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

Berlin, den 14.05.2024

Übungen zur Vorlesung
Statistik und Data Science für die Informatik

Übungsblatt 2

Abgabe: Bis Montag, den 27.05.2024, 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 (Grundlagen der Wahrscheinlichkeitsrechnung) $2 \cdot 5 = 10$ Punkte

In dieser Aufgabe beschäftigen wir uns mit verschiedenen Fragestellungen aus der Wahrscheinlichkeitstheorie. Sie sollen die Wahrscheinlichkeit von Ereignissen in verschiedenen Szenarien berechnen. Geben Sie Ihre Rechenwege an.

(a) Sie ziehen aus einem gemischten Poker-Kartenspiel 2 Karten (ohne zurücklegen). Wie hoch ist die Wahrscheinlichkeit $P(A_B)$, dass Sie zwei Bildkarten ziehen? *Hinweis:* Das Kartenspiel enthält viermal 10 Zahlkarten (1-10) und 3 Bildkarten (Bube, Dame, König), also insgesamt 52 Karten.

(b) Sie würfeln mit einem fairen sechseitigen Würfel 3 mal. Wie groß ist die Wahrscheinlichkeit $P(A_5)$, dass die größte gewürfelte Zahl genau 5 ist?

(c) Sie spielen Kopf (K) oder Zahl (Z) mit einem Partner oder einer Partnerin und einer fairen Münze. Angenommen, Sie gewinnen in der ersten Runde. Wie hoch ist die Wahrscheinlichkeit $P(W_3)$, dass Sie nach zwei weiteren Runden insgesamt häufiger gewonnen haben, als Ihr Partner oder Ihre Partnerin?

(d) In einer Urne befinden sich 12 eindeutig nummerierte Kugeln (von 1 bis 12). Sie ziehen nacheinander eine zufällige Kugel, merken sich ihre Nummer, und legen sie zurück in die Urne. Dies tun Sie insgesamt 4 Mal. Wie hoch ist die Wahrscheinlichkeit $P(A_4)$, dass Sie innerhalb der 4 Züge mindestens zweimal die gleiche Kugel ziehen?

(e) Sie verkaufen eine jährliche Produktversicherung an 1.215 Kunden und Kundinnen, die alle das gleiche Produkt zu einem Preis von 995,95 Euro gekauft haben. Wenn ein Produkt kaputt geht, zahlen Sie den Kaufpreis aus. Die Wahrscheinlichkeit, dass ein Produkt in einem Jahr kaputt geht, beträgt 0,46%. Der jährliche Versicherungsbeitrag pro Produkt kostet 12,95 Euro. Wie hoch ist Ihr erwarteter Gewinn in einem Jahr?

Aufgabe 2 (Bedingte Wahrscheinlichkeit) **3 + 3 + 4 = 10 Punkte**

In dieser Aufgabe betrachten wir verschiedene Kriminalfälle. Berechnen Sie die gesuchten Wahrscheinlichkeiten. Nennen Sie jede getroffene Annahme und geben Sie Ihre Rechenwege an.

(a) Es wurde ein Diebstahl begangen und eine Person wird verdächtigt. Sie gehört zu einer Gruppe Fußball-Hooligans, von denen bereits mehrere Diebstähle begangen wurden. Daher: a priori Wahrscheinlichkeit, dass sie schuldig ist beträgt 0.6. Weiterhin ist bekannt, dass sowohl der Dieb als auch die verdächtigte Person grüne Augen haben. Der Anteil der Personen mit grüner Augenfarbe in der Bevölkerung liegt bei 33%. Mit welcher Wahrscheinlichkeit ist die verdächtigte Person schuldig?

(b) Es wurde ein Mord begangen. Eine Person steht unter dringendem Verdacht. Am Tatort wurde die Tatwaffe mit Blutspuren der Gruppe B negativ gefunden. Die verdächtigte Person hat ebenfalls Blutgruppe B negativ. Der Anteil der Bevölkerung mit dieser Blutgruppe liegt bei 2%. Mit welcher Wahrscheinlichkeit ist der Verdächtige auf Basis dieser Fakten der Täter? Beantworten Sie die Frage (1) allgemein in Abhängigkeit von der Anzahl der theoretisch als Täter infrage kommenden Personen, n, sowie (2) für den Fall, dass alle Einwohner Berlins (3,8 Millionen) in Frage kommen.

