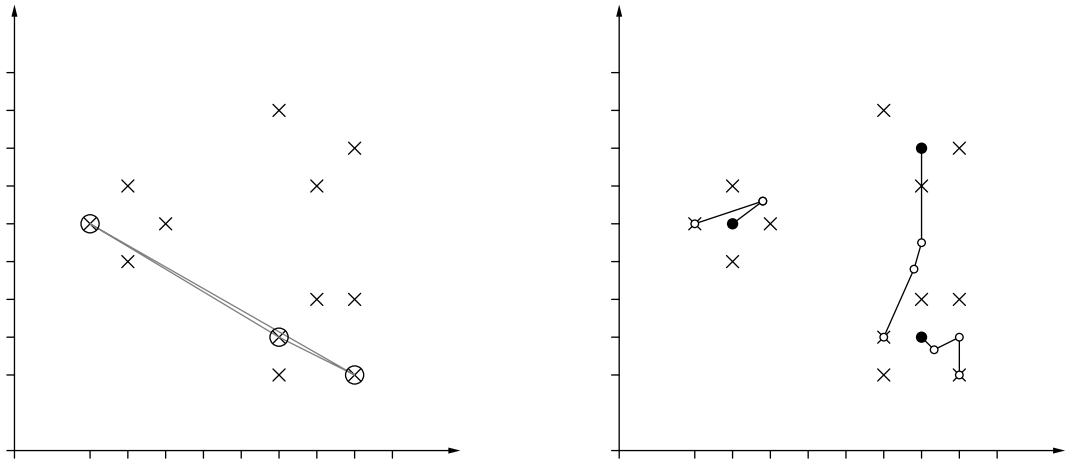
Data Mining / Intelligente Datenanalyse WS 2017/2018
PD Dr.-Ing. habil. Christian Borgelt

Lösung des 6. Übungsblattes

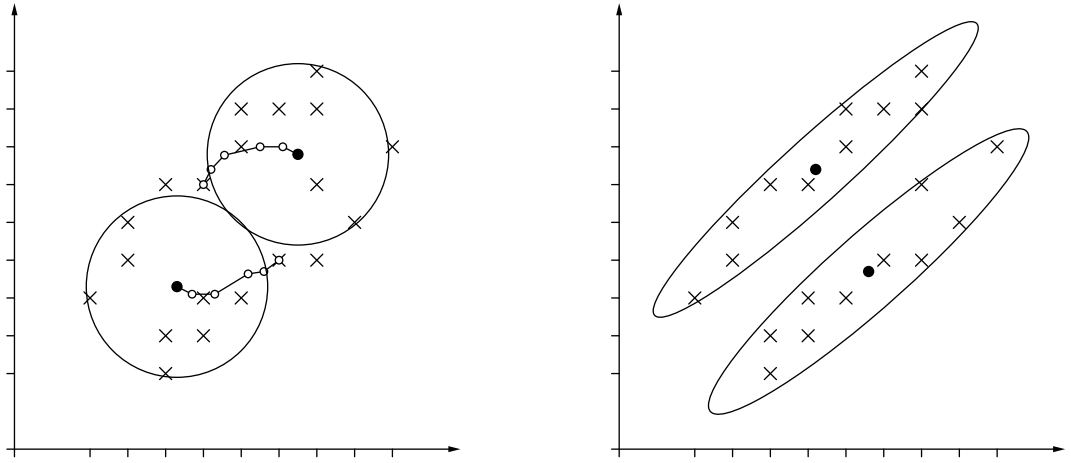
Aufgabe 29 k-Means-Clustering

Die beiden unten gezeigten Diagramme zeigen die Anfangssituation mit einer Delaunay-Triangulation der Cluster-Zentren (links) sowie die Bahnen, auf denen sich die Cluster-Zentren während der Aktualisierung bewegen. Offenbar wird das gewünschte Ergebnis erreicht, das Verfahren bleibt also nicht in einem lokalen Optimum hängen.



