5. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 28.11.2017, 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 16: *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.

5. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 28.11.2017, 23:59 Uhr

WS 2017/2018
Prof. W. Rhode

Aufgabe 17: *kMeans per Hand* — 5 P.

Population: (1;4) (1;5) (1;6) (3;3) (3;2) (4;1) (5;1) (6;2) (6;3) (8;4) (8;5) (8;6)

a) Führen Sie den kMeans-Algorithmus (euklidisches Abstandsmaß) per Hand durch, um die Punkte der Population in Cluster zu gruppieren. Verwenden Sie als Startwerte die zufällig gewählten Clusterzentren (3;4), (7;4) und (3;7). Berechnen Sie die Abstände nur, wenn die Zugehörigkeit zum Clusterzentrum nicht offensichtlich ist. Skizzieren Sie die neuen Clusterzentren sowie die Grenzen zwischen den Clustern in der vorgefertigten Abbildung 1.

b) Führen Sie 4 weitere Iterationen von kMeans durch. Fertigen Sie bei jeder Iteration wieder eine Skizze an.

c) Nach wie vielen Iterationen konvergiert der Algorithmus? Entspricht das Ergebnis Ihren Erwartungen?

Aufgabe 18: *Hauptkomponentenanalyse (PCA)* — 5 P.

a) Erzeugen Sie mit der Funktion `sklearn.datasets.make_blobs` einen Datensatz. Nutzen sie dabei folgende Einstellungen: n_samples=1000, centers=2, n_features=4, random_state=0. Stellen Sie nun zwei beliebige Dimensonen des Datensatzes in einem Scatterplot dar.

b) Beschreiben Sie kurz die Funktionsweise der Hauptkomponentenanalyse. Geben Sie in Worten und in der richtigen Reihenfolge die notwendigen Berechnungen zur Durchführung der Hauptkomponentenanalyse an.

c) Wenden Sie nun die Hauptkomponentenanalyse auf den in a) erzeugten Datensatz an. Nutzen Sie dazu das Paket `sklearn.decomposition.PCA`. Wie lauten die Eigenwerte der Kovarianzmatrix? Wie interpretieren Sie die Eigenwerte?

d) Histrogrammieren Sie nun x' in jeder Dimension und stellen sie x'_1 und x'_2 in einem Scatterplot dar.

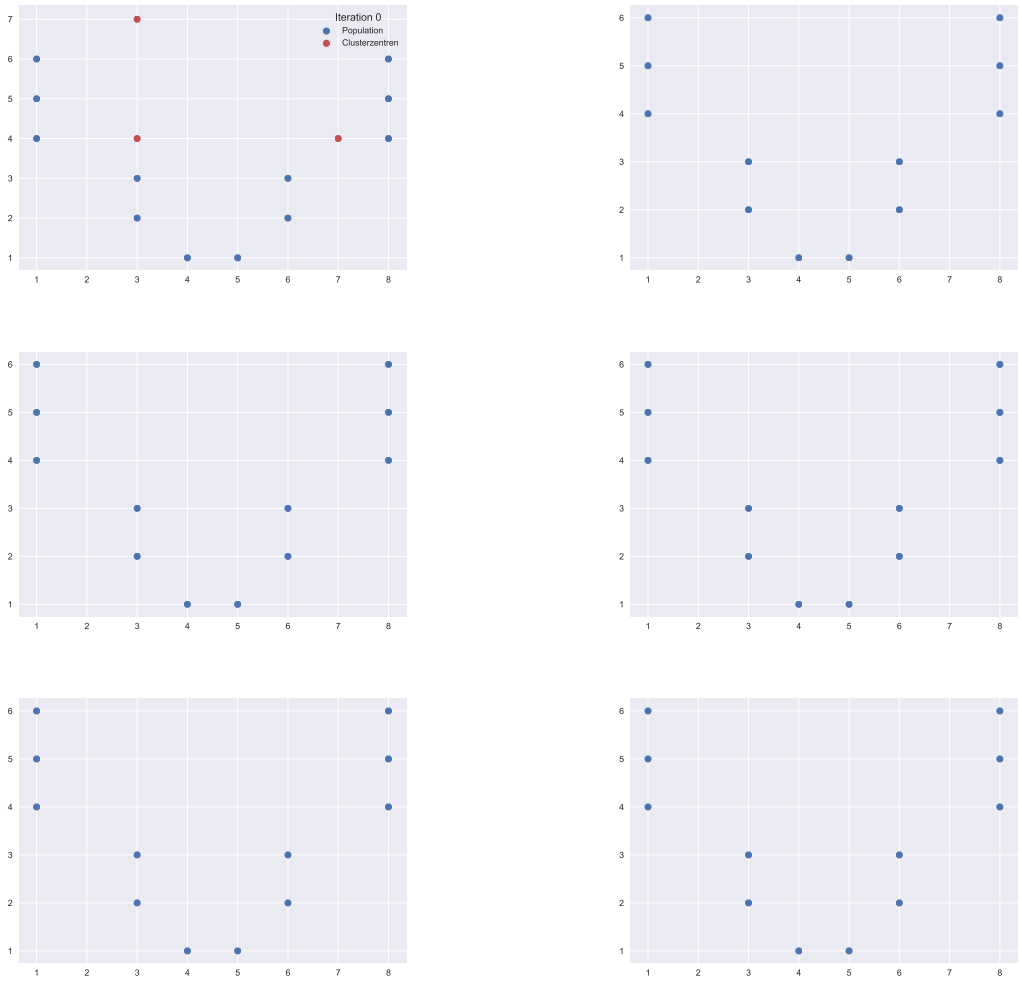

Abbildung 1: Population zum Einzeichnen der Clusterzentren und Clustergrenzen. Zu Aufgabe 17