

Knowledge Discovery in Databases
WS 2007/08
Übungsblatt 7

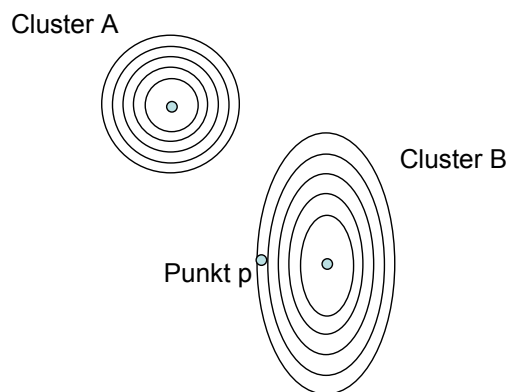
Aufgabe 7-1 EM-Algorithmus
Hausaufgabe

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

Der Cluster A ist repräsentiert durch den Mittelwert aller seiner Punkte $(1,1)$ und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$.
60 Prozent aller Punkte gehören zu diesem Cluster.

Der Cluster B ist repräsentiert durch den Mittelwert aller seiner Punkte $(3,3)$ und die Kovarianzmatrix $\begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}$.
40 Prozent aller Punkte gehören zu diesem Cluster.

Der Punkt p ist durch die Koordinaten $(2.5,2.5)$ gegeben. Geben Sie die beiden Wahrscheinlichkeiten an, mit der p zum Cluster A bzw. B gehört.



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

Lösung

Geg: Cluster A: $W_A = 0.6$, $\mu_A = (1, 1)$, $\Sigma_A = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$

Cluster B: $W_B = 0.4$, $\mu_B = (3, 3)$, $\Sigma_B = \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}$

Punkt: $p = (2.5, 2.5)$

Gesucht: $P(A|p)$ und $P(B|p)$

$P(p) = W_A \cdot P(p|A) + W_B \cdot P(p|B)$

$$\begin{aligned}
P(p|A) &= \frac{1}{\sqrt{(2\pi)^2 |\Sigma_A|}} \cdot e^{-\frac{1}{2}((p-\mu_A)^T \Sigma_A^{-1} (p-\mu_A))} \\
&= \frac{1}{\sqrt{(2\pi)^2 \cdot 16}} \cdot e^{-\frac{1}{2} \left((1.5, 1.5)^T \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix} \right)} \\
&= 0.0379 \cdot 0.5698 = 0.02267
\end{aligned}$$

$$\begin{aligned}
P(p|B) &= \frac{1}{\sqrt{(2\pi)^2 |\Sigma_B|}} \cdot e^{-\frac{1}{2}((p-\mu_B)^T \Sigma_B^{-1} (p-\mu_B))} \\
&= \frac{1}{\sqrt{(2\pi)^2 \cdot 64}} \cdot e^{-\frac{1}{2} \left((-0.5, -0.5)^T \begin{pmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix} \right)} \\
&= 0.01989 \cdot 0.96169 = 0.0191
\end{aligned}$$

$$W_A \cdot P(p|A) = 0.6 \cdot 0.02267 = 0.013602$$

$$W_B \cdot P(p|B) = 0.4 \cdot 0.0191 = 0.00764$$

$$P(p) = 0.021242$$

$$P(A|p) = \frac{0.013602}{0.021241} = 0.64$$

$$P(B|p) = \frac{0.00764}{0.021241} = 0.36$$

Aufgabe 7-2 DBSCAN

Das dichte-basiertes Clustermodell von DBSCAN definiert Kernpunkt und Randpunkte eines Clusters. Randpunkte sind Punkte, die zu einem Cluster gehören, weil sie dichte-erreichbar von Kernpunkten sind, aber selbst keine Kernpunkte sind. Wie der Name schon sagt, sind das Punkte, die am Rand eines Clusters liegen.

Kann es Randpunkte geben, die eigentlich zu verschiedenen Clustern gehören und wie würde DBSCAN mit solchen Randpunkten umgehen, d.h. welchem Cluster würden diese Punkte zugeordnet werden? Nennen Sie eine vernünftige Lösung zur Zuordnung dieser Randpunkte. Begründen Sie Ihre Entscheidung.