Woche 04 Theorie: Statistische Konzepte

Skript

Erarbeitet von
Katja Theune

Lernziele

- Benennen wichtiger statistischer Konzepte
- Anwenden dieser Konzepte auf ein Beispiel

Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Statistische Konzepte begegnen uns überall in unserem Alltag und wir verwenden sie ständig, oft vielleicht auch unbewusst. Viele dieser Konzepte sind auch für den Umgang mit Künstlicher Intelligenz von großer Bedeutung.

Bedeutung statistischer Konzepte für KI

Statistik und Künstliche Intelligenz lassen sich nicht voneinander trennen. Sogar ganz im Gegenteil. Statistische Konzepte sind ein grundlegender Bestandteil von Künstlicher Intelligenz. Sie helfen uns beim gesamten Analyseprozess, vom Umgang mit Daten und Methoden bis hin zur Interpretation der Ergebnisse. Darüber hinaus basieren natürlich auch die Methoden selbst auf Statistik. Hier möchte ich aber zunächst nur ein paar wenige zentrale Konzepte vorstellen.

Methoden der Künstlichen Intelligenz lernen aus Daten, also z. B. Befragungsdaten, Bilddaten oder Textdaten. Diese liegen uns meist in Form von Datentabellen vor. Um mit ihnen arbeiten zu können, benötigen wir das statistische bzw. mathematische Konzept der Vektoren und Matrizen.

Darüber hinaus sind Methoden der KI nur so gut, wie die Trainingsdaten, mit denen sie gefüttert werden. Wir benötigen also grundlegendes Wissen über die Daten, die wir verwenden. Das gilt sowohl für das Zustandekommen der Daten, die Beurteilung der Datenqualität, die Datenaufbereitung, als auch für die Datenbeschreibung, Visualisierung und Interpretation. Hierbei helfen z. B. statistische Kennzahlen wie der Mittelwert oder auch umgangssprachlich der Durchschnitt.

Vektoren und Matrizen

Zunächst betrachten wir zwei eher mathematische Konzepte. Vektoren und Matrizen sind insbesondere für den Umgang mit Daten wichtig. Unsere Trainingsdaten liegen uns häufig als Datentabelle vor. Dabei sind die Beobachtungen, das können Personen oder Objekte sein, in den Zeilen und die features in den Spalten sortiert. Zur Veranschaulichung sehen wir hier einmal eine solche Datentabelle. Hier haben wir z. B. Daten über fünf Beobachtungen, hier wären es Personen, und über die drei features Alter, Bildungsjahre und Einkommen vorliegen. Die Einträge der Tabelle sind hier die Werte bzw. Ausprägungen von unseren verschiedenen features für alle unsere Beobachtungen.

Die Zeilen repräsentieren dabei eine bestimmte Beobachtung und ihre jeweiligen Ausprägungen der verschiedenen features. Z. B. hätte die erste Beobachtung, hier sind es ja Personen, ein Alter von 20 Jahren, 10 absolvierte Bildungsjahre und ein Einkommen von 3200 Euro.

Eine Spalte repräsentiert für ein bestimmtes feature die Ausprägungen für alle Beobachtungen. Hier hat das feature Alter z. B. die Ausprägungen 20,67,35,28 und 35.

Datentabelle

	Alter	Bildungsjahre	Einkommen
Beobachtung 1	20	10	3200
Beobachtung 2	67	15	4300
Beobachtung 3	35	12	3500
Beobachtung 4	28	17	5400
Beobachtung 5	35	13	4000

Die genaue Bedeutung der Zeilen, Spalten und der Zusammenhang der Werte ist allerdings nur für uns selbst ersichtlich und inhaltlich sinnvoll, aber nicht für einen Computer. Daher extrahiert man für den computerbasierten Umgang mit den Daten nur die Werte der features. Man behält die Struktur von Zeilen und Spalten aber so bei, damit ihr Zusammenhang und ihre Bedeutung für uns nicht verloren geht. Wir haben jetzt eine sogenannte Datenmatrix. Eine Matrix wird häufig in eckigen oder auch runden Klammern dargestellt.

