

Special Engineering – IKT (M)

1. Projekt – Boston Housing Dataset

Für den Datensatz “Boston Housing Dataset” soll ein Regressionsmodell für den Median der Hauspreise in einer Nachbarschaft erstellt, ausgewertet und dokumentiert werden. Das Modell soll ausschließlich mithilfe der Routinen aus dem 2. und 3. Übungsblatt trainiert und ausgewertet werden. Weiterhin können die Pakete `numpy` und `matplotlib` verwendet werden.

Abzugeben sind

1. ein Paket `lineare_regression.py` mit den Routinen des 2. und 3. Übungsblattes
2. eine Dokumentation und Programmcode des Modells und aller Auswertungen in Form
 - eines PDFs und eines lauffähigen Python-Skripts **oder**
 - eines lauffähigen Jupyter-Notebooks.

Das Projekt ist bis zum Abgabetermin des aktuellen Prüfungszeitraums auf Moodle einzureichen. Es gelten die allgemeinen Kriterien für eine schriftliche Arbeit. Achten Sie z.B. darauf, ganze Sätze zu formulieren und Ihre Entscheidungen zu begründen, bzw. zu belegen.

Gehen Sie wie folgt vor:

1. Importieren Sie den Datensatz (siehe 2. Übungsblatt) und teilen Sie ihn in Trainings-, Validierungs- und Testdaten ein.
2. Erkunden Sie den Trainingsdatensatz. Visualisieren Sie zum Beispiel die verschiedenen Features und den Zusammenhang mit dem Preis. Beschreiben Sie Ihre Beobachtungen.
3. Bestimmen Sie für polynomiale Features ($p = 1$ und $p = 2$) jeweils den optimalen Parameter für eine regularisierte lineare Regression und den entsprechenden Fehler. Wählen Sie das beste Modell aus und begründen Sie Ihre Entscheidung.
4. Erstellen Sie ein entsprechendes Modell auf den gesamten Trainings- und Validierungsdaten und evaluieren Sie es auf den Testdaten. Geben Sie Kennzahlen an und visualisieren Sie Ihr Ergebnis. Diskutieren Sie Ihr Resultat.
5. Untersuchen Sie Ihr Modell mithilfe einer Lernkurve. Diskutieren Sie, wie Sie Ihr Modell verbessern könnten.