(c) Im Mordfall aus Teilaufgabe 2 (b) gibt es neue Beweise. Eine Überwachungskamera hat den Tatort gefilmt. Auf den Aufnahmen sind 100 verschiedene Menschen zu sehen, darunter die verdächtigte Person. Es wird davon ausgegangen, dass der Mörder eine dieser 100 Personen ist. Angenommen, die a priori Chance, dass eine Person schuldig ist, liegt bei $1 : 10^6$. Der Bevölkerungsanteil mit Blutgruppe B negativ liegt bei 2%. Berechen Sie auf Grundlage aller Fakten mit welcher Wahrscheinlichkeit die verdächtigte Person schuldig ist. *Hinweis:* Sie können annehmen, dass die Chance am Tatort gefilmt zu werden bei $100 : 10^6$ liegt.

Aufgabe 3 (Metal Detektor mit Naive Bayes) **3 + 3 + 4 = 10 Punkte**

In Moodle finden Sie die Datei 'song_lyrics.zip'. Dieser Datensatz enthält Informationen über Titel verschiedener Musikgenres[1]. Er ist in zwei CSV-Dateien aufgeteilt, einen Trainings- und einen Testdatensatz. Die CSV-Dateien enthalten je Titel unter anderem den Songtext und das Genre. Nutzen Sie den Datensatz, um ein binäres Textklassifikationsmodell mit Naive Bayes zu trainieren. Es soll Titel des Genres 'Metal' anhand ihres Songtextes von Titeln anderer Genres unterscheiden können.

Implementieren Sie den Klassifikator selbst und nutzen Sie ausschließlich die Open-Source-Bibliotheken NumPy, Pandas sowie die Klasse `CountVectorizer`[2] aus *scikit-learn*.

(a) Nutzen Sie den Trainingsdatensatz, um ein "Bag of Words" Modell zu erstellen. Wandeln Sie anschließend den Trainings- und Testdatensatz mittels des erstellen Modells in numerische Features um. Die Rückgabe ist das "Bag of Words" Modell, sowie die numerischen Features des Trainings- und Testdatensatzes. Hier soll je Wort und Dokument ein Feature 1 sein, wenn das Wort im Dokument vorkommt, sonst 0.

(b) Nutzen Sie den Trainingsdatensatz, um einen Naive Bayes Klassifikator zu trainieren. Berechnen Sie dazu (1) die a priori Wahrscheinlichkeiten jeder Klasse sowie (2) die bedingten Wahrscheinlichkeiten des Vorkommens eines Worts aus dem "Bag of Words" Modell in jeder Klasse. Verwenden Sie die Laplace Glättung mit $\lambda = 1$.

(c) Nutzen Sie den zuvor trainierten Naive Bayes Klassifikator um Vorhersagen für einen Datensatz zu treffen. Bestimmen Sie für jedes Element eines Datensatzes (1) das vorhergesagte binäre Klasse (Metal / kein Metal) sowie (2) den natürlichen Logarithmus der Klassenzugehörigkeit. Die Klassenzugehörigkeit ist proportional zur Wahrscheinlichkeit, dass ein Songtext zur vorhergesagten Klasse gehört. Nutzen Sie in Ihren Berechnungen die natürliche logarithmische Skala, um die numerische Stabilität sicherzustellen.

Nutzen Sie zur Beantwortung der Fragen aus (a), (b) und (c) 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-Bibliotheken NumPy, Pandas sowie die genannte `CountVectorizer` Klasse aus *scikit-learn* nutzen.

[1] Teilmenge des Datensatzes https://www.kaggle.com/datasets/mateibejan/multilingual-lyrics-for-genre-classification

[2] Siehe https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html