Lineare und Logistische Regression
Dr. Lucia Kleint
HS 2021

Fachhochschule Nordwestschweiz

Übungsblatt LE1

Lineare und Logistische Regression

Aufgabe 1.
Geben Sie an, ob die folgenden Variablen kategorial oder kontinuierlich sind:

(a) Familienstand

(b) Alter

(c) Körpergewicht

Aufgabe 2.
Die folgenden Werte seien gegeben:

x	-2	-1	3	4	6
y	0	0.5	2	2	5

(a) Stellen Sie das lineare Gleichungssystem $Ax = b$ auf und geben Sie die einzelnen Vektoren und Matrizen an.

(b) Berechnen Sie die Regressionsgerade.

(c) Erstellen Sie einen Plot und zeichnen Sie sowohl die Daten, wie die Regressionsgerade und die Residuen ein.

Aufgabe 3.
Als Data Scientist haben Sie den Auftrag, in einer Firma herauszufinden, welche Mitarbeiter öfters abwesend sind, wovon dies abhängt und wie man Absenzen voraussagen kann. Dazu erhalten Sie den Datensatz `employees.csv` (Qulle: fiktiver Datensatz von kaggle.com), den Sie vorzugsweise mittels Python analysieren.

(a) Daten sind oft nicht bereinigt. Die Firma sagt Ihnen, dass der Praktikant beim Erheben des Datensatzes bestimmt ein paar Tippfehler gemacht hat. Ihnen wird auch gesagt, dass jeder Mitarbeiter mindestens einmal gefehlt hat. Bereinigen Sie den Datensatz.

(b) Wovon hängen die Absenzen eher ab, von der Anzahl Jahren im Betrieb oder dem Alter der Mitarbeiter? Führen Sie jeweils eine Regression durch.

(c) Plotten Sie die Residuen mittels Tukey-Anscombe Digrammen. Wie erklären Sie die scharfe Kante im Residuenplot des Alters?

(d) Prüfen Sie, ob die Bedingungen der Residuenanalyse erfüllt sind.

(e) Berechnen und interpretieren Sie die R^2 Werte für die beiden Fits.

(f) Was würden Sie der Firma raten?

Aufgabe 4.
Zwei Fachexperten führen abwechslungsweise die Prüfung im gleichen Fach durch. Die folgende Statistik an Prüfungsresultaten wurde erstellt

	Fachexperte	
	A	B
bestanden	45	21
nicht bestanden	16	15

Hängen die Chancen zum Bestehen vom Fachexperten ab?

Aufgabe 5.
Eine grosse Befragung der Bevölkerung wurde durchgeführt und ihre Resultate sind in der Datei `Umfragedaten_v1_an.csv` (Quelle: allgemeine Bevölkerungsumfrage der Sozialwissenschaften 2014 (ALLBUS 2014)). Als Data Scientist sollen Sie klären, ob das Rauchverhalten der Bevölkerung vom Einkommen abhängt.

(a) Berechnen Sie ein logistisches Modell

(b) Benützen Sie das Modell, um die Wahrscheinlichkeit zu berechnen, dass eine Person, die 2000 netto pro Monat verdient, raucht.

Aufgabe 6.
In der Data Science hat man es oft mit Problemen zur Klassifikation zu tun. In dieser Aufgabe betrachten wir ein einleitendes Beispiel zur Spracherkennung. Wir wollen ein Modell finden, welches automatisch feststellt, ob ein Anrufer weiblich oder männlich ist. Dazu haben wir den Datensatz `voice.csv` (Quelle: kaggle.com), der unter anderem folgende Angaben zu den vorhandenen Frequenzen in 3168 Sprachbeispielen enthält:

- meanfreq: mean frequency (in kHz)
- median: median frequency (in kHz)
- maxfun: maximum fundamental frequency measured across acoustic signal

(a) Bestimmen Sie, welche der drei obigen Variablen am besten zur Klassifikation weiblich/ männlich geeignet ist.

(b) Bestimmen Sie die Klassifikationsmatrix und den Prozentsatz der richtigen Klassifikationen.