal	schwach	Ja
14	Regen	mittel	hoch	stark	Nein
15	sonnig	kalt	hoch	stark	?
16	bewölkt	kalt	hoch	stark	?
17	bewölkt	mittel	hoch	schwach	?

Tabelle 3: Übersicht, ob unter bestimmten Wetterbedingungen Tennis gespielt wurde

Aufgabe 8 Wir haben zwei Wochen lang notiert, welche Temperatur herrscht, wie hoch die Luftfeuchtigkeit ist und ob es windig ist. Und an diesen Tagen haben wir auch die Wettervorhersage notiert. Dann haben wir geschaut, an welchen Tagen Tennis gespielt wurde – und an welchen Tagen nicht. Alle diese Daten sind in der Tabelle 3 aufgeführt. Erstelle mit der Bayes-Methode eine Vorhersage, ob am Tag 15 Tennis gespielt wird oder nicht.

3.4 Laplace-Glättung

Wenn wir in der Tabelle 3 für den 16ten Tag vorhersagen wollen, ob wohl Tennis gespielt wird oder nicht, dann ergeben sich die folgenden bedingten Wahrscheinlichkeiten:

$$P(\text{Tennis} = \text{Ja}|\text{bewölkt, kalt, hoch, stark}) \propto \frac{9}{14} \cdot \frac{4}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \approx 0,011$$

$$P(\text{Tennis} = \text{Nein}|\text{bewölkt, kalt, hoch, stark}) \propto \frac{5}{14} \cdot \frac{0}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} = 0$$

Diese Gleichungen suggerieren, dass bei bewölktem Wetter stets Tennis gespielt wird, und zwar unabhängig von der Temperatur, der Luftfeuchtigkeit und dem Wind. Aber dieser Schluss ist sicher nicht gerechtfertigt, weil wir sehen, dass auch

an regnerischen Tagen gespielt wird - und dann ist es ja ebenfalls bewölkt. Um das Problem mit der verschwindenden Wahrscheinlichkeit aufzulösen, benutzen wir die Methode der Laplace-Glättung. Dabei wird die Wahrscheinlichkeit durch einen Glättungsparameter α angepasst.

Angenommen, $q_1, ..., q_s$ seien die Teilnehmer*innen, die eine Beurteilung für ein Item (dem wir hier die willkürliche Nummer 2 geben) abgeben konnten; sie konnten eine der Beurteilungen $v_1, ..., v_l$ abgeben. Dann ergibt sich:

$$P(\text{Item } 2 = v_1) = \frac{\text{Anzahl der Beurteilungen von Item 2 mit } v_1}{\text{Gesamtzahl aller Beurteilungen}}$$

Definition 3.8 Mit der Laplace-Glättung erhalten wir den folgenden Ausdruck:

$$P(\text{Item } 2 = v_1) = \frac{\text{Anzahl der Beurteilungen von Item 2 mit } v_1 + \alpha}{\text{Gesamtzahl aller Beurteilungen} + l \cdot \alpha}$$

Die gleiche Annäherung kann auch benutzt werden, um die bedingten Wahrscheinlichkeiten abzuschätzen: wir fügen den Summanden α zum Zähler und den Summanden $l \cdot \alpha$ zum Nenner hinzu.

Definition 3.9 Die bedingte Wahrscheinlichkeit mit Laplace-Glättung sieht dann wie folgt aus:

$$P(\text{Item } 2 = v_1 | \text{Item } 1 = v_2)$$

$$= \frac{\text{Anzahl der Teilnehmer*innen, die Item 1 mit } v_2 \text{ beurteilen und Item 2 mit } v_1 + \alpha}{\text{Anzahl der Teilnehmer*innen, die Item 2 mit } v_2 \text{ beurteilen} + l \cdot \alpha}$$

Wenn wir die Laplace-Glättung anwenden, dann entscheiden wir uns dafür, das nur dann zu tun, wenn die ungeglättete Berechnung schiefgeht, wir also verschwindende Wahrscheinlichkeiten erhalten. Wir könnten uns auch dafür entscheiden, die Glättung konsequent auf alle Terme $P(X|Y)$ anzuwenden. Bei der Berechnung mit Hilfe eines Computers macht das die Sache einfacher, denn es spart viel Zeit: man muss nicht erst untersuchen, wo etwas nicht funktioniert, sondern wendet die Glättung überall an. Der Nachteil ist, dass man die Glättung vielleicht anwendet, obwohl das an keiner Stelle notwendig gewesen wäre. Daher wollen wir erst schauen, wo etwas schiefgeht und dann an diesen Stellen die Glättung anwenden. Das bedeutet im obenstehenden Fall, dass wir für die Vorhersage, ob Tennis gespielt werden wird oder nicht, eine Korrektur anbringen müssen. Wie das genau abläuft, zeigt das folgende Beispiel 3.10.

Beispiel 3.10 Angenommen, wir gehen von $\alpha = 1$ aus. Dann können wir mit der Laplace-Glättung vorhersagen, ob am Tag 16 Tennis gespielt wird. Wir gehen wieder von der Tabelle 3 aus.

Wir müssen die Laplace-Glättung auf die Wettervorhersage anwenden. Diese konnten wir auf drei verschiedene Arten treffen: Regen, bewölkt oder sonnig. Also ist $l = 3$. Das liefert dann die folgenden Wahrscheinlichkeiten:

Die Wahrscheinlichkeit, dass am Tag 16 nicht gespielt wird, beträgt:

$$P(\text{Tennis} = \text{Nein}|\text{bewölkt, kalt, hoch, stark}) \propto \frac{5}{14} \cdot \frac{0+1}{5+3} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \approx 0,0043$$

Und die Wahrscheinlichkeit, dass am Tag 16 wohl gespielt wird, beträgt:

$$P(\text{Tennis} = \text{Ja}|\text{bewölkt, kalt, hoch, stark}) \propto \frac{9}{14} \cdot \frac{4+1}{9+3} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \approx 0,0099$$

Uns interessiert das Wahrscheinlichkeitsverhältnis von 'Es wird nicht gespielt' : 'Es wird gespielt'. Das ergibt sich nun zu 0,0043:0,0099 oder 1:2,3. Wir würden nun schließen, dass unter den gegebenen Umständen am Tag 16 doch Tennis gespielt werden wird.

Wir haben in diesem Beispiel vorgegeben, dass $\alpha = 1$ ist. Nun nehmen wir an, wir hätten $\alpha = 1000$ vorgegeben; unter dieser veränderten Vorgabe schauen wir uns den Schritt bei der Anwendung der Laplace-Glättung an und erhalten Folgendes:

Wenn $\alpha = 1$,

$$P(\text{bewölkt}|\text{Tennis} = \text{Nein}) = \frac{0+1}{5+3} \approx 0,125$$

$$P(\text{bewölkt}|\text{Tennis} = \text{Ja}) = \frac{4+1}{9+3} \approx 0,417$$

Wenn $\alpha = 1000$,

$$P(\text{bewölkt}|\text{Tennis} = \text{Nein}) = \frac{0+1000}{5+3000} \approx 0,333$$

$$P(\text{bewölkt}|\text{Tennis} = \text{Ja}) = \frac{4+1000}{9+3000} \approx 0,334$$

Wir sehen, dass große Werte von α zu einer Annäherung der Wahrscheinlichkeiten führen; wir erhalten also kaum noch Informationen, die wir für eine Vorhersage nutzen können. Dann wird die Laplace-Näherung sinnlos. Darum ist es üblich, mit $\alpha = 1$ zu arbeiten: so eliminiert man die verschwindenden Wahrscheinlichkeiten und erhält doch genug Informationen für die Erstellung einer Vorhersage.

Aufgabe 9 Erstelle mit der Laplace-Näherung eine Vorhersage, ob am Tag 17 Tennis gespielt werden wird oder nicht. Nutze die Vorgaben der Tabelle 3 und gehe von $\alpha = 1$ aus.