Aufgabe 30 k-Means-Clustering

Das linke der beiden folgenden Diagramm zeigt den Ablauf des k-Means-Clustering für Startpositionen, die sehr nah an den Mittelpunkten der beiden langgestreckten Punktwolken liegen. Obwohl die Initialisierung sehr günstig ist, verlassen die Clusterzentren jedoch ihre Lage und nehmen schließlich Positionen auf der Längsachse des gesamten Datensatzes ein. Der Grund für dieses Verhalten liegt in der Isotropie des euklidischen Abstandes, durch die das k-Means-Verfahren stets versucht, kreisförmige (allgemein: hyperkugelförmige) Cluster zu finden.



Um die beiden langgestreckten Punktwolken zu trennen, ist ein Verfahren nötig, mit dem die Form eines Clusters angepaßt werden kann, so daß ellipsoide Cluster erkannt werden können. Ein solches Verfahren ist das Gustafsson-Kessel-Fuzzy-Clustering. Es arbeitet mit dem Mahalanobis-Abstand, der eine (veränderbare) Kovarianzmatrix als Parameter erhält. Das Ergebnis dieses Verfahrens zeigt das rechte Diagramm.

Aufgabe 31 Lagrange-Theorie

Formal entspricht dieses Problem der Minimierung der Funktion

$$f(r,h) = 2\pi rh + 2\pi r^2 \qquad \text{(Oberfläche eines Zylinders)}$$

mit $r = \frac{d}{2}$ unter der Bedingung

$$\pi r^2 h = V \qquad \text{(Volumen eines Zylinders).}$$

Folglich lautet die Lagrange-Funktion

$$L(r,h,\lambda) = 2\pi rh + 2\pi r^2 + \lambda(\pi r^2 h - V).$$

Notwendige Bedingungen für ein Minimum sind

$$\begin{array}{rcccl} \frac{\partial L}{\partial r} & = & 2\pi h + 4\pi r + \lambda 2\pi rh & = & 2\pi(h + 2r + \lambda rh) \stackrel{!}{=} 0, \\ \frac{\partial L}{\partial h} & = & 2\pi r + 2\pi r^2 & = & \pi r(2 + \lambda r) \stackrel{!}{=} 0, \\ \frac{\partial L}{\partial \lambda} & = & \pi r^2 h - V & & \stackrel{!}{=} 0. \end{array}$$

Aus der zweiten Gleichung folgt $\lambda = -\frac{2}{r}$, was eingesetzt in die erste Gleichung $h = 2r$ ergibt. Folglich ergibt sich mit der dritten Gleichung

$$2\pi r^3 = V, \qquad \text{also} \qquad r = \sqrt[3]{\frac{V}{2\pi}} \qquad \text{und damit} \qquad d = h = 2\sqrt[3]{\frac{V}{2\pi}}.$$

Aufgabe 32 Fuzzy Clustering

Dem Hinweis der Aufgabenstellung folgend, betrachten wir einen Datenpunkt $\vec{x}_j$, der die Abstände d_{1j} und d_{2j} zu den beiden Clustern haben möge. Sei o.B.d.A. $d_{1j} < d_{2j}$ (anderenfalls vertausche man die Clusterindizes). Wir betrachten nun die Zugehörigkeitsgrade u_{1j} und u_{2j} und nehmen $u_{2j} > 0$ an. Dann können wir den Wert der Zielfunktion J durch die neue Zuweisung $u'_{2j} = 0$ und $u'_{1j} = u_{1j} + u_{2j}$ verringern, denn dies ändert die Summe der quadratischen Abstände (die Zielfunktion J) um

$$\begin{array}{rcl} \Delta & = & \underbrace{\left(u'_{1j}d_{1j}^2 + u'_{2j}d_{2j}^2\right)}_{\text{neue Terme}} - \underbrace{\left(u_{1j}d_{1j}^2 + u_{2j}d_{2j}^2\right)}_{\text{alte Terme}} \\ & = & (u_{1j} + u_{2j})d_{1j}^2 + 0 \cdot d_{2j}^2 - u_{1j}d_{1j}^2 - u_{2j}d_{2j}^2 \\ & = & u_{2j}(d_{1j}^2 - d_{2j}^2) \quad < \quad 0, \end{array}$$

