

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD19]
Do.14–16	CP-03–150	kevin3.schmidt@udo.edu und maximilian.sackel@udo.edu
Fr. 10–12	CP-03–150	tobias.hoinka@udo.edu und noah.biederbeck@udo.edu
Fr. 16–18	CP-03–150	felix.geyer@udo.edu und rune.dominik@udo.edu

Aufgabe 11: *Fisher-Diskriminante: Per Hand*

5 P.

Führen Sie eine lineare Diskriminanzanalyse 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) Projizieren 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 erstellte 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_{P_0} und μ_{P_1} der beiden Populationen.
- (b) Berechnen Sie die Kovarianzmatrizen V_{P_0} und V_{P_1} der beiden Populationen, sowie die kombinierte Kovarianzmatrix V_{P_0, P_1} .
- (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 P_0 als Signal und P_1 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 P_0 nun die Population P_{0_1000} bezeichnet. Was fällt Ihnen auf? Interpretieren Sie die Ergebnisse.

Aufgabe 13: *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 Dimensionen 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) Histogrammieren Sie nun x' in jeder Dimension und stellen sie x'_1 und x'_2 in einem Scatterplot dar.