Datenmatrix & -vektor

$$\begin{bmatrix} 20 & 10 & 3200 \\ 67 & 15 & 4300 \\ 35 & 12 & 3500 \\ 28 & 17 & 5400 \\ 35 & 13 & 4000 \end{bmatrix}$$

Dimension:

m x n
m: Zeilen
n: Spalten
hier: 5 x 3

$$\begin{pmatrix} 20 & 10 & 3200 \\ 67 & 15 & 4300 \\ 35 & 12 & 3500 \\ 28 & 17 & 5400 \\ 35 & 13 & 4000 \end{pmatrix}$$

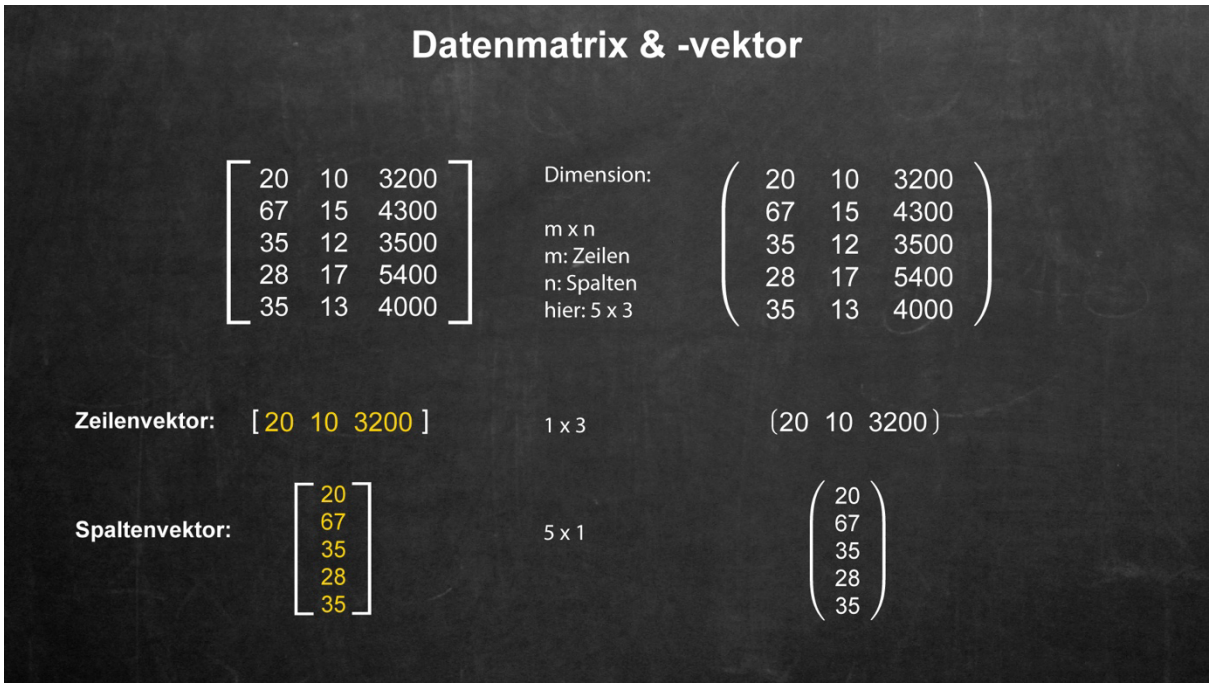
Allgemein sind Matrizen, einfach und reduziert auf unsere Zwecke ausgedrückt, eine Menge von Einträgen, die in Zeilen und Spalten sortiert werden. Sie haben eine sogenannte Dimension, die wir mit mxn angeben. Sie gibt die Anzahl an Zeilen m und Spalten n an. In unserer Datenmatrix haben wir also fünf Zeilen und drei Spalten, wir haben damit eine 5x3 Matrix. Die sogenannten Matrix-Einträge sind hier wieder die Werte bzw. Ausprägungen von unseren verschiedenen features für alle unsere Beobachtungen aus unserer Datentabelle.

Die Zeilen repräsentieren dabei weiterhin eine bestimmte Beobachtung und ihre jeweiligen Ausprägungen der verschiedenen features. Wie in dem Beispiel der Datentabelle hätte die erste Beobachtung ein Alter von 20 Jahren, 10 absolvierte Bildungsjahre und ein Einkommen von 3200 Euro.

Eine Spalte repräsentiert auch hier für ein bestimmtes feature die Ausprägungen für alle Beobachtungen. Hier hat wieder das feature Alter, das stand ja in der ersten Spalte, die Ausprägungen 20,67,35,28 und 35.

Eine Matrix, die aus nur einer Zeile oder nur einer Spalte besteht, bezeichnen wir als Vektor. Nehmen wir wie vorhin wieder die erste Zeile aus der obigen Datenmatrix, dann erhalten wir einen sogenannten Zeilenvektor mit der Dimension 1x3. Er gibt uns dann wieder alle Werte für die features der ersten Person an.

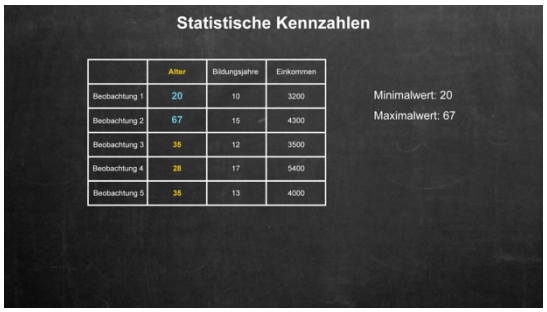
Nehmen wir jetzt wieder die erste Spalte, erhalten wir einen sogenannten Spaltenvektor mit der Dimension 5x1. Er beinhaltet in unserem Fall wieder für das erste feature, hier das Alter, die Ausprägungen für alle Personen.

