9. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 09.01.2017, 23:59 Uhr

WS 2017/2018
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD1718]
Di. 10-12	CP-03-150	philipp2.hoffmann@udo.edu und jan.soedingrekso@udo.edu
Di. 16-18	P1-02-110	felix.neubuerger@udo.edu und tobias.hoinka@udo.edu
Di. 16-18	CP-03-150	simone.mender@udo.edu und maximilian.meier@udo.edu

Aufgabe 28: *Data-Mining Challenge* **20 P.**

In dieser Aufgabe soll eine reale Problemstellung aus dem Gebiet des Immobiliengeschäfts behandelt werden. Es handelt sich dabei um die Schätzung von Haus-Verkaufspreisen. Es steht Ihnen wie immer frei, wie viel Zeit und Anstrengung Sie für diese Aufgabe verwenden, es wird jedoch **Amazon-Gutscheine** für die besten Regressionen zu gewinnen geben.

Problemstellung

Die Problemstellung ist ausführlich hier beschrieben:
`https://www.kaggle.com/c/house-prices-advanced-regression-techniques`
Kaggle ist eine Plattform, auf der Datensätze und Problemstellungen hochgeladen werden können und Benutzer versuchen die besten Modelle für die Probleme zu finden.

Datensatz

Die Daten finden Sie im Moodle, sie sind bereits vorverarbeitet. Wie der Problemstellung zu entnehmen ist, kann es sinnvoll sein, eigene Attribute zu erzeugen und ggfs. zusätzliche Informationen in den Datensatz mit einzubringen.

Auswertung

Die Qualität der Regression wird über das *Kaggle* Leaderboard bestimmt. **Dafür muss eine CSV-Datei mit der ID der Häuser und dem erwarteten Verkaufspreis erstellt werden.** Die Datei `sample_submission.csv` zeigt die korrekte Formatierung. Sie reichen die CSV-Datei und alle Skripte/Prozesse wie gewohnt über das Moodle ein. Wir werden dann Ihre finalen Regressionen auf *Kaggle* hochladen. Die drei Regressionen mit den höchsten Rängen im Leaderboard gewinnen Preise. Beachten Sie, dass für die Aufgabe ebenfalls Punkte für den Übungsbetrieb verteilt werden, die Gutscheine sind nur ein Bonus.

Hinweise

Benutzen Sie Python für die Lösung. Die Bibliotheken `pandas` und `sklearn` können dabei hilfreich sein. Sollte `sklearn` benutzt werden, achten Sie darauf die Daten geeignet zu formatieren (`numpy.array`, `float32`).

Sie müssen sich keine *Kaggle* Account für diese Aufgabe erstellen. Alle erforderlichen Daten sind im Moodle bereitgestellt.

Zur Vorbereitung gibt es zusätzlich kleine Aufgaben, um sich mit dem Datensatz vertraut zu machen:

a) Lesen Sie die Datei `train.csv` ein. Bestimmen Sie die drei Attribute mit der höchsten Korrelation zum Attribute `SalePrice`. Stellen Sie die Korrelationen jeweils in einem Scatter-Plot dar.

b) Führen Sie eine lineare Regression zwischen dem Attribut mit der höchsten Korrelation und dem Attribut `SalePrice` durch. Stellen Sie die Regressionsgerade zusammen mit den Datenpunkten dar.

c) Stellen Sie den relativen Abstand zwischen den geschätzten Verkaufspreisen aus der linearen Regression in Teil **b)** und den wahren Verkaufspreisen in einem Histogramm dar.

d) Anschließend die Aufgabe der Data-Mining Challenge: Führen Sie eine Regression mit Methoden Ihrer Wahl im Rahmen einer Validierung auf den Trainingsdaten (`train.csv`) aus und wenden das Model auf den Testdatensatz (`test.csv`) an. Speichern Sie die regressierten Verkaufspreise zusammen mit den IDs der Häuser in einer `csv`-Datei gemäß der Beispieldatei `sample_submission.csv` von *Kaggle* ab.

Geben Sie jeglichen Code, der zur Lösung genutzt wurde mit ab. Zusätzlich muss eine korrekt formatierte `csv`-Datei mit den finalen Regressionsergebnissen abgegeben werden.