

Programmierabgabe 1 - Gruppe 1

Regression zur Bewertung von Weinqualität mit Scikit Learn

Datenset

Das Wine Quality Data set beinhaltet 11 chemische Eigenschaften für roten (1599 Samples) oder weißen (4898 Samples) "Vinho Verde" Wein aus Portugal. Jedem Sample ist ein Qualitäts-Score als Label zugeordnet. Beachten Sie, dass es sich bei dem Datenset um ein unbalanced Datenset handelt, d.h. dass sie dies bei Ihrem Splitting in Trainings- und Testdaten sowie den gewählten Evaluationsmethoden berücksichtigen müssen. Führen Sie alle Aufgabenschritte sowohl für weißen als auch für roten Wein durch.

Sie können das Datenset unter <https://archive.ics.uci.edu/ml/datasets/wine+quality> downloaden.

Aufgabe

Implementieren Sie ein ML-System zur Regression für die Vorhersage der Weinqualität mithilfe der Python-Bibliothek Scikit Learn. Bearbeiten Sie dabei die folgenden Schritte:

1. Laden Sie die Daten und führen Sie ein Splitting in Trainings- und Testdaten (Verhältnis 75:25) durch.
2. Führen Sie eine Principal Component Analysis zur Merkmalsreduktion durch.
3. Implementieren Sie mithilfe von Scikit Learn zwei unterschiedliche Regressionsverfahren: Nearest Neighbor Regression und Support Vector Regression. Vergleichen Sie die Performance der beiden Regressionsmodelle miteinander.
4. Evaluieren Sie Ihr ML-System und stellen Sie die Ergebnisse in geeigneter Weise dar. Nutzen Sie zur Evaluation eine Kreuzvalidierung (3-fold). Vergleichen Sie die Ergebnisse für roten und weißen Wein miteinander.