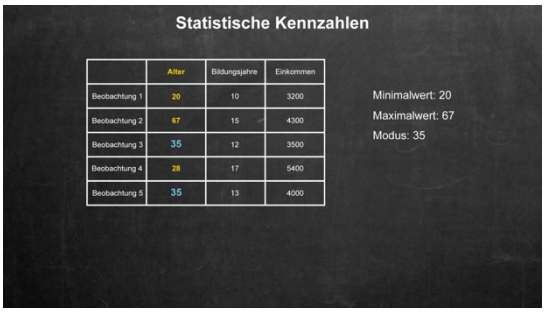


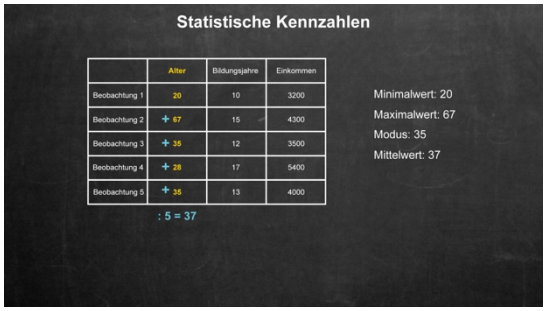
Statistische Kennzahlen

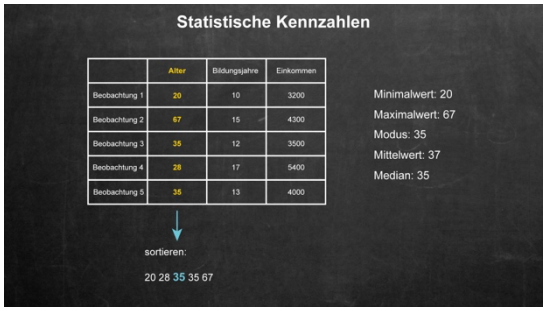
Deskriptive bzw. beschreibende statistische Kennzahlen sind für viele Aspekte in unseren Analysen relevant. Insbesondere auch, wie bereits gesagt, um unsere Trainingsdaten zu verstehen.

Schauen wir uns mal ein paar wichtige Kennzahlen an, die sich nur mit einem feature beschäftigen und seine Merkmalsausprägungen beschreiben. Dazu betrachten wir jetzt nur die Spalte mit dem feature Alter.









Der Minimal- und Maximalwert beschreibt die niedrigste und höchste Ausprägung eines features. Bei unserem feature Alter ist der Minimalwert also 20 und der Maximalwert 67.

Der Modus ist die häufigste Ausprägung eines features. Hier wäre es der Wert 35, er kommt zweimal vor, alle anderen nur einmal.

Der arithmetische Mittelwert gibt den uns aus dem Alltag bekannten Durchschnitt an. Für unser feature Alter erhalten wir ihn, wenn wir einfach die fünf Werte aufsummieren und durch die Anzahl an Beobachtungen, also fünf, teilen. Wir erhalten ein Durchschnittsalter von 37.

Der Median oder auch Zentralwert gibt den Wert an, der in der Mitte stehen würde, wenn wir die Werte für das feature Alter der Größe nach sortieren. Es liegen dann 50 % der Werte unter und 50 % der Werte über diesem Wert. Er liegt hier bei 35.

Es gibt aber natürlich noch sehr viele weitere Kennzahlen. Viele davon können z. B. für das Daten-Preprocessing genutzt werden, um die Repräsentativität der Daten zu beurteilen, Plausibilitätsprüfungen zu machen oder Ausreißer festzustellen.

Abschluss

Wir haben gesehen, dass wenn wir uns mit KI beschäftigen möchten, wir uns auch mit einigen statistischen Konzepten auseinandersetzen müssen. Sie helfen uns, Daten, Methoden und unsere Ergebnisse zu verstehen.

Weiterführendes Material

Fachbücher:

Kosfeld, R., Eckey, H. F., & Türck, M. (2016). *Deskriptive Statistik: Grundlagen-Methoden-Beispiele-Aufgaben*. Springer-Verlag.

Videos/Kurse:

Von der Datenanalyse zur Datengeschichte – Datenanalyseergebnisse adressatengerecht kommunizieren, Datenanalyse I, Datenanalyse Theorie I: Deskriptive Statistik vs. Inferenzstatistik.
https://learn.ki-campus.org/courses/Datenanalyse-unibi2021/items/1lgMIuSBdZuC32b3JXrK4H

Disclaimer