B!G B4NG challenge, 19. Wettbewerb
Aufgabe 3

Diese Aufgabe wird vom Fachgebiet Software Engineering an der Fakultät für Elektrotechnik und Informatik der Leibniz Universität Hannover gestellt.

Weitere Informationen zum Fachgebiet Software Engineering findet ihr unter
http://www.se.uni-hannover.de

Big Data: Analyse großer Datensätze

Quelle: https://upload-magazin.de/files/2016/03/symbol-daten-big-data-e1555774546475-940x529.jpg

Daten gibt es heutzutage in vielen verschiedenen Bereichen zu unterschiedlichsten Zwecken. Browser erheben Daten, um Nutzerprofile zu erstellen und personalisierte Werbung zu ermöglichen, Unfallstatistiken helfen Versicherungen, angemessene Beiträge zu berechnen und Ampelschaltungen nutzen Daten über den Verkehrsfluss, um die Wartezeit zu optimieren.

Das Sammeln von Daten ist meist nicht schwer, sofern die Anwender*innen dem zustimmen. Allerdings helfen die Daten allein nicht weiter. Die Datensätze sind zu groß, um darin mit bloßem Auge etwas zu erkennen. Aus diesem Grund gibt es Analyseverfahren, die „Licht ins Dunkel“ bringen. In dieser Aufgabe geht es darum, wo Big Data vorkommt und was man mit so vielen Daten anfangen kann. Dabei werdet ihr auch einen echten Datensatz mit über 300 Variablen und fast 1500 Einträgen selbst analysieren.

---

a) Grundlagen (10 Punkte)

1) Beschreibt in eigenen Worten, was Big Data für euch ist.
2) Gebt 5 Beispiele an, in denen große Datensätze gesammelt und wofür sie genutzt werden. (Für die Beispiele, die bereits in der Einleitung genannt wurden, gibt es keine Punkte.)

In der Forschung vom Software Engineering nutzen wir KI und Methoden wie zum Beispiel neuronale Netze, wenn wir große Datensätze analysieren wollen. Ein Beispiel für einen großen Datensatz findet ihr unter
https://www.researchgate.net/publication/329246554_Complementing_Materials_for_the_HELENA_Study_Stage_2. Der Datensatz enthält fast 1500 Datenpunkte aus einer Umfrage, die Aufschluss über aktuell in der Industrie verwendete Entwicklungsansätze bieten. Mit der so genannten HELENA-Studie, in deren Rahmen der Datensatz entstanden ist, möchten wir untersuchen, wie Softwareentwicklung in der Praxis aussieht (also wie die Unternehmen die Software entwickeln) und von welchen Kontextfaktoren wie Unternehmens- oder Projektgröße der Entwicklungsansatz beeinflusst wird.

Da es praktisch unmöglich ist, Regelmäßigkeiten in derart großen Datensätzen zu finden, greifen wir auf Methoden des maschinellen Lernens zurück, die uns bei der Analyse der Daten unterstützen. Diese Methoden helfen uns, Regelmäßigkeiten und Zusammenhänge in den Daten zu sehen, die vom menschlichen Betrachter nicht ohne Weiteres entdeckt werden können.

Bei der Analyse auf Zusammenhänge und der Anwendung von maschinellen Lernverfahren ist es vor allem wichtig, die Ergebnisse richtig zu interpretieren. Ein Fehler, der dabei häufig auftritt, ist zum Beispiel die Verwechslung von Korrelation und Kausalität.

3) Beschreibt kurz in eigenen Worten die beiden Begriffe „Korrelation“ und Kausalität“ sowie deren Unterschied.
4) Gebt zwei Beispiele für Korrelationen an, die keinen kausalen Zusammenhang haben.

---

b) Methoden zur Analyse von großen Datensätzen (10 Punkte)

Für die Analyse großer Datensätze gibt es eine Vielzahl an Algorithmen, die bei der Auswertung helfen.

1) Nennt drei Methoden, die häufig bei der Analyse großer Datensätze verwendet werden. Die Methoden können, müssen aber nicht aus den Bereichen der künstlichen Intelligenz und dem maschinellen Lernen kommen.
2) Beschreibt eine der zuvor genannten Methoden genauer.

Bei vielen Methoden wird mit Trainings- und Testdaten für die Validierung gearbeitet.

3) Wofür braucht man die Validierung und warum kommen Trainings- und Testdaten zum Einsatz?

