Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Arthur Zimek
Dr. Tobias Emrich

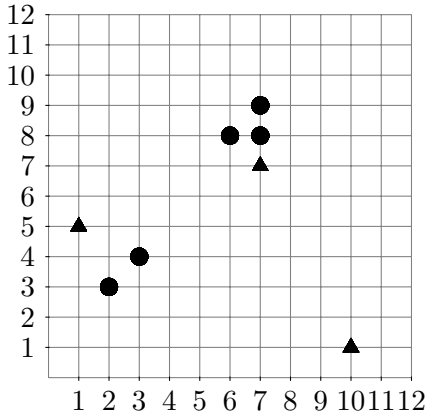
Knowledge Discovery in Databases

SS 2015

Übungsblatt 4: Clusteranalyse – k-means, PAM und EM

Aufgabe 4-1 Clustering durch Varianzminimierung

Gegeben sei folgender Datensatz mit 8 Punkten (Featurevektoren in $\mathbb{R}^2$).



Im folgenden sollen vollständige Partitionierungen des Datensatzes in $k = 2$ Cluster berechnet werden. Dabei wird jedes Objekt x demjenigen Cluster zugewiesen, bei dem die Summe der Quadrate der Abweichungen vom Clusterzentrum c minimal ist:

$$SSQ(x, c) = \sum_{i=1}^{d} |x_i - c_i|^2$$

(a) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem einfachen Verfahren "Clustering durch Varianzminimierung" (nach Lloyd, siehe Skript). Die initiale Partitionierung der Daten ist durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. Zeichnen Sie nach jedem Schritt die Zentroiden ein und markieren Sie die Punkte anhand ihrer Clusterzugehörigkeit.
Tipp: Hierzu können Sie die Vorlage auf der letzten Seite benutzen.

(b) Erzeugen Sie eine Partitionierung in $k = 2$ Cluster mit dem k-means Verfahren (nach MacQueen, siehe Skript). Die initiale Partitionierung der Daten ist auch hier durch die Dreiecke und Punkte gegeben (die Dreiecke bilden einen initialen Cluster, genauso die Punkte). Beschreiben Sie jede Aktion des Algorithmus. Zeichnen Sie nach jedem Schritt die Zentroiden ein und markieren Sie die Punkte anhand ihrer Clusterzugehörigkeit. Die Reihenfolge der Zuodnung bleibt Ihnen überlassen.
Tipp: Auch hierzu können Sie die Vorlage auf der letzten Seite benutzen.

(c) Begründen Sie kurz, warum k-means reihenfolgeabhängig ist.

(d) Optional: probieren Sie auch k-medoids, bspw. mit Euklidischer Distanz

Aufgabe 4-2 PAM

Zeigen Sie, dass der Algorithmus PAM konvergiert.

Aufgabe 4-3 Multivariate Dichte und Mahalanobis-Distanz

Die Dichte der multivariaten Normalverteilung (mit Kovarianzmatrix Σ und Mittelwert μ) wird mit der folgenden Formel berechnet:

$$prob(p, \mu, \Sigma) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot e^{-\frac{1}{2}\left((p-\mu)^T \Sigma^{-1} (p-\mu)\right)}$$

Finden und diskutieren Sie den Zusammenhang dieser Formel zu der Formel der Mahalanobis-Distanz mit Matrix Σ von p zu μ.

$$d_{\textit{Mahalanobis}}(x, y, \Sigma) := \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$$

Hinweis: betrachten Sie auch die multivariate Standardnormalverteilung, mit der Dichtefunktion

$$prob(p) := \frac{1}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}\langle p,p \rangle} = \frac{1}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}\|p\|^2}$$

Aufgabe 4-4 Zuweisung im EM-Algorithmus

Gegeben sei eine Datenmenge mit 100 Punkten, die drei Gausscluster A, B und C und den Punkt p enthält.

Der Cluster A enthält 30% aller Punkte und ist repräsentiert durch den Mittelwert aller seiner Punkte $\mu_A = (2, 2)$ und die Kovarianzmatrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

Der Cluster B enthält 20% aller Punkte und ist repräsentiert durch den Mittelwert aller seiner Punkte $\mu_B = (5, 3)$ und die Kovarianzmatrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

Der Cluster C enthält 50% aller Punkte und ist repräsentiert durch den Mittelwert aller seiner Punkte $\mu_C = (1, 4)$ und die Kovarianzmatrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$.

Der Punkt p ist durch die Koordinaten $(2.5, 3.0)$ gegeben. Geben Sie die beiden Wahrscheinlichkeiten an, mit der p zu den Clustern A, B bzw. C gehört.

Achtung: Folgende Skizze ist nur zu Veranschaulichungszwecken gedacht und nicht maßstabsgetreu!