Alfatraining
Machine Learning Kurs
Abschlussprojekt
Dozent: Mathias Schmerling

Projekt: Kartografierung des Sonnensystems

06. April 2021 - 09. April, 2021

Projektbeschreibung

In unserem Sonnensystem sind, Stand 17. August 2020, bereits 991,909 Asteroiden bekannt (Quelle: link). Ziel dieses Projekts ist es, zu untersuchen, ob sich mit den Methoden des Machine Learning die Kartografierung des Sonnensystems erleichtern lässt. Dazu untersuchen Sie, ob sich die Größe von Asteroiden anhand anderer Kenngrößen vorhersagen lässt. Betrachten Sie dazu den folgenden öffentlich verfügbaren Datensatz, der auf der Datenbank des *Jet Propulsion Laboratory* der NASA beruht: Asteroid Dataset.
Neben verschiedenen anderen Kenngrößen, mit denen Sie sich selbstständig vertraut machen sollen, enthält der Datensatz für jeden Asteroid eine Variable *diameter*, also den Durchmesser des Himmelskörpers, der die Zielvariable Ihres Vorhersagemodells sein soll. Ihre Aufgabe ist es, eine Exploration des Datensatzes vorzunehmen und anschließend mit einem geeigneten Modell die *diameter*-Variable vorherzusagen. Neben einem Modell sollen Sie auch eine Einschätzung der Generalisierungsfähigkeit des Modells bereitstellen, sowie einen Bericht ausarbeiten, der Ihr Vorgehen und ihre Entscheidungen dokumentiert.

Aufgaben

1. **Datenexploration** Explorieren Sie die Daten auf eine in Ihren Augen geeignete Weise. Folgende Aspekte könnten unter anderem dabei relevant sein:
   - Dateiformat(e), Anzahl der Datenpunkte und Messwerte
   - Art der Messwerte und Verteilungen der Messwerte
   - Missing Values
   - Korrelationen zwischen den Features und zwischen den Features und der Zielvariablen.
2. **Datenrepräsentation** Repräsentieren Sie die Daten auf eine für Ihr Modell geeignete Weise. Untersuchen Sie, welche Vorverarbeitungen sinnvoll oder nötig sind.
3. **Training** Trainieren Sie eines oder mehrere Modelle zur Vorhersage der Zielvariablen.
4. **Evaluation** Evaluieren Sie das Modell und schätzen Sie die Generalisierungsfähigkeit des Modells. Diskutieren Sie das Ergebnis in Hinblick auf Over- bzw. Underfitting.

5. ***Bonus*** *Einige der Merkmale im Datensatz sollten eigentlich nicht für die Vorhersage verwendet werden. Untersuchen Sie den Teil der Daten, für die die Zielvariable unbekannt ist, in Hinblick auf fehlende Werte. Was fällt Ihnen auf? Recherchieren Sie, wie der Durchmesser von Asteroiden normalerweise bestimmt wird. Warum sollten bestimmte Merkmale eigentlich nicht verwendet werden? Treffen Sie eine Vorhersage auf dem Teil der Daten mit unbekannter Zielvariable. Vergleichen Sie Modelle mit und ohne diese Merkmale.*

Ergebnisse

Einzureichen und vorzubereiten sind:

- Code (Jupyter-Notebook)
- Bericht (Jupyter-Notebook)
- Präsentation (5-7 Minuten)

Die Ergebnisse Ihrer Untersuchungen sollten Sie in einem Bericht zusammenfassen. Dieser Bericht sollte sowohl Ihre Vorgehensweise als auch Ihren Code dokumentieren. Verwenden Sie dafür am besten ein Jupyter-Notebook, das Sie einmal als Notebook (`.ipynb`) und einmal als `.html` einreichen.
Neben lesbarem und gut dokumentiertem Code sollten aus Ihrem Bericht auch Ihre Entscheidungsprozesse hervorgehen. Dokumentieren Sie stets, für welches Vorgehen Sie sich entschieden haben und warum. Dazu gehört zum Beispiel:

- Welche Explorationen haben Sie vorgenommen und warum?
- Haben Sie sämtliche Daten verwendet oder nur einen Teil? Warum oder warum nicht?
- War es nötig die Daten vorzuverarbeiten? Auf welche Weise? Warum oder warum nicht? Wie hängt das mit dem Modell zusammen?
- Für welches Modell haben Sie sich entschieden und warum? Handelt es sich bei dem Problem um eine Regression oder Klassifikation? Welche Hyperparameter haben Sie wie eingestellt und warum? Haben Sie mehrere Modelle getestet? Warum oder warum nicht?
- Auf welche Weise haben Sie die Generalisierung geschätzt? Welche Metrik haben Sie verwendet? Welche Alternativen gibt es?

Präsentieren Sie anschließend Ihre Ergebnisse auf Grundlage Ihres Notebooks den anderen TeilnehmerInnen (5-7 Minuten Präsentationszeit). ‘