4 Vorhersagen aus k-Nachbarschaften

Eine andere Art, um zu Vorhersagen zu kommen, besteht darin, dass man die k nächstliegenden Nachbarn anschaut. Wir können diese k-Nachbarschaft aus den vorliegenden Daten der anderen Teilnehmer*innen bestimmen: man sucht also die k Personen, die die meisten Übereinstimmungen mit der neu hinzugekommenen Person aufweisen. Dabei kann es wieder um Filme gehen, aber wir können auch andere Eigenschaften in Betracht ziehen. Man kann bei der Bestimmung der k Nachbarn auch bestimmte Eigenschaften stärker gewichten: sie werden dann bei der Beurteilung der Nähe wichtiger als andere Eigenschaften. Die Formel, mit der man die Nähe zweier Teilnehmer a und b bestimmt, lautet:

$$\sum_{i=1}^{n} \frac{w_i \cdot \partial(a_i, b_i)}{\sum_i w_i}.$$

Dabei gibt n die Anzahl der Eigenschaften an, und der Summationsindex i durchläuft alle Items, denen für jede Person die dazugehörigen Eigenschaften zugeordnet werden. w_i ist das Gewicht, das der Eigenschaft mit dem Index i zugeordnet ist. In der Formel steht a_i für die Eigenschaft der neuen Person a hinsichtlich des Items i und b_i für die Eigenschaft der anderen Person b hinsichtlich des Items i. Es gilt $\partial(a_i, b_i) = 1$ genau dann, wenn die Person a und die Person b in der Eigenschaft i übereinstimmen, und in allen anderen Fällen ist $\partial(a_i, b_i) = 0$. $\sum_i w_i$ ist die Summe aller verteilten Gewichte.

Diese Formel liefert stets ein Ergebnis zwischen 0 und 1. Anders gesagt: Das Maß für die Nähe liegt immer zwischen 0 und 1. Der kleinste Wert 0 bedeutet, dass es keine Ähnlichkeit zwischen den Teilnehmern a und b gibt, und der Wert 1 signalisiert die Übereinstimmung der beiden Personen in allen relevanten Eigenschaften.

In der Folge kann man mit Hilfe dieses Maßes eine Prognose erstellen: Man betrachtet nämlich die k Personen, die die größte Nähe zur neuen Person aufweisen; dann schaut man auf die Beurteilungen dieser k nächsten Nachbarn.

Definition 4.1 Die abgegebene Prognose wird die Beurteilung sein, die die meisten nächsten Nachbarn abgegeben haben. Diese Nachbarn weisen nämlich die größte Ähnlichkeit mit der neuen Person auf, so dass anzunehmen ist, dass sie die gleiche Beurteilung abgeben.

Beispiel 4.2 Von vier verschiedenen Personen wissen wir ihr Geschlecht, wissen, ob sie Schulden haben und ob sie in einer Partnerschaft leben. Weiterhin wissen wir, ob sie ein Haus kaufen wollen. Von Miss Marple wissen wir auch das Geschlecht, ihren Schuldenstand und ihren Partnerschaftsstatus. Nicht bekannt ist, ob sie ein Haus kaufen will. Das wollen wir durch die k-Nachbarschaftsmethode vorhersagen. In Tabelle 4 sind die bekannten Daten zusammengefasst.
Natürlich finden wir den Schuldenstand sehr wichtig und ordnen ihm deswegen das Gewicht 2 zu; der Rest bekommt jeweils das Gewicht 1. Wir erstellen die Prognose

Teilnehmer*in	Geschlecht	Schulden	Partner	Kauft ein Haus
1	Frau	Ja	Nein	Nein
2	Frau	Nein	Nein	Ja
3	Mann	Ja	Ja	Ja
4	Mann	Nein	Ja	Ja
Miss Marple	Frau	Nein	Ja	?

Tabelle 4: Information über Hauskäufe

anhand der beiden Teilnehmer*innen, die am nächsten bei Miss Marples Voraussetzungen liegen. Die Berechnung für Teilnehmer*in 1 sieht wie folgt aus:

Für die Items Geschlecht, Schulden und Partner müssen wir bestimmen, ob diese mit den Gegebenheiten von Miss Marple übereinstimmen oder nicht. Weil Teilnehmer*in 1 und Miss Marple beide weiblich sind und das Gewicht des Items den Wert 1 hat, ergibt sich:

$$w_i \cdot \partial(a_i, b_i) = 1 \cdot \partial(\text{Frau, Frau}) = 1 \cdot 1 = 1.$$

So verfahren wir auch mit den Items Schulden und Partner und erhalten:

$$w_i \cdot \partial(a_i, b_i) = 2 \cdot \partial(\text{Nein, Ja}) = 2 \cdot 0 = 0$$

$$w_i \cdot \partial(a_i, b_i) = 1 \cdot \partial(\text{Ja, Nein}) = 1 \cdot 0 = 0$$

Schließlich folgt dann:

$$\text{Nähe von Teilnehmer*in 1 zu Miss Marple} = \sum_{i=1}^{3} \frac{w_i \cdot \partial(a_i, b_i)}{\sum_i w_i} = \frac{1}{4} + \frac{0}{4} + \frac{0}{4} = \frac{1+0+0}{4} = \frac{1}{4}$$

In der Tabelle 5 stehen dann die Maße für die Nähe aller Teilnehmer*innen zu Miss Marple.

Teilnehmer*in	Kauft ein Haus	Nähe zu Miss Marple
1	Nein	$(1+0+0)/4 = 1/4$
2	Ja	$(1+2+0)/4 = 3/4$
3	Ja	$(0+0+1)/4 = 1/4$
4	Ja	$(0+2+1)/4 = 3/4$

Tabelle 5: Übersicht über die Nähe von Miss Marples Daten zu denen anderer Teilnehmer*innen

Wir sehen, dass die Teilnehmer*innen 2 und 4 die beiden nächstgelegenen Nachbarn von Miss Marple sind, denn das Maß für die Nähe beträgt 3/4. Wir wissen, dass die Teilnehmer*innen 2 und 4 beide planen ein Haus zu kaufen; das macht die Prognose plausibel, dass Miss Marple dies auch tun wird.

Aufgabe 10 Mit der Tabelle 4 aus Beispiel 4.2 können wir auch mit der Bayes-Methode bestimmen, ob Miss Marple wahrscheinlich ein Haus kaufen wird. Benutze hier die Laplace-Glättung mit $\alpha = 1$ und bestimme das Verhältnis der Wahrscheinlichkeiten Miss Marple kauft ein Haus : Miss Marple kauft kein Haus.

Personen	Lebensalter	Einkommen	Schulden	Geschlecht	Kauft einen Computer
1	≤ 30	Hoch	Nein	Mann	Nein
2	≤ 30	Hoch	Nein	Frau	Nein
3	$31 \ldots 40$	Hoch	Nein	Mann	Ja
4	> 40	Mittel	Nein	Mann	Ja
5	> 40	Niedrig	Ja	Mann	Ja
6	> 40	Niedrig	Ja	Frau	Nein
7	$31 \ldots 40$	Niedrig	Ja	Frau	Ja
8	≤ 30	Mitte l	Nein	Mann	Nein
9	≤ 30	Niedrig	Ja	Mann	Ja
10	> 40	Mittel	Ja	Mann	Ja
11	≤ 30	Mittel	Ja	Frau	Ja
12	$31 \ldots 40$	Mitte l	Nein	Frau	Ja
13	$31 \ldots 40$	Hoch	Ja	Mann	Ja
14	> 40	Mittel	Nein	Frau	Nein
Achim	≤ 30	Mittel	Ja	Mann	?
Angela	> 40	Mittel	Ja	Frau	?
Olaf	31...40	Mittel	Nein	Mann	?

