Arbeitsblatt 4 Lösungen

Aufgabe 1: Geschlechterbasierte Vorlieben beurteilen

Die Kreuze wurden aufgrund einer Analyse der durchschnittlichen Filmbewertungen von Frauen und Männern gesetzt. Jeder der 10 Filme wurde von mindestens 40 Frauen und 40 Männern bewertet. Das Ergebnis sieht folgendermassen aus:

Titel	Bewertung M	Bewertung F
Return of the Jedi (1983)	4.01	4.01
Volcano (1997)	2.72	3.09
2001: A Space Odyssey (1968)	4.10	3.49
The Silence of the Lambs (1991)	4.28	4.32
A Fish Called Wanda (1988)	3.79	3.78
My Fair Lady (1964)	3.60	4.24
E.T. the Extra-Terrestrial (1982)	3.83	3.84
Star Trek: The Wrath of Khan (1982)	3.92	3.38
Indiana Jones and the Last Crusade (1989)	3.93	3.92
Grease (1978)	3.14	3.75

Analyse: *My Fair Lady* hat mit 0.64 Sternen Differenz zwischen Männern und Frauen die zweitgrösste Abweichung aller Filme, die mindestens von je 40 Frauen und Männern bewertet wurden. *Return of the Jedi* weist die geringste Abweichung all dieser Filme auf (keine messbare Abweichung).

Es scheint nicht so einfach zu sein, vorherzusagen, welche Filme vor allem von Männern und welche vor allem von Frauen gut bewertet werden. Bei einigen Filmen stimmen unsere Vermutungen (z.B. My Fair Lady und Grease als von Frauen besser bewertete und Star Trek und 2001: Space Odyssey für von Männern besser bewertete) mit den berechneten Daten überein. Andere Filme erstaunen aber (z.B. Return of the Jedi, Volcano, The Silence of the Lambs).

Wir können nun einen «Leap of Faith» machen und daraus folgende geschlechtsbasierte Vorlieben ableiten:

Titel	klar M	va M	?	va F	klar F
Return of the Jedi (1983)			x		
Volcano (1997)				x	
2001: A Space Odyssey (1968)	x				
The Silence of the Lambs (1991)			x		
A Fish Called Wanda (1988)			x		
My Fair Lady (1964)					x
E.T. the Extra-Terrestrial (1982)			x		
Star Trek: The Wrath of Khan (1982)	x				
Indiana Jones and the Last Crusade (1989)			x		
Grease (1978)					x

Es bleibt aber die Frage, ob dieser Sprung von den Filmbewertungen zu Aussagen über die Filmvorlieben von Männern und Frauen zulässig ist.

Aufgabe 2: Aussagekraft

a) Nicht für alle Filme liegen gleich viele einzelne Bewertungen vor. Einige wurden nur von einer Handvoll Personen bewertet, andere von einer Vielzahl. Wahrscheinlich liegt dies daran, dass nicht alle Filme gleich bekannnt sind. Filme, die oft gezeigt werden bzw welche viele Leute kennen, werden wohl mehr Bewertungen erhalten als Filme, die nicht so verbreitet und bekannt sind. Die Filme die nicht genug Bewertungen haben, fallen aus der Liste heraus; deshalb werden es immer weniger Filme, je mehr Bewertungen gefordert sind.

b) Bekanntere/verbreitetere Filme sind wahrscheinlich beliebter. Das macht ja auch Sinn: Kinos und private TV-Stationen zeigen am liebsten Filme, die Geld einbringen. Das heisst, dass sie vor allem Filme zeigen, die beim Publikum gut ankommen.

c) Die Standardabweichung für den Unterschied zwischen Bewertungen, die ausschliesslich von Männern erfolgt, und solchen, die ausschliesslich von Frauen stammen, beträgt 0.19. Der Mittelwert des Unterschieds ist -0.02. Wenn wir mit dem gleichen Signifikanzniveau arbeiten wie viele Wissenschaften, können wir also davon ausgehen, dass alle Abweichungen, die im Bereich von -0.40 bis 0.36 liegen (-0.02 ± 2*0.19), nicht signifikant sind (d.h. wohl zufällige Abweichungen sind) und deshalb keine Aussagekraft besitzen. Damit **fallen alle Filme weg**. Grundsätzlich heisst das also, dass unsere Methode, Filme zu identifizieren, die vor allem von Frauen oder von Männern gemocht werden, ziemlich unzuverlässig ist.

   Interessant ist übrigens, dass wir, wenn wir uns auf weniger Bewertungen beschränken (mindestens je 40 von Männern und Frauen), andere Ergebnisse erhalten (Mittelwert -0.06, Standardabweichung 0.24, und es gibt doch einige Filme mit einer Bewertungsdifferenz von über 0.5). Das sollte dich nachdenklich stimmen. Wenn wir mit einer kleineren Stichprobe arbeiten (wir geben uns mit weniger Bewertungen pro Film zufrieden, um die Durchschnittswerte für Männer und Frauen zu berechnen), *fällt uns vielleicht gar nicht auf*, dass unsere Methode nicht besonders gut ist!

Aufgabe 3: Star Wars

Return of the Jedi ist der am ähnlichsten bewertete Film (praktisch kein Unterschied zwischen männlichen und weiblichen Bewertungen). *The Empire Strikes Back* gehört zu den Top 5 der von Frauen und Männern unterschiedlich bewerteten Filme und ist der Film, der Männern (gemäss unserem Modell) am deutlichsten besser gefällt als Frauen. Warum liefern zwei Filme, die in der gleichen Fantasiewelt spielen (George Lucas meinte dazu: Cowboys im Weltraum), teil der gleichen Geschichte sind, die gleichen Helden und Schurken haben und vom gleichen Filmstudio produziert wurden, so unterschiedliche Ergebnisse?

a) Dieses Resultat ist weiterer Hinweis darauf, dass unser Bewertungsmodell schlicht **nicht besonders gut** ist. Die Differenz zum Mittelwert ist bei beiden Filmen weniger als zwei Standardabweichungen; es wäre also nicht sehr erstaunlich, wenn die errechneten Durchschnittswerte **schlicht und einfach Zufall** wären.

b) Falls *Return of the Jedi* tatsächlich von Männern und Frauen gleichermassen gemocht wird, *The Empire Strikes Back* aber eher ein «Männerfilm» ist, dann können wir nur aufgrund der vorliegenden Daten keinerlei Hypothesen aufstellen, weshalb das so ist. Die Daten helfen uns grundsätzlich nicht dabei zu erklären, weshalb die Ergebnisse so ausfallen, wie sie ausfallen (liegt es an bestimmten Schauspielern? An der Geschichte? Sind Frauen Happy Endings in Filmen wichtig? Spielt es eine Rolle, wieviel On-Screen-Time weibliche Hauptfiguren haben? Macht der Schnitt des Films den Unterschied aus, oder der Regisseur?) Wir müssten unser Modell stark erweitern (z.B. um Informationen zum Regisseur, zu den Schauspielern usw), um Antworten auf solche Hypothesen erhalten zu können.

   Dass unser Modell **keine Erklärungen** liefert, ist eine sehr wichtige Eigenschaft von praktisch allen modernen Machine Learning-Modellen. Die Modelle selbst sind vielleicht nützlich oder nutzlos, sehr gut oder schlecht, aber allen gemeinsam ist, dass sie uns **keinen Erkenntnisgewinn** verschaffen, sondern nur Muster in bestehenden Daten finden.