Fachhochschule Graubünden
University of Applied Sciences

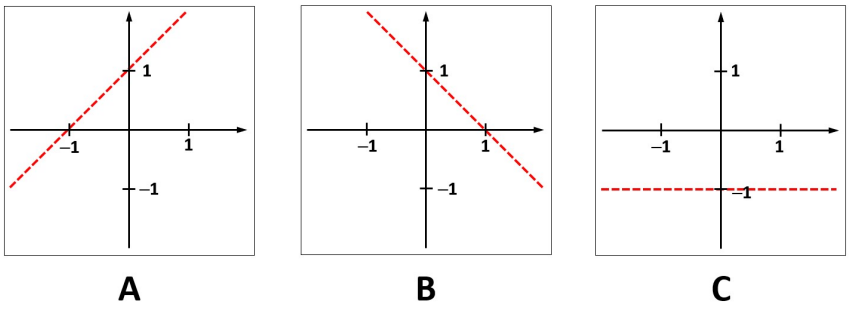
Introduction to Data Science HS 2021

Aufgabenblatt 1: Lineare Regression

Die Bearbeitung der Aufgaben ist freiwillig; es erfolgt keine Bewertung.

Aufgabe 1

a) Gegeben sind drei verschiedene Regressionsgeraden (rot-gestrichelte Linie). Bestimmen Sie für jede der Geraden A, B und C die beiden Parameter β_0 und β_1 durch Ablesen aus der Grafik.



Vorbemerkung:

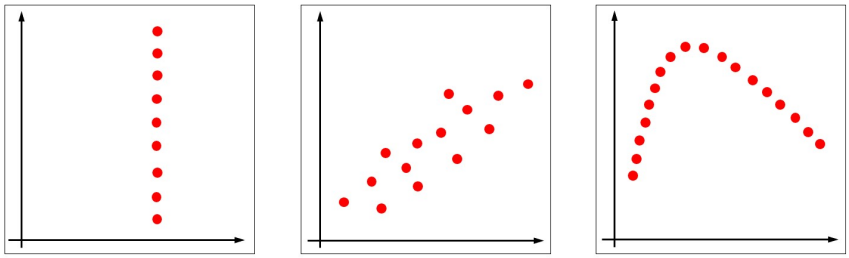
i. Achsenabschnitt β_0: Schnittpunkt der Geraden mit der y-Achse

ii. Steigung β_1: Verhältnis $\Delta y/\Delta x$ der Differenz zweier Punkte in x- und y-Richtung

Lösung:

- Gerade A: $\beta_0 = 1, \beta_1 = 1$
- Gerade B: $\beta_0 = 1, \beta_1 = -1$
- Gerade C: $\beta_0 = -1, \beta_1 = 0$

b) Gegeben sind folgende Datenerhebungen A, B und C mit den zugehörigen Datenpunkten. Entscheiden Sie für jede Erhebung, ob eine Analyse mittels linearer Regression möglich ist und begründen Sie kurz Ihre Antwort, falls dies nicht geht.



– Lösung –

- Bei Datenerhebung A ist keine lineare Regression möglich. Zwar liegen alle Punkte auf einer Geraden, jedoch ist die Steigung der zugehörigen Regressionsgeraden unendlich, d.h. damit lässt sich keine brauchbare Vorhersage machen. Ausserdem liefert das Modul **`LinearRegression`** aus dem Paket **`sklearn.linear_model`** hier falsche Ergebnisse (→ Zweifel…? Ausprobieren…! ☺).
- Bei Datenerhebung B ist eine lineare Regression möglich und sinnvoll.
- Bei Datenerhebung C ist eine lineare Regression zwar möglich, jedoch gibt es offensichtlich keinen linearen Zusammenhang, insofern wird das Ergebnis nur einen sehr niedrigen R^2-Wert aufweisen.

Aufgabe 2

Der „California Housing"-Datensatz aus dem Paket **`sklearn.datasets`** beinhaltet verschiedene Informationen aus 20.640 Haushalten in Kalifornien. Laden Sie diesen Datensatz und nutzen Sie daraus die Spalten 2 (durchschnittliche Anzahl an Räumen) und 3 (durchschnittliche Anzahl an Schlafzimmern) für eine Regressionsanalyse. Gehen Sie dazu wie folgt vor:

```
>>> from sklearn import datasets
>>> data = datasets.fetch_california_housing()
>>> data_x = data.data[:,2]
>>> data_y = data.data[:,3]
```

Teilen Sie die Daten im Verhältnis 80:20 in Trainings- und Testdaten auf und führen Sie damit eine Regressionsanalyse durch, um festzustellen, ob es einen linearen Zusammenhang zwischen Anzahl an Räumen und Anzahl an Schlafzimmern gibt. Bestimmen Sie für den Testdatensatz zudem den MSE sowie R^2-Wert und plotten Sie Ihre Ergebnisse in einem Diagramm.

– Lösung –

```
>>> x_train = data_x[:-4128]
>>> x_test = data_x[-4128:]
>>> y_train = data_y[:-4128]
```

```
>>> y_test = data_y[-4128:]
>>> from sklearn.linear_model import LinearRegression as lr
>>> model = lr()
>>> model.fit(x_train.reshape((-1,1)), y_train)
>>> y_pred = model.predict(x_test.reshape((-1,1)))
>>> from sklearn.metrics import mean_squared_error as mse
>>> from sklearn.metrics import r2_score
>>> mse(y_test, y_pred, squared=True)
0.0530112229657212
>>> r2_score(y_test, y_pred)
0.2270383929711829
>>> import matplotlib.pyplot as plt
>>> plt.plot(x_test, y_test, 'ro')
>>> plt.plot(x_test, y_pred, 'b-')
>>> plt.show()
```