Tabelle 6: Übersicht darüber, welche Personen einen Computer gekauft haben

Aufgabe 11 Von 14 Personen haben wir Daten vorliegen. 9 von ihnen haben einen neuen Computer gekauft. ≤ 30 bedeutet, dass die Person höchstens 30 Jahre alt ist, und > 40, dass die Person älter als 40 Jahre ist. Alle Daten sind in der Tabelle 6 zusammengestellt.

a) Bestimme das Wahrscheinlichkeitsverhältnis Achim kauft einen neuen Computer : Achim kauft keinen neuen Computer

b) Betrachte die fünf nächstgelegenen Nachbarn von Achim. Alle Items tragen das Gewicht 1 - außer dem Einkommen, das mit dem Faktor 2 gewichtet wird. Wie wird er sich wahrscheinlich beim Kauf eines Computers entscheiden?

c) Ermittle die sechs nächstgelegenen Nachbarn von Angela; alle Gewichte sind wieder 1 bis auf das Einkommen, das wieder mit 2 gewichtet wird.

d) Prognostiziere mit Aufgabenteil c), ob Angela einen Computer kaufen wird.

e) Prognostiziere unter Nutzung der Laplace-Glättung, ob Olaf einen Computer kaufen wird. Wieder ist $\alpha = 1$ gegeben.

5 Gemeinsame Filter

Durch gemeinsames Filtern können wir Vorhersagen über unbekannte Werte in Tabellen oder Matrizen erstellen. Das kann man auf zwei Arten tun: Man vergleicht Personen miteinander oder man vergleicht Items miteinander. Beide Arten werden hier erklärt, aber vorher müssen wir noch ein paar dazu notwendige Begriffe einführen.

5.1 Begriffe

Aus einer Matrix können wird - wie es in Abschnitt 2.2 schon dargelegt wurde - Informationen wie einzelne Werte, das Zeilen- oder Spaltenmittel auslesen. Nun benötigen wir einige weitere Begriffe zu Matrizen.

5.1.1 Übereinstimmungsmaß

Wenn man eine Kaufempfehlung unterbreiten will, müssen wir das Maß der Übereinstimmungen zwischen zwei Personen betrachten. Im Kapitel 4 wurde bereits ein Maß für die Nähe definiert, aber dieses Maß basiert auf den Beurteilungen etwa eines Films. Solche Beurteilungen werden jedoch meist durch ganze Zahlen zwischen 1 und 5 erfasst, während im Kapitel 4 lediglich die Werte 0 und 1 zulässig waren.

Definition 5.1 Um ein Übereinstimmungsmaß (Englisch: similarity) zwischen zwei Personen oder zwei Gegenständen zu bestimmen, benutzen wir die folgende Formel:

$$Sim(u,v) = \frac{\sum_i r_{u,i} r_{v,i}}{\sqrt{\sum_i (r_{u,i})^2} \sqrt{\sum_i (r_{v,i})^2}},$$

wobei u und v zwei verschiedene Personen sind und i die Items durchläuft. Das gleiche können wir mit zwei Items (zum Beispiel Filmen und Getränken) tun; dann lautet die Formel

$$Sim(i,j) = \frac{\sum_u r_{u,i} r_{u,j}}{\sqrt{\sum_u (r_{u,i})^2} \sqrt{\sum_u (r_{u,j})^2}},$$

wobei i und j für die interessanten Items stehen und u alle Personen durchläuft.

Mit den genannten Formeln bekommen wir ein Übereinstimmungsmaß mit Werten zwischen 0 und 1. Wenn das Maß den Wert 0 hat, dann bedeutet das, dass die zwei Personen wenig oder keine Übereinstimmungen besitzen, während das Maß 1 angibt, dass die zwei Personen das gleiche Getränk gerne mögen oder den gleichen Film gut finden.

Im folgenden Beispiel wird die Definition genauer erklärt:

Beispiel 5.2 Gegeben ist eine Matrix, worin Peter, Hassan und Kim von vier verschiedenen Getränken angeben, wie lecker sie diese finden. Eine 0 bedeutet, dass sie das Getränk nicht mögen, während eine 5 anzeigt, dass sie das Getränk wirklich lecker finden:

$$\begin{array}{r} \\ \text{Peter} \\ \text{Hassan} \\ \text{Kim} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Cola} & \text{Fanta} & \text{Sprite} & \text{Eistee} \end{array} \\ \left[\begin{array}{cccc} 4 & 3 & 1 & 5 \\ 0 & 1 & 5 & 3 \\ 3 & 2 & 3 & 1 \end{array}\right] \end{array}$$

Wir rechnen nun $Sim(Peter, Hassan)$ in ausführlicher Darstellung aus.

Als erstes berechnen wir die Summe im Zähler; das u ersetzen wir durch Peter und das v durch Hassan.

$$\sum_i r_{Peter,i} \cdot r_{Hassan,i}$$

Das i steht in diesem Beispiel für die vier Getränke; diese Summe stellt sich also dar als

$$\begin{array}{lrll} \sum_i r_{Peter,i} \cdot r_{Hassan,i} = & & r_{Peter,Cola} \cdot r_{Hassan,Cola} & + r_{Peter,Fanta} \cdot r_{Hassan,Fanta} \\ & + & r_{Peter,Sprite} \cdot r_{Hassan,Sprite} & + r_{Peter,Eistee} \cdot r_{Hassan,Eistee} \end{array}$$

Alle Werte können wir aus der Matrix ablesen und erhalten:

$$\sum_i r_{Peter,i} \cdot r_{Hassan,i} = 4 \cdot 0 + 3 \cdot 1 + 1 \cdot 5 + 5 \cdot 3 = 23$$

Nun schauen wir uns den Nenner in der Formel an; dort finden wir die Summe

$$\sum_i (r_{u,i})^2 = \sum_i (r_{Peter,i})^2 = (r_{Peter,Cola})^2 + (r_{Peter,Fanta})^2 + (r_{Peter,Sprite})^2 + (r_{Peter,Eistee})^2$$

Auch diese Werte können wir wieder aus der Matrix ablesen:

$$\sum_i (r_{Peter,i})^2 = 4^2 + 3^2 + 1^2 + 5^2 = 16 + 9 + 1 + 25 = 51$$

Auf die gleiche Art berechnen wir die zweite Summe im Nenner:

$$\sum_i (r_{v,i})^2 = \sum_i (r_{Hassan,i})^2 = 0^2 + 1^2 + 5^2 + 3^2 = 0 + 1 + 25 + 9 = 35$$

Da wir nun alle Summen kennen, können wir diese in die Formel einsetzen:

$$Sim(Peter, Hassan) = \frac{\sum_i r_{Peter,i} \cdot r_{Hassan,i}}{\sqrt{\sum_i (r_{Peter,i})^2} \sqrt{\sum_i (r_{Hassan,i})^2}} = \frac{23}{\sqrt{51}\sqrt{35}} = 0,54$$

Aufgabe 12 Berechne aus der Matrix aus dem Beispiel 5.2 die Übereinstimmungsmaße $Sim(Peter, Kim)$ und $Sim(Cola, Fanta)$

5.1.2 Umgebungen

Immer, wenn Netflix einen Algorithmus anwendet, um dir einen Film zu empfehlen, arbeitet das Programm nicht den gesamten Datenbestand aller Netflix-Abonnent*innen durch. Das würde viel zu viel Rechenzeit kosten. Stattdessen beschränkt sich das Programm auf sogenannte Umgebungen.

Definition 5.3 Wenn wir von einer Umgebung einer Person sprechen, dann bezeichnen wir damit die Gruppe der Personen, die das größte Übereinstimmungsmaß mit dieser Person besitzen, für die also $Sim(u, v)$ am größten ist. Die Größe dieser Umgebung kann verschieden sein, wird aber stets durch die Fragestellung angegeben.

Diese Definition wird analog verwendet, wenn es um Umgebungen von Items statt um Umgebungen von Personen geht.

Beispiel 5.4 Wir betrachten erneut die Matrix aus Beispiel 5.2..Wir wollen nun die Umgebung der Größe 2 von Cola berechnen; das bedeutet, dass man die zwei Getränke sucht, die das höchste Übereinstimmungsmaß mit Cola haben.

