Maschinelles Lernen
Aufgabenblatt 02

Prof. Dr. David Spieler
Hochschule München

14. Oktober 2019

Aufgabe 1 (Lineare Regression und Gradientenabstiegsverfahren) *In dieser Aufgabe erstellen Sie ein Modell mit Hilfe (eindimensionaler) linearer Regression und trainieren das Modell selbst mit Hilfe des Gradientenabstiegsverfahren. Als Trainingsdaten verwenden wir einen Datensatz mit der Anzahl der verkauften Eiskugeln bei bestimmten Außentemperaturen angelehnt an den Datensatz* `https://kenandeen.wordpress.com/2015/01/21/dissecting-a-dataset/`.

1. *Laden Sie die CSV* `IceCream.csv` *in einen Pandas DataFrame.*
2. *Erstellen Sie mit* `matplotlib.pyplot.scatter` *einen Scatterplot der* `SoldIceCream` *über die* `Temperature`*. Beschriften Sie die Achsen passend.*
3. *Beschreiben Sie kurz den Zusammenhang der beiden Features.*
4. *Implementieren Sie eine Funktion* `train(X, Y, steps, eta)` *welche den Lernalgorithmus für die eindimensionale lineare Regression mit Hilfe des Gradientenabstiegsverfahren darstellen soll. Die Parameter* `X` *und* `Y` *sind Listen der Eingabe- bzw. Ausgabewerte,* `steps` *ist die Anzahl der Lernschritte und* `eta` *ist die Lernrate. Die Funktion soll die beiden Gewichte* $\mathbf{w}_0$ *und* $\mathbf{w}_1$ *zurückgeben.*
5. *Trainieren Sie ein lineares Regressionsmodell mit Hilfe dieser Funktion auf dem Feature* `Temperature` *und der Ausgabe* `SoldIceCream` *mit* 200000 *Schritten und Lernrate* $\eta = 0.0001$*. Geben Sie* $\mathbf{w}_0$ *und* $\mathbf{w}_1$ *aus.*
6. *Was bedeuten die Gewichte konkret in diesem Fall?*
7. *Implementieren Sie eine Funktion* `predict(x, w0, w1)` *welche eine Vorhersage für die Eingabe* `x` *unter den Modellparametern* `w0` *und* `w1` *zurückgibt.*
8. *Berechnen Sie den kleinste und größte Temperatur, die in den Daten vorkommt als Variablen* `xmin` *und* `xmax` *und die entsprechenden Vorhersagen des Modells als Variablen* `ymin` *und* `ymax`*.*

9. *Wiederholen Sie den Scatterplot und zeichnen Sie zusätzlich via* `matplotlib.pyplot.plot` *die Modellvorhersage als Linie mit Hilfe der zuvor berechneten Variablen* `xmin`, `xmax`, `ymin` *und* `ymax` *ein.*

10. *Interpretieren Sie den Plot.*

Aufgabe 2 (SciKit-Learn, R^2 und mehrdimensionale Regression) *Die* `Boston` *Housing Daten beschreiben den Median der Hauspreise ca. um 1978 in Boston (Einheit: 1000$) abhängig von den Features wie dargestellt in Tabelle 1. Sie werden den Umgang mit SciKit-Learn üben, Modelle anhand der R^2-Statistik vergleichen und mehrdimensionale Regressionsmodelle anwenden.*

1. *Laden Sie den Boston Housing Datensatz mit Hilfe von* `load_boston` *aus* `sklearn.datasets` *in eine Variable* `boston`.

2. *Was ist* `boston`*?*

3. *Laden Sie die Boston Feature-Daten in einen DataFrame namens* `X`.

4. *Zeigen sie die ersten Zeilen von* `X` *mit Hilfe von* `X.head()` *an.*

5. *Was stellen Sie fest?*

6. *Laden Sie die Beschriftung aus dem* `boston` *Objekt nach und verifizieren Sie das Ergebnis.*

7. *Laden Sie die Ausgabewerte aus dem* `boston` *Objekt in einen neuen DataFrame* `y` *und nennen Sie die Spalte* `MEDV` *(Median Value). Verifizieren Sie das Ergebnis mit* `y.head()`.

8. *Fügen Sie die beiden DataFrames* `X` *und* `y` *mit Hilfe von* `pd.concat([X, y], axis=1, sort=False)` *in einem neuen DataFrame* `full` *zusammen und erstellen Sie eine Scatter-Matrix.*

9. *Bei welchen Features vermuten Sie einen direkten Zusammenhang mit den Hauspreisen?*

10. *Erstellen Sie einen Scatterplot der Hauspreise über das Feature* `LSTAT`. *Achten Sie auf eine sinnvolle Achsenbeschriftung.*

11. *Erstellen Sie die Variable* `simple_model` *als neues* `sklearn.linear_model.LinearRegression` *Objekt und trainieren Sie das lineare Regressionsmodell via* `simple_model.fit` *nur mit dem Feature* `LSTAT` *auf die Ausgabewerte* `y`. ***Tipp**: Mit* `X[['LSTAT']]` *erhalten Sie einen DataFrame, welcher nur die Spalte* `LSTAT` *enthält.*

12. *Lesen Sie aus dem Modell die beiden Parameter aus.*

13. *Erstellen Sie eine neue Abbildung mit dem Scatterplot der Hauspreise über das Feature* `LSTAT` *und zeichnen Sie eine Gerade ein mit Hilfe der zuvor berechneten Modellparametern.*

Feature	Bedeutung
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population

Tabelle 1: Features des Boston Housing Datensatzes. Beschreibung übernommen von `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`.

14. *Interpretieren Sie den Plot. Glauben Sie, dass der Zusammenhang tatsächlich linear ist?*

15. *Berechnen Sie mit Hilfe von* `simple_model.score` *den* R^2*-Wert.*

16. *Interpretieren Sie den Wert.*

17. *Wiederholen Sie die lineare Regression diesmal im Mehrdimensionalen mit den beiden Features* `LSTAT` *und* `RM`. ***Tipp**: Mit* `X[['LSTAT','RM']]` *erhalten Sie einen DataFrame, welcher nur die Spalten* `LSTAT` *und* `RM` *enthält. Berechnen Sie den* R^2*-Wert.*

18. *Fitten Sie ein neues lineares Regressionsmodell mit allen Features und berechnen Sie den* R^2*-Wert.*

19. *Trennen Sie mit Hilfe von* `train_test_split` *aus* `sklearn.model_selection` *den kompletten Datensatz zufällig in einen Trainingsdatensatz (80%) und einen Testdatensatz (20%). Setzen Sie dabei den* `random_state` *auf* 0.

20. *Was ist nun der entsprechende* R^2*-Wert, warum und was bedeutet das?*