Programmentwurf – Maschinelles Lernen

Prof. Dr. Alexander Dück

Übersicht

In diesem Projekt werden Sie an einem tabellarischen Datensatz zum Thema Brustkrebs arbeiten. Der Datensatz enthält Features aus digitalisierten Bildern einer Feinnadelbiopsie von Brusttumoren. Sie charakterisieren die Merkmale der im Bild enthaltenen Zellkerne. Ziel ist die Klassifizierung des Tumors in “gutartig” (benign) und “bösartig” (malignant). Beachten Sie, dass der Datensatz fehlende Werte enthalten kann, die während Ihrer Vorverarbeitungsschritte behandelt werden müssen.
Bitte betrachten Sie dazu den Ihnen zugeteilten individuellen Datensatz. Dieser kann bzgl. der Attribute und der tatsächlichen Datenpunkte abweichen.
Falls nötig erhalten Sie weitere Details zum Datensatz unter https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic .

Aufgaben

A. Wenden Sie die folgenden ML-Algorithmen an:
   a. Multilayer-Perceptron (Fully Connected Neural Network)
   b. Logistische Regression
   c. Support Vector Machine

   Das Ziel ist es vorherzusagen, ob der Tumor “gutartig” (benign) oder “bösartig” (malignant) ist. Die entsprechende Spalte im Datensatz ist “Diagnosis”.

B. Bereiten Sie dazu die Daten geeignet vor, um sie durch die Lernverfahren gut nutzbar zu machen.

C. Bewerten Sie das Lernergebnis bzgl. der diskutierten Qualitätskriterien und vergleichen Sie die Modelle und verschiedene Konfigurationen zur Optimierung des Ergebnisses.

Verwenden Sie geeignete Bibliotheken wie pandas, sklearn, TensorFlow, PyTorch oder andere freie und offene Bibliotheken, um dieses Projekt abzuschließen.

Abgabe

Reichen Sie ein Jupyter-Notebook ein, das Folgendes enthält:

1. Python-Code zum Einlesen und Vorverarbeiten der Daten, Anwenden der oben genannten ML-Algorithmen und Evaluieren dieser mit geeigneten Metriken.
2. Dokumentation Ihrer Konfigurationsentscheidungen für jeden Algorithmus, sowie der Vergleich, die Analyse und ein Fazit des Ergebnisses.

Die Abgabe erfolgt über das Moodle System.

Bewertungskriterien

Die Bewertung setzt sich aus den folgenden Kriterien zusammen:

1. Korrektheit und Strukturiertheit (inkl. Kommentierung) des Programmcodes
2. Plausibilität und Qualität des Lernergebnisses
3. Angemessene Datenvorbereitung
4. Passende und gut begründete Konfiguration und Aufgabenanpassung der Lernverfahren
5. Darstellung und Interpretation der Lernergebnisse