Um diese beiden Getränke ermitteln zu können, berechnen wir die Übereinstimmungsmaße von Cola mit allen anderen Getränken:

$$Sim(Cola, Fanta) = \frac{4 \cdot 3 + 0 \cdot 1 + 3 \cdot 2}{\sqrt{25}\sqrt{14}} = 0,96$$

$$Sim(Cola, Sprite) = \frac{4 \cdot 1 + 0 \cdot 5 + 3 \cdot 3}{\sqrt{25}\sqrt{35}} = 0,44$$

$$Sim(Cola, Eistee) = \frac{4 \cdot 5 + 0 \cdot 3 + 3 \cdot 1}{\sqrt{25}\sqrt{35}} = 0,78$$

Wir ordnen die Übereinstimmungsmaße von hohen zu niedrigen Werten und stellen fest, dass Fanta und Eistee die beiden Getränke mit den höchsten Übereinstimmungsmaßen mit Cola sind. Da die Umgebungsgröße genau 2 sein soll, bilden diese beiden Getränke die gesuchte Umgebung.

Aufgabe 13 Berechne aus der Matrix von Beispiel 5.2 die Umgebung von Hassan mit der Größe 1.

Es kann vorkommen, dass eine Person einen bestimmten Film noch niemals gesehen oder ein Getränk noch niemals probiert hat. Wir können auf zwei verschiedene Arten berechnen, welche Bewertung eines solchen Items wir von der Person erwarten. Entweder können wird das tun, indem wir auf Personen schauen, die der gegebenen Person möglichst ähnlich sind oder wir können nach Items suchen, die dem interessierenden Item möglichst ähnlich sind. Dabei ist 'ähnlich' immer im Sinne eines hohen Übereinstimmungsmaßes gedacht. Beide Methoden werden im folgenden Abschnitt erklärt.

5.2 Personenbasierte gemeinsame Umgebung

Um festzustellen, ob wir jemandem ein bestimmtes Getränk oder einen bestimmten Film empfehlen würden, können wir andere Personen heranziehen, die dieser Person sehr ähnlich sind. Wenn wir uns dann ansehen, wie diese Personen das Getränk oder den Film bewerten, können wir eine erwartete Bewertung berechnen.

Definition 5.5 Zur Berechnung der personenbasierten gemeinsamen Umgebung für die Person u und das Item i verwenden wir die folgende Formel

$$r_{u,i} = \overline{r}_u + \frac{\sum_v Sim(u,v)(r_{v,i} - \overline{r}_v)}{\sum_v Sim(u,v)},$$

wobei v alle Personen sind, die sich in der Nähe von Person u befinden.

Wir nehmen den gewichteten Durchschnitt von $r_{v,i} - \overline{r}_v$ mit dem Gewicht $Sim(u,v)$. Dies geschieht, damit die Personen, die der Person u ähnlicher sind, bei der Berechnung des unbekannten Wertes stärker gewichtet werden.
Auch in dieser Formel verwenden wir $\overline{r}_u$ und $\overline{r}_v$. Der Grund, warum wir von jedem Wert den Durchschnitt abziehen, ist, dass jeder Mensch anders urteilt. Eine Person kann sehr schnell eine 4 oder eine 5 als Bewertung vergeben, während eine andere Person nur dann eine 5 vergibt, wenn sie nie wieder etwas anderes trinken möchte. Wir betrachten also alle $r_{v,i} - \overline{r}_v$ und nehmen daraus den gewichteten Durchschnitt.

Beispiel 5.6 Drei Personen wurden gebeten, anzugeben, wie sehr sie ein bestimmtes Getränk mögen. Siehe die nachstehende Matrix. Noa gab an, dass sie noch nie Sprite getrunken hat.

	Cola	Fanta	Sprite	Eistee
Theo	2	5	4	1
Bente	1	2	4	2
Noa	3	3	?	5

Wir werden den Wert berechnen, den wir bei $r_{Noa,Sprite}$ erwarten würden, indem wir die personenbasierte gemeinsame Umgebung verwenden. Als Umgebungsgröße nehmen wir 2 (d.h. die beiden anderen Personen).
Zuerst werden wir $\sum_i Sim(Noa,i)$ berechnen, wobei i die anderen Personen bezeichnet. Man beachte, dass wir $r_{Noa,Sprite}$ nicht kennen und daher auch die Bewertungen der anderen Personen von Sprite nicht hinzuziehen, wenn wir die Umgebung berechnen.

$$Sim(Noa,Theo) = \frac{2\cdot 3 + 5\cdot 3 + 1\cdot 5}{\sqrt{30}\sqrt{43}} = 0,72$$

$$Sim(Noa,Bente) = \frac{1\cdot 3 + 2\cdot 3 + 2\cdot 5}{\sqrt{9}\sqrt{43}} = 0,97$$

Außerdem benötigen wir das Zeilenmittel

$$\overline{r}_{Theo} = \frac{2+5+4+1}{4} = 3$$

$$\overline{r}_{Bente} = \frac{1+2+4+2}{4} = 2,25$$

$$\overline{r}_{Noa} = \frac{3+3+5}{3} = 3,67$$

Nun können wir die Formel von Definition 5.5 verwenden:

$$r_{Noa,Sprite} = \overline{r}_{Noa} + \frac{\sum_v Sim(Noa,v)(r_{v,Sprite} - \overline{r}_v)}{\sum_v Sim(Noa,v)}$$

$$= \overline{r}_{Noa} + \frac{Sim(Noa,Theo)(r_{Theo,Sprite} - \overline{r}_{Theo}) + Sim(Noa,Bente)(r_{Bente,Sprite} - \overline{r}_{Bente})}{Sim(Noa,Theo) + Sim(Noa,Bente)}$$

$$= 3,67 + \frac{0,72 \cdot (4-3) + 0,97 \cdot (4-2,25)}{0,72 + 0.97} = 4,52$$

Beachte, dass die Matrix nur aus ganzen Zahlen besteht, das Ergebnis aber eine Dezimalzahl sein kann. Wir runden diese Dezimalzahl nicht ab, sondern nehmen sie als endgültige Antwort.

Aufgabe 14 Betrachte folgende Matrix

	Cola	Fanta	Sprite	Eistee
Theo	2	5	4	1
Bente	1	2	4	2
Dion	?	1	2	5

Berechne den Wert bei $r_{Dion,Cola}$ mit Umgebungsgröße 2 unter Verwendung der personenbasierten gemeinsamen Umgebung.

5.3 Itembasierte gemeinsame Umgebung

Wir können nicht nur nach Personen suchen, die einen hohen Ähnlichkeitsgrad aufweisen, sondern auch nach Gegenständen, die einen hohen Ähnlichkeitsgrad aufweisen. Denke an Filme, die denselben Genres angehören. Wenn jemand einen Kriminalfilm mag, besteht die Chance, dass er auch einen anderen Kriminalfilm mag.

Definition 5.7 Um die gemeinsame Umgebung für Person u und Gegenstand i zu berechnen, verwenden wir die folgende Formel

$$r_{u,i} = \overline{r}_i + \frac{\sum_j Sim(i,j)(r_{u,j} - \overline{r}_j)}{\sum_j Sim(i,j)}$$

wobei j alle Gegenstände sind, die sich in der Nähe von Gegenstand i befinden.

Wie bei der Formel zur personenbasierten gemeinsamen Umgebung verwenden wir auch hier den gewichteten Durchschnitt, in diesem Fall der $r_{u,j} - \overline{r}_j$.

