4. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 21.11.2017, 23:59 Uhr

WS 2017/2018
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD1718]
Di. 10-12	CP-03-150	philipp2.hoffmann@udo.edu und jan.soedingrekso@udo.edu
Di. 16-18	P1-02-110	felix.neubuerger@udo.edu und tobias.hoinka@udo.edu
Di. 16-18	CP-03-150	simone.mender@udo.edu und maximilian.meier@udo.edu

Aufgabe 12: *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:

$$\mu_{y|x} = a + bx \quad \text{mit} \quad a = -0.5, b = 0.6 \quad \text{und} \quad \sigma_{y|x} = 1$$

a) Zeigen Sie mithilfe der Formel für die bedingte Wahrscheinlichkeit der 2D Normalverteilung [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.

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.

[1] Von Blatt 1, Aufgabe 4: Zweidimensionale Gaußverteilung

Aufgabe 13: *Trennende Geraden* **5 P.**

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

$$g_1(x) = 0 \tag{1}$$
$$g_2(x) = -\frac{3}{4}x \tag{2}$$
$$g_3(x) = -\frac{5}{4}x \quad . \tag{3}$$

gegeben.

a) Stellen Sie die beiden Populationen zusammen in einem zweidimensionalen Scatterplot dar und zeichnen Sie die drei Projektionsgeraden ein. Im Folgenden werden diese beiden Populationen mit $P0$ und $P1$ bezeichnet.

b) Projizieren Sie die Punkte aus Population $P0$ und $P1$ jeweils auf die Geraden g_1, g_2 und g_3. Bestimmen und normieren Sie vorher den Projektionsvektor und wählen sie das Vorzeichen so, dass die projizierte Population $P0$ rechts (zu größeren Werten) von $P1$ liegt. Stellen Sie die projizierten Populationen $P0$ und $P1$ für jede Projektion in einem eigenen, eindimensionalen Histogramm dar.

c) 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 den projizierten Räumen und stellen Sie die Ergebnisse für jede Projektion in einem eigenen Plot dar.

Aufgabe 14: *Fisher-Diskriminante: Per Hand* **5 P.**

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

a) Berechnen Sie die Mittelwerte $\vec{\mu}$ und Streumatrizen S_i, sowie die Streuung innerhalb der Klassen und zwischen den Klassen (S_W und S_B).

b) Berechnen Sie die Eigenwerte und Eigenvektoren der Matrix $S_W^{-1} \cdot S_B$.

c) Wie lautet die Projektion $\vec{\lambda}$? Verifizieren Sie dies, indem Sie $S_W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$ berechnen.

d) Projizieren Sie die Punkte der Populationen auf den Projektionsvektor (denken Sie an die Normierung). Skizzieren Sie die Positionen der einzelnen Punkte auf der Projektionsgeraden ($\vec{\lambda} \cdot \vec{x}_i$) (eindimensional).

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

Aufgabe 15: *Metropolis-Hastings-Algorithmus* **5 P.**

a) Zeigen Sie analytisch, dass der Metropolis-Hastings-Algorithmus für eine gaußförmige Schrittvorschlags-PDF in den Metropolis-Algorithmus übergeht.

b) Implementieren Sie den Metropolis-Algorithmus um aus einer Gaußverteilung zu samplen. Nutzen Sie dafür die in der Datei `MCMC_template.py` vorgegebene Klassenstruktur und implementieren Sie die vorgegebenen Methoden anhand ihrer Docstrings.

c) Erzeugen Sie mit ihrer implementierten `sample`-Methode 10 000 Zufallszahlen aus einer Gaußverteilung mit $\mu = -3$ und $\sigma = 2$. Nutzen Sie `x_0 = 15` als Startwert und `step_size = 2` als Schrittweite. Vergleichen Sie die erzeugten Zufallszahlen mit der zugrundeliegenden Verteilung.

d) Erzeugen Sie einen sogenannten „Trace Plot“, indem Sie die erzeugten Zufallszahlen gegen die Iteration, in der sie erzeugt wurden, auftragen. Interpretieren Sie das Ergebnis.