Humboldt-Universität zu Berlin
Institut für Informatik
Prof. Dr. Ulf Leser
Arik Ermshaus, Hermann Stolte

Berlin, den 30.04.2024

Übungen zur Vorlesung
Statistik und Data Science für die Informatik

Übungsblatt 1

Abgabe: Bis Montag, den 13.05.2024, bis 23:59 Uhr über Moodle. Die Übungsblätter sind in Gruppen von drei (in Ausnahmefällen zwei) Studierenden zu bearbeiten. Die Lösungen sind, sofern nicht anders angegeben, auf nach Aufgaben getrennten **PDFs** über Moodle abzugeben. Alle Teilaufgaben einer Aufgabe sind in einem PDF hochzuladen. Pro Gruppe genügt es, wenn eine Person die Lösung der Gruppe abgibt. Zur Bewertung wird der zuletzt hochgeladene Stand herangezogen. Vermerken Sie auf allen Abgaben Ihre **Namen**, Ihre **CMS-Benutzernamen** und Ihre **Abgabegruppe (z.B. Gruppe 123)** aus Moodle. Benennen Sie die hochgeladenen PDF-Dateien nach dem Schema: A<Aufgabe>-<Person1>-<Person2>- <Person3>.pdf, bspw. A3-Musterfrau-Mustermann-Beispiel.pdf für Aufgabe 3 von Lisa Musterfrau, Peter Mustermann und Karla Beispiel. Die Auflistung der Namen kann in beliebiger Reihenfolge erfolgen. Beachten Sie die Informationen im Moodle-Kurs https://hu.berlin/sds24.

Aufgabe 1 (Features eines Datensatzes) — 4 + 2 + 3 = 9 Punkte

Die folgende Tabelle enthält Informationen über verschiedene Tornados in den USA im Jahr 2021[1]. Es gibt sieben Spalten: Jahr, Monat, Bundesstaat, Magnitude (auf der Fujita-Skala), Streckenlänge (in Meilen), Breite (in Metern) und Anzahl der Verletzten.

Jahr	Monat	Bundesstaat	Magnitude	Streckenlaenge	Breite	Verletzte
2021	1	FL	1	21.37	200	0
2021	2	GA	0	0.50	50	0
2021	2	GA	2	1.81	600	5
2021	3	KS	2	12.15	100	0
2021	4	TX	0	2.61	50	0
2021	6	OH	2	5.60	200	0
2021	7	WI	1	2.44	125	0
2021	10	TN	0	2.51	25	0
2021	11	LA	1	0.16	100	0
2021	12	TN	4	168.53	2600	515
2021	12	KY	3	29.26	440	63
2021	12	TX	1	1.57	100	0
2021	12	FL	1	1.29	50	1
2021	12	AL	1	0.95	50	0
2021	12	GA	1	2.75	150	0
2021	12	GA	1	2.50	75	6

Tabelle 1: Tornados in den USA im Jahr 2021

(a) Nennen Sie jeweils die Spalten, auf die die folgenden Eigenschaften zutreffen:
(1) diskret, (2) nominal, (3) ordinal, (4) kontinuierlich.

[1] Teilmenge des Datensatzes: https://www.kaggle.com/datasets/michaelbryantds/tornadoes

(b) Bestimmen Sie die relative Frequenz von (1) Tornados mit einer Magnitude ≥ 2 und (2) Tornados mit mindestens einer verletzten Person. Geben Sie Ihren Rechenweg an.

(c) Berechnen Sie (1) den Median und Mittelwert der Anzahl von verletzten Personen und (2) den Modus des Monats. Geben Sie Ihren Rechenweg an.

Aufgabe 2 (Korrelation) **4 + 5 + 2 = 11 Punkte**

Wir betrachten einen Datensatz mit zwei Spalten X und Y, die die Zeit, die ein Benutzer oder eine Benutzerin in einer Messenger-App pro Tag verbringt (in Stunden), und die mittlere Anzahl der Nachrichten, die er oder sie pro Tag verschickt, repräsentieren.

X (Stunden)	Y (Nachrichten)
1	9.50
2	21.86
3	30.65
4	45.52
5	49.77

Tabelle 2: Zeit in Messenger-App und gesendete Nachrichten

(a) Wie hoch ist die Kovarianz s_{XY} sowie die Pearson Korrelation r_{XY} zwischen X und Y? Geben Sie Ihren Rechenweg an. *Hinweis:* Nutzen Sie zur Berechnung der Standardabweichung sowie Kovarianz die Bessel-Korrektur.

(b) Berechnen Sie die lineare Regression $f : \mathbb{N} \to \mathbb{R}$, die Stunden (X) auf gesendete Nachrichten (Y) abbildet. Welchen Wert liefert die Funktion für $f(3)$? Geben Sie Ihren Rechenweg an.

(c) Wie viele Nachrichten werden laut Ihrer Funktion $f : \mathbb{N} \to \mathbb{R}$ aus Aufgabe 2 (b) pro Stunde verschickt? Begründen Sie.

Aufgabe 3 (Multivariate Lineare Regression) **6 + 4 = 10 Punkte**

In Moodle finden Sie die Datei "honda_city.csv". Dieser Datensatz enthält Informationen zu Gebrauchtwagenverkäufen von Honda City Kleinwagen[2]. Die CSV-Datei enthält drei Spalten: Baujahr, Kilometerstand und Verkaufspreis (in Tausend Dollar).

(a) Nutzen Sie den Datensatz, um eine multivariate lineare Regression zu implementieren, die mithilfe des Baujahres und des Kilometerstandes den Verkaufspreis schätzt. Implementieren Sie die algebraische Lösung zur Berechnung der Regression. Geben Sie die Schätzwerte für den Datensatz zurück.

(b) Berechnen Sie den "Root Mean Square Error" der Schätzwerte (bezüglich der echten Verkaufspreise) aus Aufgabe 3 (a). Notieren Sie sitchhaltig, als Kommentar, was dieser Fehlerwert im Kontext der Aufgabe bedeutet.

[2]Teilmenge des Datensatzes: https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

Nutzen Sie zur Beantwortung der Fragen aus (a) und (b) das in Moodle bereitgestellte Python-Template ”aufgabe3.py”. Im Template gibt es jeweils eine Funktion je Teilaufgabe, in welcher die Antwort programmatisch ermittelt werden soll. Die erwarteten Datentypen der Rückgabe entnehmen Sie der Beschreibung zu Beginn jeder Funktion.

Verändern Sie **nicht** die Signatur der definierten Funktionen. Darüber hinaus können Sie beliebige Hilfsfunktionen definieren. Sie dürfen lediglich die Open-Source-Bibliothek NumPy nutzen.