Beispiel 5.8 Betrachte erneut folgende Matrix

$$\begin{array}{r c} & \begin{array}{cccc} \text{Cola} & \text{Fanta} & \text{Sprite} & \text{Eistee} \end{array} \\ \begin{array}{r} \text{Theo} \\ \text{Bente} \\ \text{Noa} \end{array} & \left[\begin{array}{cccc} 2 & 5 & 4 & 1 \\ 1 & 2 & 4 & 2 \\ 3 & 3 & ? & 5 \end{array}\right] \end{array}$$

Wir werden nun den Wert $r_{Noa,Sprite}$ berechnen, aber mit der itembasierten Methode mit Umgebungsgröße 2. Zunächst werden wir die Gemeinsamkeiten feststellen:

$$Sim(Sprite, Cola) = \frac{2 \cdot 4 + 1 \cdot 4}{\sqrt{5}\sqrt{32}} = 0,95$$
$$Sim(Sprite, Fanta) = \frac{4 \cdot 5 + 4 \cdot 2}{\sqrt{29}\sqrt{32}} = 0,92$$
$$Sim(Sprite, Eistee) = \frac{4 \cdot 1 + 4 \cdot 2}{\sqrt{5}\sqrt{32}} = 0,95$$
$$\overline{r}_{Cola} = \frac{2 + 1 + 3}{3} = 2$$
$$\overline{r}_{Fanta} = \frac{5 + 2 + 3}{3} = 3,33$$
$$\overline{r}_{Sprite} = \frac{4 + 4}{2} = 4$$
$$\overline{r}_{Eistee} = \frac{1 + 2 + 5}{3} = 2,67$$

Nun können wir die Formel aus Definition 5.7 verwenden, wobei wir berücksichtigen, dass die Umgebungsgröße 2 ist und wir somit nur Cola und Eistee beachten.

$$r_{Noa,Sprite} = \overline{r}_{Sprite} + \frac{\sum_j Sim(Sprite, j)(r_{Noa,j} - \overline{r}_j)}{\sum_j Sim(Sprite, j)} =$$
$$\overline{r}_{Sprite} + \frac{Sim(Sprite,Cola)(r_{Noa,Cola} - \overline{r}_{Cola}) + Sim(Sprite,Eistee)(r_{Noa,Eistee} - \overline{r}_{Eistee})}{Sim(Sprite, Cola) + Sim(Sprite, Eistee)}$$
$$= 4 + \frac{0,95 \cdot (3 - 2) + 0,95 \cdot (5 - 2,67)}{0,95 + 0.95} = 5,67$$

Beachte, dass das Ergebnis höher ist als die höchstmögliche Bewertung. Das ist möglich und bedeutet, dass wir bei dieser Methode davon ausgehen können, dass Noa Sprite tatsächlich mag. Es ist auch möglich, Werte unter 1 (und sogar negative) zu erhalten.

Aufgabe 15 Beachte folgende Matrix:

$$\begin{array}{r c} & \begin{array}{cccc} \text{Cola} & \text{Fanta} & \text{Sprite} & \text{Eistee} \end{array} \\ \begin{array}{r} \text{Theo} \\ \text{Bente} \\ \text{Dion} \end{array} & \left[\begin{array}{cccc} 2 & 3 & 4 & 1 \\ 3 & 2 & 3 & 2 \\ ? & 3 & 2 & 5 \end{array}\right] \end{array}$$

Berechne den Wert bei $r_{Dion,Cola}$ mithilfe der itembasierten gemeinsamen Umgebung. Verwende dabei die Umgebungsgröße 3.

6 k-Clustermittelung

Wenn z. B. Netflix eine Empfehlung für eine bestimmte Person aussprechen möchte, verfügt es über zahlreiche Informationen von anderen Teilnehmer*innen, von denen viele für die Empfehlung überflüssig sind. Eine Lösung besteht darin, das Umfeld des*der Teilnehmer*in zu bestimmen und - wie zuvor beschrieben - eine kleinere Gruppe zu bilden. Der Nachteil ist, dass auch dann immer noch alle Daten berechnet werden müssen, und genau das soll vermieden werden. Dementsprechend lohnt es sich mit dem Clustering oder der Gruppierung zu beginnen. Es wird eine Auswahl getroffen und das verwendete Angebotssystem muss nicht mehr auf alle Informationen angewandt werden. Das spart Zeit und ist deutlich effektiver, weil jetzt lediglich die Informationen verwendet werden, die interessant sind, um die Empfehlung auszusprechen. Normalerweise mögen Menschen mit demselben Geschmack dieselben Dinge und befinden sich aufgrund der k-Clustermittelung iim selben Cluster. Bei diesen Personen schauen wir uns also an, was sie noch nicht gesehen haben, und machen eine Vorhersage, welcher Film oder welches Genre ihnen wahrscheinlich gefallen wird. Dies geschieht auf der Grundlage der anderen Teilnehmer*innen der Gruppe. Wir wollen uns nun ansehen, wie das funktioniert.

Um Cluster zu erstellen, benötigt man einen Graphen mit einer endlichen Anzahl von Punkten oder eine Punktwolke. Jeder dieser Punkte stellt eine*n Teilnehmer*in dar. Betrachtet man ein zweidimensionales Beispiel, so stehen die X- und Y-Achsen beispielsweise für die Bewertungen von Filmen oder Genres. Dies kann auf höhere Dimensionen ausgedehnt werden, aber wir werden dies hier nicht tun.

Wenn ein Nutzer Filme des Genres Romantik mit einer durchschnittlichen Bewertung von 3 und Filme des Genres Action mit einer durchschnittlichen Bewertung von 3,5 bewertet hat, ergibt dies den Punkt (3,5;3).

Man kann diese Punktwolke nun in verschiedene Cluster unterteilen. Dies kann auf verschiedene Weise geschehen und hängt von der Anzahl der Cluster ab, die erstellt werden sollen. Anschließend kann dann ein Angebotssystem für eine*n Teilnehmer*in innerhalb eines Clusters angewandt werden.

Um mit dem Clustering zu beginnen, verwenden wir einen Algorithmus. Dies ist ein Fahrplan, der schrittweise vom Ausgangspunkt zum Endpunkt durchgeführt wird. Algorithmen finden ihre Anwendung auch im Alltag. Man denke beispielsweise an ein Kochrezept, in dem genau beschrieben ist, was Schritt für Schritt getan werden muss.

Es gibt verschiedene Möglichkeiten des Clusterns, und wir werden uns die k-Clustermittelung ansehen. Es ist eine Methode, die unsere n Punkte in der Punktwolke in k Cluster unterteilt. Die Anzahl der Cluster muss im Voraus geschickt gewählt werden, sodass jedes Cluster eine Gruppe von Teilnehmer*innen mit der gleichen Filmauswahl darstellt. Auf diese Weise kann eine gute Vorhersage getroffen werden. Der Algorithmus

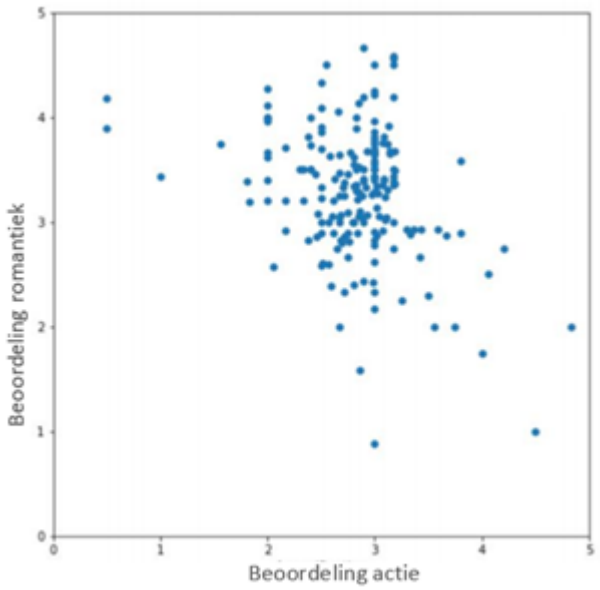

Abbildung 1: Punktwolke.

der k-Clustermittelung ist eine iterative Methode, d. h. ein Algorithmus, der sich immer und immer wiederholt. In diesem Fall bedeutet dies, dass die Schritte 3 bis 5 so lange wiederholt werden, bis die Cluster stabil sind, d. h. sie sich nicht mehr verändern. Bevor der iterative Teil des Algorithmus durchlaufen werden kann, muss die Anzahl der Cluster und die Position der Clusterzentren bestimmt werden. Das Cluster-Zentrum kann selbst gewählt oder per Zufallsprinzip bestimmen werden. Es kann also ein Datenpunkt sein, muss es aber nicht.

Der Algorithmus der k-Clustermittelung geht wie folgt vor:
Schritt 1: Lege fest, wie viele Cluster erstellt werden sollen; dies ist k.
Schritt 2: Bestimme für jedes Cluster die Position des Cluster-Zentrums.