Mit diesem Vorwissen könnt ihr nun selbst einen Blick in die Daten werfen, um sie zu analysieren. Ladet euch dazu zunächst den HELENA-Datensatz vom oben angegebenen Link herunter und verschafft euch einen Überblick über die Daten. In dem Ordner findet ihr neben dem Datensatz als csv-Datei auch eine Übersicht über die gestellten Fragen (in codebooks/de/ bzw. in codebooks/“Variables Listing From SoSci.pdf“ (auf Englisch)).

4) Welche Variable (das heißt Spalte) gibt die Unternehmensgröße an? Visualisiert die Verteilung in einem Diagramm. Wie viele Personen haben die Frage beantwortet?

Wenn man eine statistische Analyse durchführen möchte, fängt man meistens mit Hypothesen an. Diese geben erwartete Zusammenhänge wieder.

5) Formuliert basierend auf den Daten und dem Codebook drei Hypothesen über erwartete Zusammenhänge und begründet, warum ihr den Zusammenhang vermutet.

---

c) KNN (10 Punkte)

Eine weit verbreitete Methode ist der k-nearest-neighbour-Algorithmus, der basierend auf Eingabewerten im Datensatz den Datenpunkt (bzw. die k Datenpunkte) sucht, die am ehesten mit den Eingabewerten übereinstimmen.

Einen Pseudocode für den KNN findet ihr zum Beispiel unter https://towardsdatascience.com/k-nearest-neighbours-introduction-to-machine-learning-algorithms-18e7ce3d802a

```
1. Load the training and test data
2. Choose the value of K
3. For each point in test data:
        - find the Euclidean distance to all training data points
        - store the Euclidean distances in a list and sort it
        - choose the first k points
        - assign a class to the test point based on the majority of
          classes present in the chosen points
4. End
```

Betrachtet den Pseudocode und versucht, nachzuvollziehen, was in den einzelnen Schritten gemacht wird.

1) Gebt die Arbeitsweise des Algorithmus in euren eigenen Worten wieder.
2) Was ist die euklidische Distanz?

Im letzten Schritt sollt ihr nun den KNN selbst programmieren.

3) Entwickelt ein Programm, das ein paar Eingabeparameter abfragt und die 3 Datenpunkte zurückgibt, die am ehesten mit den Eingabewerten übereinstimmen. Berücksichtigt dabei die folgenden Hinweise und Anforderungen:

   [1] Es bleibt euch überlassen, ob ihr den Algorithmus über Abfragen in Excel implementiert oder ob ihr Java, R oder eine andere Programmiersprache nehmt.
   [2] Es ist nicht notwendig, dass ihr eine GUI (also eine graphische Benutzeroberfläche) entwickelt. Es muss aber möglich sein, Eingaben zu machen.
   [3] Das Programm soll als Eingabewerte die Unternehmensgröße, die Projekt-/Produktgröße, die Industriedomäne (Branche), die Verteilung des Unternehmens und das Vorhandensein unternehmensweiter Prozesse (PU01) abfragen.
   [4] Die Rückgabe soll eine Liste mit verwendeten Methoden und Praktiken für jeden gefundenen Datenpunkt sein.
   [5] Als Distanz könnt ihr die Anzahl der übereinstimmenden Eingabeparameter nehmen.
   [6] Gebt den Quellcode sowie eine ausführbare Datei ab.

Viel Erfolg!

Allgemeine Hinweise

Einsendeschluss: Sonntag, 05. Januar 2020, 19:59 Uhr.

Gebt eure Lösungen über unser Portal ab: https://portal.studienberatung.uni-hannover.de/anmeldungen/users/login

Zulässige Dateiformate sind: PDF für die zusammengeschriebene Lösung (mit eingebetteten Bildern) sowie unter Windows gängige Videoformate, die sich ohne Installation von zusätzlicher Software abspielen lassen, z. B. mp4.

Die Dateien sollten nicht größer als 7,5 MB sein (die Dateien können gezippt sein)! Bitte gebt auch euren Teamnamen, die Namen der Gruppenmitglieder sowie deren Schulen an. Bitte benennt eure hochgeladenen Dateien nach dem Gruppennamen.

ACHTUNG bei Zip-Dateien! Um sicherzugehen, dass eure Dateien wirklich fehlerfrei und für die Korrektor*innen zu öffnen sind, solltet ihr eure Zip-Dateien etc. noch mal von eurem Account herunterladen und öffnen. Dateien, die sich nicht öffnen lassen, können nicht bewertet werden!

Gebt eure Lösungen auch dann ab, wenn ihr nicht alle Fragen beantworten konntet! Vielleicht gelingt euch das ja bei der kommenden Aufgabe.

Die Teilnahmebedingungen und weitere Informationen findet ihr unter: www.uni-hannover.de/bigbangchallenge

Der Rechtsweg ist ausgeschlossen.