

Maschinelles Lernen

Aufgabenblatt 01

Prof. Dr. David Spieler
Hochschule München

14. Oktober 2019

Organisatorisches

Arbeitsumgebung

Um mit den praktischen Übungen beginnen zu können, benötigen Sie eine entsprechende Arbeitsumgebung. Wir verwenden die Programmiersprache **Python 3.7** innerhalb der interaktiven Notebook-Umgebung **Jupyter**. Die Distribution **Anaconda** bringt Python, Jupyter und alle relevanten Bibliotheken bereits mit. Sie finden Anaconda zum Download unter <https://www.anaconda.com/download> für Linux, Mac OS und Linux. Sie werden sich im Verlauf des Praktikums mit der Programmiersprache Python und den Bibliotheken

- **numpy** (schnelle n-dimensionale Arrays, lineare Algebra)
- **pandas** (schnelle Datenstrukturen und Analyse)
- **matplotlib** (Datenvisualisierung)
- **seaborn** (statistische Datenvisualisierung aufbauend auf matplotlib)
- **scikit-learn** (Machine Learning)

beschäftigen. Für alle diese Bibliotheken ist eine hervorragende Dokumentation online verfügbar.

Aufgabe 1 (Datenexploration und Visualisierung) *In dieser Aufgabe beschäftigen wir uns mit dem **College** Datensatz, welchen Sie im Moodle finden bzw. von der Quelle <http://www-bcf.usc.edu/~gareth/ISL/data.html> beziehen können. Es handelt sich um eine einzige CSV-Datei **College.csv** mit 777 Datensätzen über amerikanische Universitäten und Colleges. Jede Zeile entspricht einem Datensatz mit je 19 Features, welche durch Kommata getrennt sind – CSV steht für comma separated values. Diese Features sind in Tabelle 1 aufgeführt. Erstellen Sie ein neues Jupyter-Notebook und laden Sie die Pakete wie im folgenden Skript angegeben.*

1. Laden Sie die CSV-Datei in einen Pandas **DataFrame** mit Hilfe von `pandas.read_csv()`.
2. Geben Sie den so erhaltenen DataFrame in einer Zelle aus.
3. Erklären Sie in eigenen Worten, was ein DataFrame ist?
4. Führen Sie die Methode `describe` auf dem DataFrame aus.
5. Erläutern Sie das Ergebnis, was bedeuten die einzelnen Zeilen?
6. Extrahieren Sie in eine neue Variable nur einen Teil des DataFrames mit den ersten zehn Features.
7. Importieren Sie die Funktion `scatter_matrix` von `pandas.plotting` und führen sie die Funktion auf das Extrakt des DataFrames mit den ersten zehn Features aus.
8. Was macht `scatter_matrix`?
9. Mit den zusätzlichen Parametern `figsize=(15,15)`, `s=50`, `marker='D'` bekommen Sie eine bessere Darstellung. Was bedeuten die Zusatzparameter?
10. Erstellen Sie mit `matplotlib.bboxplot` einen Boxplot des Features `Outstate`.
11. Interpretieren Sie den Boxplot.
12. Erstellen Sie mit `seaborn.countplot` einen CountPlot des Features `Private`.
13. Interpretieren Sie den CountPlot.
14. Sie können mit `df['NeuesFeature'] = ...` einem DataFrame auch ein neues Feature hinzufügen. Tun Sie das mit einem neuen boolschen Feature `Elite`, welches genau dann wahr ist, wenn bei einer Hochschule mehr als 50% der neuen Studenten unter den Top 10% der jeweiligen High School waren.
15. Erstellen Sie danach einen CountPlot von `Elite`.
16. Wie viele Elite-Hochschulen gibt es in etwa im Datensatz?

Feature	Bedeutung
<Leerstring>	Name der Universität bzw. des Colleges
Private	Indikator, ob es sich um eine private Einrichtung handelt
Apps	Anzahl der erhaltenen Bewerbungen
Accept	Anzahl der angenommenen Bewerber
Enroll	Anzahl der Neueinschreibungen
Top10perc	Anteil der neuen Studenten, die zu den 10% besten ihrer High School gehören (in Prozent)
Top25perc	Anteil der neuen Studenten, die zu den 25% besten ihrer High School gehören (in Prozent)
F.Undergrad	Anzahl der Vollzeit-Studierenden
P.Undergrad	Anzahl der Teilzeit-Studierenden
Outstate	Studiengebühren für Studierende von außerhalb des Bundesstaats
Room.Board	Kosten für Miete und Verpflegung
Books	Kosten für Bücher (geschätzt)
Personal	Persönliche Ausgaben (geschätzt)
PhD	Anteil der Mitarbeiter mit Dokortitel (in Prozent)
Terminal	Anteil der Mitarbeiter mit berufsqualifizierendem Abschluss (in Prozent)
S.F.Ratio	Verhältnis Anzahl Studenten zu Mitarbeiter
perc.alumni	Anteil der Alumni, die spenden (in Prozent)
Expend	Ausgaben für die Lehre pro Student
Grad.Rate	Anteil der Absolventen mit Abschluss (in Prozent)

Tabelle 1: Features des College Datensatzes.