Es folgen nun die iterativen Schritte des Algorithmus:
Schritt 3: Ordne jeden Punkt der Punktwolke dem Cluster zu, dessen Zentrum ihm am nächsten ist.
Schritt 4: Bestimme für jedes Cluster ein neues Zentrum, welches sich aus dem Durchschnitt der Koordinaten der einzelnen Punkte des Clusters berechnen lässt.
Schritt 5: Überprüfe, ob es Cluster gibt, die mindestens einen neuen Punkt haben.

a) Wenn ja, gehe zurück zu Schritt 3 und führe den Algorithmus erneut aus.

b) Ansonsten wird der Algorithmus beendet und man ist fertig.

Wir müssen den Abstand jedes Punktes zu jedem Clusterzentrum bestimmen, damit wir wissen, zu welchem Cluster er in Schritt 3 hinzugefügt werden soll. Wir

werden dazu den euklidischen Abstand verwenden. Der euklidische Abstand ist der kürzeste Abstand zwischen zwei Punkten. Sie können diesen Abstand mit Hilfe des Satzes von Pythagoras berechnen. Sie benötigen also den Unterschied in Höhe und Breite der beiden Punkte.

Beispiel 6.1 Abbildung 2 zeigt beispielhaft das Durchlaufen der Schritte der k-Clustermittelung.

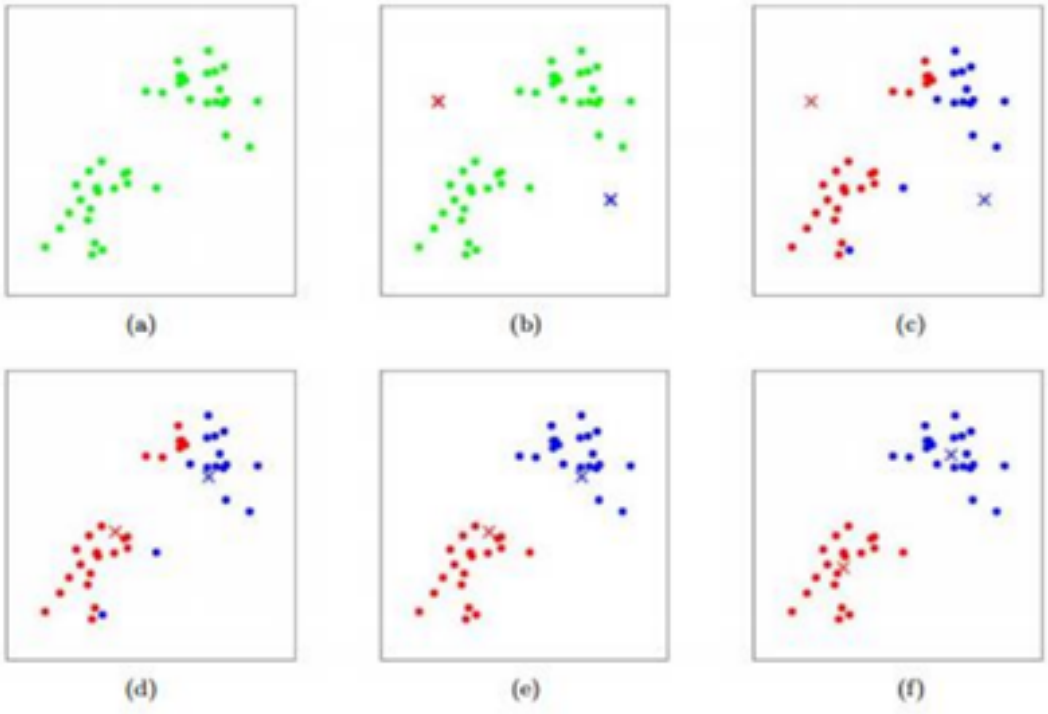

Abbildung 2: Die k-Clustermittelung Schritt für Schritt.

Schritt 1: In Abbildung (a) sehen wir die Datenpunkte. Man kann bereits zwei Cluster erkennen, daher wählen wir k = 2.
Schritt 2: Wir wählen die Clusterzentren nach dem Zufallsprinzip aus. In Abbildung (b) sind diese durch ein blaues und ein rotes Kreuz gekennzeichnet.
Schritt 3: Von jedem Punkt wird der euklidische Abstand zu den Clusterzentren berechnet, und es wird ausgewählt, zu welchem Cluster ein Punkt gehört. Die Punkte, die dem roten Kreuz am nächsten sind, werden rot gefärbt und die Punkte, die dem blauen Kreuz am nächsten sind, werden blau gefärbt, wie in Abbildung (c) zu sehen ist.
Schritt 4: Die Clusterzentren werden neu berechnet. In Abbildung (d) ist zu erkennen, dass die Kreuze verschoben wurden.
Schritt 5: Die Lage der Clusterzentren wurde in Schritt 4 geändert, sodass wir wieder mit Schritt 3 fortfahren.
Schritt 3: In Abbildung (e) ordnen wir die Punkte wieder den Clusterzentren zu.

Schritt 4: Wir bestimmen die neuen Clusterzentren.
Schritt 5: Wir setzen den Algorithmus so lange fort, bis die Clusterzentren unverändert bleiben, was nach Abbildung (f) der Fall ist.

Beispiel 6.2 Angenommen, wir haben die folgenden Punkte: (1,1), (2,1), (4,3) und (5,4). Außerdem ist k=2 gegeben und wir wählen (1,1) und (2,1) als Clusterzentren. Wir werden nun den Algorithmus der k-Clustermittelung anwenden. Dann erhalten wir:

Schritt 3: Für jeden Punkt berechnen wir den euklidischen Abstand zum Clusterzentrum und ordnen den Punkt dann einem Cluster zu. Dies geht aus der Tabelle 7 hervor.

Punkt	Abstand zum Cluster 1	Abstand zum Cluster 2	Zugewiesenes Cluster
(1,1)	0	1	1
(2,1)	1	0	2
(4,3)	$\sqrt{3^2+2^2}=\sqrt{13}$	$\sqrt{2^2+2^2}=\sqrt{8}$	2
(5,4)	$\sqrt{4^2+3^2}=5$	$\sqrt{3^2+3^2}=\sqrt{18}$	2

Tabelle 7: Abstand der Punkte zu den Clustern.

Schritt 4: Die Koordinaten des Clusterzentrums 1 betragen jetzt (1,1) und die Koordinaten des Clusterzentrums 2 betragen jetzt $(\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (3\frac{2}{3}, 2\frac{2}{3})$
Schritt 5: Die Clusterzentren haben sich geändert, also fahren wir mit Schritt 3 fort.

Aufgabe 16 Vervollständige das Beispiel 6.2 mittels des Algorithmus der k-Clustermittelung, bis sich die Cluster nicht mehr verändern.

Wir verwenden hier nur sehr wenige Punkte. Wenn wir uns Netflix anschauen, betrachten wir viel mehr Punkte und dann kann man diesen Algorithmus nicht von Hand machen. Er wird mit dem Computer durchgeführt. Um sich dies zu veranschaulichen, lohnt sich ein Besuch folgender Website:
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

7 The Netflix Prize

Dieses Kapitel ist nur als Hintergrundinformation gedacht und ist nicht notwendig, um sich für das Mathe-Turnier vorzubereiten.

Die in diesem Material behandelten Methoden zur Berechnung unbekannter Werte bilden relativ einfache Versionen von Angebotssystemen. Große Unternehmen wie Netflix verwenden äußerst ausgeklügelte Systeme. Diese Systeme bestimmen, welche Filme oder Serien unter „empfohlen“ aufgeführt werden. Dies ist also sehr wichtig für solche Unternehmen. Bis 2006 nannte Netflix sein Angebotssystem „CineMatch“. Netflix war mit CineMatch recht zufrieden; die Hälfte der Nutzer, die sich einen von CineMatch empfohlenen Film angesehen haben, gaben diesem Film sogar 5 Sterne. Dennoch sah Netflix noch Raum für Verbesserungen.
Im Jahr 2006 veranstaltete Netflix einen Wettbewerb, um einen besseren Algorithmus als CineMatch zu finden. Dieser Wettbewerb trug den Namen „Der Netflix-Preis“ („The Netflix Prize“). Bei diesem Wettbewerb gab es auch etwas zu gewinnen. Die erste Person oder Gruppe, die das System von CineMatch um mindestens 10 Prozent verbessern konnte, sollte 1 Million Dollar gewinnen.

