Data Mining / Intelligente Datenanalyse WS 2017/2018
PD Dr.-Ing. habil. Christian Borgelt

3. Übungsblatt

Aufgabe 14 Konfidenzintervalle

Unter den 87827 Lebendgeburten im Jahre 1972 in Niedersachsen waren 45195 Jungen. Bestimmen Sie aus diesen Angaben für die unbekannte Wahrscheinlichkeit p, daß ein neugeborenes Kind ein Junge ist, einen Punktschätzwert sowie Konfidenzintervalle zu den Konfidenzniveaus

a) $\alpha = 0.01$ (99%-Konfidenzintervall) und
b) $\alpha = 0.001$ (99.9%-Konfidenzintervall).

(Hinweis: Die benötigten Quantile der Normalverteilung können z.B. mit dem auf der Vorlesungsseite verfügbaren C-Program `ndqtl.c` berechnet werden. Quantil: Argumentwert zu vorgegebenem Funktionswert einer Verteilungsfunktion; analog zu den Quantilen einer Stichprobe.)

Aufgabe 15 Konfidenzintervalle

Bestimmen Sie ausgehend von dem bereits in Aufgabe 13 betrachteten Punktschätzer $W_2 = n \min_{i=1}^{n} X_i$ für den Parameter θ einer Exponentialverteilung ein Konfidenzintervall für diesen Parameter! Zur Erinnerung: In Aufgabe 13 haben wir

$$f_{W_2}(w; \theta) = \frac{1}{\theta} e^{-\frac{w}{\theta}}$$

als Wahrscheinlichkeitsdichte des Schätzers W_2 hergeleitet.

Aufgabe 16 Methode der kleinsten Quadrate/Regression

Bestimmen Sie eine Ausgleichsgerade $y = a + bx$ (Regressionsgerade) für den Datensatz aus Aufgabe 10, d.h. für

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

a) aus der Kovarianz und den Varianzen/Standardabweichungen (siehe Folien zur Vorlesung, Abschnitt zum Korrelationskoeffizienten)

b) mit Hilfe der Methode der kleinsten Quadrate/des Systems der Normalgleichungen!

Zeichnen Sie die Datenpunkte und die Ausgleichsgerade!

Aufgabe 17 Methode der kleinsten Quadrate/Regression

Bestimmen Sie mit Hilfe der Methode der kleinsten Quadrate eine Ausgleichsparabel $y = a + bx + cx^2$ für den Datensatz $(x, y) = ((0, 0), (2, 1), (3, 2), (4, 4))$ und zeichnen Sie diese Parabel!

Aufgabe 18 Multilineare Regression

Bestimmen Sie mit Hilfe der Methode der kleinsten Quadrate für den folgenden Datensatz eine Ausgleichsebene $z = a + bx + cy$:
$(x, y, z) = ((0, 1, 0), (0, 4, 2), (2, 0, 1), (3, 1, 2), (2, 3, 3), (4, 4, 4))$.

Aufgabe 19 Logistische Regression

Die folgende Tabelle zeigt die Anzahl der amerikanischen ballistischen Interkontinentalraketen (intercontinental ballistic missiles, ICBMs) in den sechziger Jahren:

Jahr, x	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Anzahl, y	18	63	294	424	834	854	904	1054	1054	1054

Finden Sie eine Ausgleichskurve mit Hilfe logistischer Regression ($Y = 1060$)! Zeichnen Sie die Originaldaten und skizzieren Sie die Kurve $y = \frac{1060}{1+e^{a+bx}}$!

Aufgabe 20 Exponentielle Regression

Radioaktive Substanzen zerfallen nach dem Gesetz $N(t) = N_0 e^{-\lambda t}$, wobei t die Zeit, λ ein subtanzabhängiger Zerfallsparameter, N_0 die zu Beginn und $N(t)$ die zum Zeitpunkt t noch vorhandenen Teilchen der radioaktiven Substanz sind. Mit Hilfe eines Geiger-Müller-Zählers werden bei einer kleinen Probe eines radioaktiven Materials die folgenden Anzahlen n der bis zu den Zeitpunkten t ausgesandten α-Teilchen gemessen:

t (in s)	0	30	60	90	120	150	180	210	240
n	0	306	552	655	768	863	901	919	956

Jedes gezählte α-Teilchen zeigt den Zerfall eines Teilchens der radioaktiven Substanz an. Bestimmen Sie die Halbwertszeit der radioaktiven Substanz! Um welche Substanz könnte es sich handeln?

Vorgehen: Legen Sie durch die Datenpunkte eine Ausgleichskurve $n = n_0(1 - e^{a+bt})$! (Hinweis: Sie müssen dazu eine Transformation finden, durch die das Problem auf die Bestimmung einer Ausgleichsgeraden zurückgeführt wird; $n_0 = 1000$.) Obwohl man so einen Wert für a erhält, der von 0 verschieden ist, wird man $-b$ als Näherungswert für den Zerfallsparameter λ ansehen dürfen, aus dem man die Halbwertszeit leicht berechnen kann. Die Halbwertszeit ist die Zeit, nach der die Hälfte der ursprünglich vorhandenen Teilchen zerfallen sind.