Programmierabgabe 1 - Gruppe 8

Klassifikation von gesprochenen Buchstaben mit Scikit Learn

Datenset

Das Datenset "ISOLET" besteht aus einer Matrix von 617 Merkmalen, die aus den Audioaufnahmen von gesprochenen Buchstaben extrahiert wurden. Dazu wurde von 150 Probanden jeder Buchstabe des Alphabets zweimal aufgenommen, d.h. es liegen ca. 7800 Samples vor. Die Anzahl der Klassen in diesem Klassifikationsproblem beträgt 26, je eine für jeden Buchstaben des Alphabets.

Sie können das Datenset unter `https://datahub.io/machine-learning/isolet` downloaden. Laden Sie die Daten in Form eines csv-Files herunter, dieses lässt sich einfach in Python importieren. Das csv-File beinhaltet eine Merkmalsmatrix, in der die Zeilen je ein Datensample darstellen und die Spalten je ein Merkmal. Die letzte Spalte enthält die Information über den gesprochenen Buchstaben, d.h. ein Klassenlabel zwischen 1 und 26.

Aufgabe

Implementieren Sie ein ML-System zur Klassifikation der gesprochenen Buchstaben mithilfe der Python-Bibliothek Scikit Learn. Bearbeiten Sie dabei die folgenden Schritte:

1. Laden der Daten und Splitting in Trainings- und Testdaten (Verhältnis 75:25).

2. Wählen Sie zur Merkmalsreduktion je eine Methoden aus dem Bereich Merkmalstransformation und Merkmalsselektion aus. Vergleichen Sie die Performance Ihres ML-Systems mit der gewählten Transformation und der gewählten Selektion.

3. Implementieren Sie mithilfe von Scikit Learn zwei unterschiedliche Klassifikatoren zur Lösung des Klassifikationsproblems: k-Means Clustering und Random Forest. Vergleichen Sie die Performance der beiden Klassifikatoren miteinander. *Hinweis: k-Means ist ein unüberwachter Klassifikator.*

4. Evaluieren Sie Ihr ML-System und stellen Sie die Ergebnisse in geeigneter Weise dar. Nutzen Sie zur Evaluation eine Kreuzvalidierung (3-fold).