7.1 Das Format des Wettbewerbs

Um festzustellen, ob ein System ordnungsgemäß funktioniert, verwendet man häufig einen Trainingsdatensatz und einen Qualifizierungsdatensatz.

Der Trainingsdatensatz bestand aus 100.480.507 Datenzeilen. Diese Daten hatten das Format (Name, Film, Datum der Bewertung, Bewertung), wobei die Bewertung eine Zahl zwischen 1 und 5 ist. Insgesamt wurden die Daten von 480.189 Nutzer*innen und 17.770 Filmen verwendet. Dieser Datensatz wird verwendet, um ein System zu entwickeln, das nach Verbindungen zwischen Personen oder Filmen sucht, wie in den Abschnitten 5.2 und 5.3 erläutert. Wenn man glaubt, dass das System gut funktioniert, kann man mit der Vorhersage des Quiz-Datensatzes beginnen.

Der Qualifizierungsdatensatz bestand aus 2.817.131 Zeilen, die aus (Name, Film, Datum der Überprüfung) bestanden. Die damit verbundenen Einschaltquoten waren nur der Netflix-Jury bekannt. Die eine Hälfte dieses Datensatzes bestand aus dem Quiz-Datensatz und die andere Hälfte aus dem Test-Datensatz.
Der Quiz-Datensatz wird verwendet, um zu sehen, wie gut ein System funktioniert. Eine Gruppe kann beliebig viele Beiträge einsenden, die dann anhand des Quiz-Datensatzes getestet werden.
Der Testdatensatz ist eigentlich derselbe wie der Quiz-Datensatz, wird aber erst am Ende verwendet, um den*die endgültige*n Gewinner*in zu ermitteln.
Ziel ist es, anhand der Trainingsdaten Systeme zu entwickeln, die möglichst genaue Auswertungen der Testdaten liefern können. Um zu berechnen, wie gut das System

funktioniert, wird der „root mean squared error“ (RMSE) verwendet. Mit anderen Worten: die Quadratwurzel aus der mittleren quadratischen Differenz. Dies funktioniert folgendermaßen.

Beispiel 7.1 Bei einem Datensatz lassen wir ein selbst entwickeltes System die Bewertungen ermitteln. Wir berechnen den RMSE anhand der folgenden Ergebnisse.

Bewertung	Systembewertung	Differenz	Quadratische Differenz
3	3.87	0.87	0.757
5	4.58	0.42	0.176
4	4.84	0.84	0.706
2	3.21	1.21	1.464
5	4.97	0.03	0.001

Dann nehmen wir den Durchschnitt der quadratischen Differenz:

$$\frac{0.757 + 0.176 + 0.706 + 1.464 + 0.001}{5} = 0.621$$

Und anschließend nehmen wir noch die Wurzel davon:

$$\text{RMSE} = \sqrt{0.621} = 0.788$$

Der mittlere quadratische Fehler unseres Systems ist also gleich 0,788.

Je niedriger der RMSE, desto besser funktioniert das System. Der RMSE misst die Fehler Ihrer Vorhersage. Netflix ließ sein eigenes System CineMatch den aktuellen Testdatensatz vorhersagen und kam auf einen RMSE von 0,9514. Da das neue System um 10% besser sein musste, bedeutet dies, dass das neue System einen RMSE von 0,8563 haben musste.

7.2 Der Wettbewerb selber

Der Wettbewerb begann am 2. Oktober 2006 und innerhalb einer Woche hatte die Gruppe „WXYZConsulting“ bereits ein besseres System als CineMatch. Es war nur eine sehr kleine Verbesserung, aber sie war sehr vielversprechend für die Zukunft.

Ein Jahr, nachdem der Wettbewerb begonnen hatte, wurde bekannt gegeben, dass sich mehr als 40.000 Teams aus 186 Ländern angemeldet hatten. Von allen Teams belegte das Team „KorBell“ mit einem RMSE von 0,8712 den ersten Platz. Dies ist eine Verbesserung um 8,43%.
Am 26. Juni 2009 gab es eine Einsendung, mit der CineMatch um 10,05% verbessert wurde. Das war der Zeitpunkt, an dem Netflix den ‚Letzten Aufruf‘ machte. Jeder hatte nun 30 Tage Zeit, sein selbst entwickeltes System einzureichen. Nach Ablauf dieser 30 Tage werden alle Systeme mit dem Testdatensatz getestet, und das System, das dabei am besten abschnittt, sollte den Wettbewerb gewinnen.
Nach diesen 30 Tagen gab es zwei Teams namens „BellKor’s Pragmatic Chaos“ und

„The Ensemble“, die beide den gleichen RMSE von 0,8567 auf dem Testdatensatz hatten, was eine Verbesserung von 10,06% bedeutet. Allerdings ist „BellKor's Pragmatic Chaos“ das Gewinnerteam, da ihr Beitrag 20 Minuten früher als der Beitrag von „The Ensemble“ eingereicht wurde.

8 Lösungen

Aufgabe 1 Die Summen haben folgende Lösungen:

a) $\sum_{i=16}^{20} \frac{i}{4} = 22\frac{1}{2}$

b) $\sum_{k=1}^{4} \frac{3k}{2} = 15$

c) $\sum_{i=2}^{5} 6i + i^3 = 308$

Aufgabe 2

$$\overline{r}_{Karen} = \frac{6+4+2}{3} = 4$$
$$\overline{r}_{Unentschieden} = \frac{2+1+3}{3} = 2$$

Aufgabe 3 $P(Y|X) = \frac{1}{13}$

Aufgabe 4 $P(W|N) = \frac{37}{89} \approx 0,416$ und $P(N|W) = \frac{37}{113} \approx 0,327$

Aufgabe 5 Die Wahrscheinlichkeit, dass sich ein Schulkind beim Wintersport ein Bein bricht, beträgt 0,5

Aufgabe 6

$$P(\text{Lied } 1 = 1|1,1,-1,-1) \propto \frac{2}{4} \cdot \frac{1}{2} \cdot \frac{1}{1} \cdot \frac{2}{2} \cdot \frac{1}{2} = \frac{4}{32} \approx 0,125$$
$$P(\text{Lied } 1 = -1|1,1,-1,-1) \propto \frac{2}{4} \cdot \frac{0}{1} \cdot \frac{0}{2} \cdot \frac{0}{2} \cdot \frac{0}{2} = \frac{0}{32} = 0$$

Wir sagen voraus, dass Teilnehmer*in 3 Lied 1 mit 1 bewerten wird.

Aufgabe 7

$$P(\text{Lied } 2 = 1|-1,-1,1,1,1) \propto \frac{2}{4} \cdot \frac{0}{1} \cdot \frac{0}{1} \cdot \frac{0}{2} \cdot \frac{0}{2} \cdot \frac{0}{1} = 0$$
$$P(\text{Lied } 2 = -1,-1,1,1,1) \propto \frac{2}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{1}{2} = \frac{4}{128} \approx 0,031$$

Wir sagen voraus, dass Teilnehmer*in 5 Lied 2 mit -1 bewerten wird.

Aufgabe 8

$$P(\text{Tennis} = \text{Ja}|\text{sonnig, kalt, hoch, stark}) \propto \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \approx 0,0053$$
$$P(\text{Tennis} = \text{Nein}|\text{sonnig, kalt, hoch, stark}) \propto \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \approx 0,0206$$

Wenn wir daraus ein Verhältnis machen wollen, erhalten wir das Verhältnis 'Es wird Tennis gespielt' : 'Es wird kein Tennis gespielt'. Dies ergibt 0,0053 : 0,0206 oder 1 : 3,8868. Daher sagen wir voraus, dass es am 15. Tag kein Tennis geben wird.

Aufgabe 9

$$P(\text{Tennis} = \text{Ja}|\text{bewölkt, mittel, hoch, schwach}) \propto \frac{9}{14} \cdot \frac{4+1}{9+3} \cdot \frac{4}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \approx 0,0265$$

$$P(\text{Tennis} = \text{Nee}|\text{bewölkt, mittel, hoch, schwach}) \propto \frac{5}{14} \cdot \frac{0+1}{5+3} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \approx 0,0057$$

Wir verwenden erneut das Verhältnis 'Es wird Tennis gespielt' : 'Es wird kein Tennis gespielt'. Wir erhalten das Verhältnis 0,0265 : 0,0057 oder 1 : 0,2151. Deswegen sagen wir voraus, dass am 17. Tag Tennis gespielt wird.

