4. Übungszettel Mustererkennung WS15/16

Prof. Raúl Rojas, Fritz Ulbrich
Institut für Informatik, Freie Universität Berlin
Abgabe Online bis Mittwoch, 25.11.15, 10 Uhr

Bitte laden Sie ihre Lösung der Aufgaben als **pdf-Datei** hoch.
Quellcode können Sie optional als Archiv anhängen.

1. Aufgabe (6 Punkte):

Gegeben seien folgende Punkte:
(1,-1),(2,1),(4,-1),(5,1)

a. (1 Punkt) Bilden Sie den Mittelwert μ der gegebenen Punkte.
b. (1 Punkt) Berechnen Sie die Kovarianzmatrix C der gegebenen Punkte.
c. (1 Punkt) Berechnen Sie die Eigenwerte λ_1 und λ_2, indem Sie die Gleichung $det(C - I\lambda) = 0$ lösen. (I ist die Identitätsmatrix). Geben Sie die einzelnen Schritte der Berechnung an.
d. (1 Punkt) Geben Sie das durch das durch die Gleichung $(C - I\lambda) \cdot x = \vec{0}$ definierte Gleichungssystem an und berechnen Sie die Eigenvektoren x_1 und x_2. Sie müssen das Ergebnis eventuell noch normieren.
e. (1 Punkt) Geben Sie die Transformationsmatrix T an, um die Punkte in das durch die Eigenvektoren aufgespannte Koordinatensystem zu transformieren. Transformieren Sie die gegebenen Punkte.
f. (1 Punkt) Geben Sie die Matrix $M = T \cdot C \cdot T^T$ an. (T^T ist T transponiert)

2. Aufgabe (4 Punkte):

Laden Sie die Datei **mouse.csv** aus dem Resources Ordner der KVV-Seite herunter. Jede Zeile dieser Datei ist ein Datensatz, der einen zweidimensionalen Punkt beschreibt: **x, y, Klasse**.

a. (1 Punkt) Clustern Sie den Datensatz mit k-means in 3 Cluster. Verwenden Sie als initiale Zentren der Cluster die Punkte **(7,4),(8,6) und (9,4)** . Verwenden Sie als Abstandsmaß die **euklidische Distanz**. Plotten Sie die Clusterzentren und zugeordnete Datenpunkte für die ersten 12 Iterationen (12 Subplots in einem Plot)
b. (1 Punkt) Clustern Sie den Datensatz mit k-means in 3 Cluster. Verwenden Sie als Abstandsmaß die **Wahrscheinlichkeitsdichtefunktion (pdf) der Normalverteilung** der jeweiligen Cluster. Wählen Sie die initialen Mittelwerte der Cluster **zufällig** und als initiale Kovarianzmatrizen die **Identitätsmatrix**. Plotten Sie die Clusterzentren (Mittelwerte) und zugeordnete Datenpunkte für die ersten 12 Iterationen (12 Subplots in einem Plot)
c. (1 Punkt) Wie b) aber mit **(7,4),(8,6) und (9,4)** als initiale Clusterzentren.
d. (1 Punkt) Wie b) aber mit k=30. Die Clusterzentren (Mittelwerte) und Kovarianzmatrizen sollen dabei **gewichtetet** über **allen** Datenpunkten berechnet werden (also nicht nur über den Punkten, die den jeweiligen Clustern zugeordnet wurden). Als Gewichte verwenden Sie dabei den Wert der jeweiligen pdf. (https://en.wikipedia.org/wiki/Weighted_arithmetic_mean). Für diese Aufgabe müssen Sie die Datenpunkte nicht einem bestimmten Cluster zuordnen.