3. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 15.11.2018 23:59

WS 2018/2019
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD1819]
Di. 10-12	CP-03-150	tobias.hoinka@udo.edu, felix.geyer@udo.edu und jan.soedingrekso@udo.edu
Di. 16-18	CP-03-150	simone.mender@udo.edu und alicia.fattorini@udo.edu
Mi. 10-12	CP-03-150	mirco.huennefeld@udo.edu und kevin3.schmidt@udo.edu

Aufgabe 10: *Zwei Populationen* **5 P.**

Gegeben seien zwei Populationen von jeweils 10 000 Punkten in einer Ebene. Die Population P_0 sei eine zweidimensionale, korrelierte Gaußverteilung mit:

$$\mu_x = 0, \quad \mu_y = 3, \quad \sigma_x = 3{,}5, \quad \sigma_y = 2{,}6 \quad \text{und Korrelation} \quad \rho = 0{,}9$$

Die zweite Verteilung P_1 ist gegeben durch eine Gaußverteilung in x mit

$$\mu_x = 6 \quad \text{und} \quad \sigma_x = 3{,}5,$$

und einer Gaußverteilung in y, deren Mittelwert linear von x abhängt:

$$\mathbf{E}[y|x] = \mu_{y|x} = a + bx \quad \text{mit} \quad a = -0.5,\, b = 0.6 \quad \text{und} \quad \mathbf{Var}[y|x] = \sigma^2_{y|x} = 1$$

a) Zeigen Sie mithilfe der Formel für die bedingte Wahrscheinlichkeit der 2D Normalverteilung[1],

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{\tilde{y}}{\sigma_y} - \rho\frac{\tilde{x}}{\sigma_x}\right]^2\right) \tag{1}$$

dass die zweite Population ebenfalls einer 2D Normalverteilung entspricht. Geben Sie an, wie die Parameter μ'_y, σ'_y und ρ der 2D Normalverteilung aus den Parametern a, b und σ_y der 1D Normalverteilung der Population P_1 bestimmt werden.

b) Stellen Sie die beiden Populationen zusammen in einem zweidimensionalen Scatter-Plot dar.

c) Berechnen Sie die Stichproben-Mittelwerte und -Varianzen von x und y sowie die Stichproben-Kovarianz und den -Korrelationskoeffizienten für die Einzelpopulationen und die Gesamtheit beider Populationen.

[1] Von Blatt 2, Aufgabe 7: Zweidimensionale Gaußverteilung

d) Erzeugen Sie eine weitere Population mit den Eigenschaften der Population P_0, diesmal jedoch nur mit 1000 Punkten. Erstellen Sie anschließend ein HDF5-File mit drei Keys und speichern Sie die drei erzeugten Populationen unter eindeutigen Bezeichnern ab. Nutzen sie dafür das Python Paket `pandas`, siehe Python Hands-On.

Aufgabe 11: *Fisher-Diskriminante: Per Hand* — 5 P.

Führen Sie eine lineare Diskriminazanalyse nach Fisher per Hand durch.
Population 0: (1;1) (2;1) (1,5;2) (2;2) (2;3) (3;3)
Population 1: (1,5;1) (2,5;1) (3,5;1) (2,5;2) (3,5;2) (4,5;2)

a) Berechnen Sie die Mittelwerte $\vec{\mu}$ und Streumatrizen S_i, sowie die kombinierte Streumatrix S_{ij}.

b) Wie lautet $\vec{\lambda}$?

c) Zeichnen Sie die Punkte der beiden Populationen in einen Graphen ein, zusammen mit der Projektionsgeraden $\vec{\lambda} = \lambda \cdot \vec{e}_{\vec{\lambda}}$.

d) Projezieren Sie die einzelnen Punkte auf diese Gerade.

e) Wählen Sie einen geeigneten Parameter λ_{cut} und berechnen Sie die dazugehörige Effizienz und Reinheit. Warum haben Sie diesen Parameter gewählt?

Aufgabe 12: *Fisher-Diskriminante: Implementierung* — 10 P.

Gegeben seien die Populationen `P_0_10000` und `P_1` aus der Aufgabe „Zwei Populationen“. Nutzen Sie das dort erstellt HDF5-File für diese Aufgabe. (Sie finden die Datei ebenfalls im Moodle.)

Hinweis: Es sei Ihnen erlaubt Pakete z.B. für lineare Algebra zu benutzen, jedoch nicht Pakete, die die Diskriminanzanalyse durchführen.

a) Berechnen Sie die Mittelwerte μ_{P0} und μ_{P1} der beiden Populationen.

b) Berechnen Sie die Kovarianzmatrizen V_{P0} und V_{P1} der beiden Populationen, sowie die kombinierte Kovarianzmatrix $V_{P0,P1}$.

c) Konstruieren Sie eine lineare Fisher-Diskriminante $\vec{\lambda} = \lambda \cdot \vec{e}_{\vec{\lambda}}$. Geben Sie diese Geradengleichung an.

d) Stellen Sie die Populationen als Projektion auf die Gerade aus **c)** in einem eindimensionalen Histogramm dar.

e) Betrachten Sie $P0$ als Signal und $P1$ als Untergrund. Berechnen Sie die Effizienz und die Reinheit des Signals als Funktion eines Schnittes λ_{cut} in λ und stellen Sie die Ergebnisse in einem Plot dar.

f) Bei welchem Wert von λ_{cut} wird nach der Trennung das Signal-zu-Untergrundverhältnis S/B maximal? Erstellen Sie auch hierzu einen Plot.

g) Bei welchem Wert von λ_{cut} wird nach der Trennung die Signifikanz $S/\sqrt{S+B}$ maximal? Erstellen Sie auch hierzu einen Plot.

h) Wiederholen Sie die Schritte **a)** bis **g)** für den Fall, dass $P0$ nun die Population `P_0_1000` bezeichnet. Was fällt Ihnen auf? Interpretieren Sie die Ergebnisse.