Aufgabe 10

$$P(\text{Haus} = \text{Ja}|\text{Frau, Nein, Ja}) \propto \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{2+1}{3+2} \cdot \frac{2+1}{3+2} = \frac{27}{300} \approx 0,09$$

$$P(\text{Haus} = \text{Nein}|\text{Frau, Nein, Ja}) \propto \frac{1}{4} \cdot \frac{1}{1} \cdot \frac{0+1}{1+2} \cdot \frac{0+1}{1+2} = \frac{1}{36} \approx 0,0278$$

Es ergibt sich das Verhältnis 0,09 : 0,0278 oder 1 : 0,309.

Aufgabe 11

a) $P(\text{Computer} = \text{Ja}| \leq 30, \text{Mittel, Ja, Mann}) \propto \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \approx 0,028$
$P(\text{Computer} = \text{Nein}| \leq 30, \text{Mittel, Ja, Mann}) \propto \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \approx 0,007$
Daraus ergibt sich ein Verhältnis von 0,028 : 0,007, d.h. 1 : 0,25

b) Achim wird sich wahrscheinlich einen neuen Computer kaufen.

c) Die sechs nächstgelegenen Nachbarn von Angela sind die Personen 4, 6, 10, 11, 12 und 14.

d) Angela wird sich wahrscheinlich einen neuen Computer kaufen.

e) $P(\text{Computer} = \text{Ja}|31...40, \text{ Mittel, Nein, Mann}) \propto \frac{9}{14} \cdot \frac{4+1}{9+3} \cdot \frac{4}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \approx 0,026$
$P(\text{Computer} = \text{Nein}|31...40, \text{Mittel, Nein, Mann}) \propto \frac{5}{14} \cdot \frac{0+1}{5+3} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \approx 0,006$
Wir erhalten als Verhältnis 0,026 : 0,006, das ergibt 1 : 0,231. Wir sagen also voraus, dass Olaf sich einen neuen Computer kaufen wird.

Aufgabe 12

$$Sim(Peter, Kim) = \frac{4 \cdot 3 + 3 \cdot 2 + 1 \cdot 3 + 5 \cdot 1}{\sqrt{4^2 + 3^2 + 1^2 + 5^2}\sqrt{3^2 + 2^2 + 3^2 + 1^2}} = \frac{26}{\sqrt{51}\sqrt{23}} = 0,76$$

$$Sim(Cola, Fanta) = \frac{4 \cdot 3 + 0 \cdot 1 + 3 \cdot 2}{\sqrt{4^2 + 0^2 + 3^2}\sqrt{3^2 + 1^2 + 2^2}} = \frac{18}{\sqrt{25}\sqrt{14}} = 0,96$$

Aufgabe 13 Zuerst werden wir alle Umgebungen mit Hassan berechnen.

$$Sim(Peter, Hassan) = 0,54 \qquad \text{(Berechnet wie in Beispiel 5.2)}$$

$$Sim(Kim, Hassan) = \frac{3 \cdot 0 + 2 \cdot 1 + 3 \cdot 5 + 1 \cdot 3}{\sqrt{35}\sqrt{23}} = \frac{20}{\sqrt{35}\sqrt{23}} = 0,70$$

Wir haben die Umgebungsgröße 1, also ist die Umgebung von Hassan mit der Umgebungsgröße 1 gleich Kim.

Aufgabe 14 Wir berechnen zunächst das Übereinstimmungsmaß mit den anderen Personen.

$$Sim(Dion, Theo) = \frac{5 \cdot 1 + 4 \cdot 2 + 1 \cdot 5}{\sqrt{42}\sqrt{30}} = \frac{18}{\sqrt{42}\sqrt{30}} = 0,51$$

$$Sim(Dion, Bente) = \frac{2 \cdot 1 + 4 \cdot 2 + 2 \cdot 5}{\sqrt{24}\sqrt{30}} = \frac{20}{\sqrt{24}\sqrt{30}} = 0,75$$

Die Zeilenmittel lauten wie folgt:

$$\overline{r}_{Theo} = \frac{2 + 5 + 4 + 1}{4} = 3$$

$$\overline{r}_{Bente} = \frac{1 + 2 + 4 + 2}{4} = 2,25$$

$$\overline{r}_{Dion} = \frac{1 + 2 + 5}{3} = 2,67$$

Anschließend können wir die Werte ausrechnen.

$$r_{Dion,Cola} = \overline{r}_{Dion} + \frac{\sum_v Sim(Dion, v)(r_{v,Cola} - \overline{r}_v)}{\sum_v Sim(Dion, v)}$$

$$= 2,67 + \frac{0,51(2 - 3) + 0,75(1 - 2,25)}{0,51 + 0,75} = 1,52$$

Aufgabe 15 Berechne zunächst das Übereinstimmungsmaß von Cola mit den anderen Getränken.

$$Sim(Cola, Fanta) = \frac{2 \cdot 3 + 3 \cdot 2}{\sqrt{13}\sqrt{13}} = \frac{12}{13} = 0,92$$

$$Sim(Cola, Sprite) = \frac{2 \cdot 4 + 3 \cdot 3}{\sqrt{13}\sqrt{25}} = \frac{17}{\sqrt{13}\sqrt{25}} = 0,94$$

$$Sim(Cola, Eistee) = \frac{2 \cdot 1 + 3 \cdot 2}{\sqrt{13}\sqrt{5}} = \frac{8}{\sqrt{13}\sqrt{5}} = 0,99$$

Und die Spaltenmittel:

$$\overline{r}_{Cola} = \frac{2 + 3}{2} = 2,5$$

$$\overline{r}_{Fanta} = \frac{3 + 2 + 3}{3} = 2,67$$

$$\overline{r}_{Sprite} = \frac{4 + 3 + 2}{3} = 3$$

$$\overline{r}_{Eistee} = \frac{1 + 2 + 5}{3} = 2,67$$

Mit der entsprechenden Formel gilt:

$$r_{Dion,Cola} = \overline{r}_{Cola} + \frac{\sum_j Sim(Cola, j)(r_{Dion,j} - \overline{r}_j)}{\sum_j Sim(Cola, j)}$$

$$= 2,5 + \frac{0,92 \cdot (3 - 2,67) + 0,94 \cdot (2 - 3) + 0,99 \cdot (5 - 2,67)}{0,92 + 0,94 + 0,99} = 2,5 - 0,59 = 1,91$$

Aufgabe 16 Wir beginnen wieder mit Schritt 3, also mit der Berechnung des euklidischen Abstandes. Es ergibt sich das Folgende:

Punkt	Abstand zu Cluster 1	Abstand zu Cluster 2	Zugewiesenes Cluster
(1,1)	0	3,14	1
(2,1)	1	2,36	1
(4,3)	3,61	0,47	2
(5,4)	5	1,89	2

Daraus ermitteln wir die neuen Clusterzentren. Es folgt, dass die Koordinaten des Cluster-Zentrums 1 $(1\frac{1}{2}, 1)$ und die Koordinaten des Cluster-Zentrums 2 $(4\frac{1}{2}, 3\frac{1}{2})$ sind.
Auch hier haben sich die Clusterzentren geändert, so dass wir die Iteration erneut durchlaufen.

Das heißt, wir beginnen erneut mit Schritt 3, berechnen also erneut den euklidischen Abstand der Punkte zu den neuen Cluster-Zentren. Wir erhalten dann:

Punkt	Abstand zu Cluster 1	Abstand zu Cluster 2	Zugewiesenes Cluster
(1,1)	0,5	4,3	1
(2,1)	0,5	3,5	1
(4,3)	3,2	0,7	2
(5,4)	4,6	0,7	2

Dies ergibt die gleichen Clusterzentren wie zuvor, wir haben also unseren Algorithmus beendet und unsere Cluster gefunden.