da $0 \leq d_{1j} < d_{2j}$ nach Voraussetzung. Daher ist es am besten, $u_{1j} = 1$ und $u_{2j} = 0$ zu setzen. Durch direkte Verallgemeinerung dieser Überlegung sieht man, daß es allgemein am besten ist, jeden Datenpunkt mit vollem Gewicht dem Cluster zuzuordnen, der ihm am nächsten liegt. Also gilt für das Minimum der Funktion J: $\forall i \in \{1,\ldots,c\} : \forall j \in \{1,\ldots,n\} : u_{ij} \in \{0,1\}$.

Alternativ kann man so vorgehen: Sei $k_j = \operatorname{argmin}_{i=1}^{c} d_{ij}^2$, sei also k_j der Index des Clusters, dem der Datenpunkt $\vec{x}_j$ am nächsten liegt. Dann ist

$$\begin{aligned} J(\mathbf{X}, \mathbf{U}, \mathbf{C}) &= \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d_{ij}^2 \\ &\geq \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d_{k_j j}^2 \quad = \quad \sum_{j=1}^{n} d_{k_j j}^2 \underbrace{\sum_{i=1}^{c} u_{ij}}_{=1 \text{ (wegen der Nebenbedingung)}} \\ &= \sum_{j=1}^{n} \left(1 \cdot d_{k_j j}^2 + \sum_{\substack{i=1 \\ i \neq k_j}}^{c} 0 \cdot d_{ij}^2 \right). \end{aligned}$$

Folglich ist es am besten $u_{k_j j} = 1$ und $u_{ij} = 0$, $1 \leq i \leq c$, $i \neq k_j$, zu setzen. Mit anderen Worten: Die Zielfunktion wird minimiert, indem man jeden Datenpunkt dem nächstgelegenen Cluster zuordnet.

Aufgabe 33 Fuzzy Clustering

Die Berechnungsformel für die Zugehörigkeitsgrade lautet (siehe Vorlesungsfolien)

$$\forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : \qquad u_{ij} = \frac{d_{ij}^{\frac{2}{1-w}}}{\sum_{k=1}^{c} d_{kj}^{\frac{2}{1-w}}},$$

also für $w = 2$:

$$\forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : \qquad u_{ij} = \frac{d_{ij}^{-2}}{\sum_{k=1}^{c} d_{kj}^{-2}}.$$

Damit ergeben sich für Clusterzentren bei 1 und 5 unter Verwendung des euklidischen Abstandes die folgenden Zugehörigkeitsgrade:

j	x_j	u_{1j}	u_{2j}
1	1	1.000	0.000
2	3	0.500	0.500
3	4	0.100	0.900
4	5	0.000	1.000
5	8	0.155	0.845
6	10	0.236	0.764
7	11	0.265	0.735
8	12	0.288	0.712

Die Berechnungsformel für die neuen Clusterzentren lautet (siehe Vorlesungsfolien)

$$\forall i; 1 \leq i \leq c : \quad \vec{\mu}_i = \frac{\sum_{j=1}^{n} u_{ij}^w\, \vec{x}_j}{\sum_{j=1}^{n} u_{ij}^w}, \quad \text{also für } w = 2: \quad \forall i; 1 \leq i \leq c : \quad \vec{\mu}_i = \frac{\sum_{j=1}^{n} u_{ij}^2\, \vec{x}_j}{\sum_{j=1}^{n} u_{ij}^2}.$$

Damit ergeben sich als neue Zentren nach dem ersten Schritt $\mu_1 = 2.8848$ und $\mu_2 = 7.3930$. Die Konvergenzpunkte liegen bei $\mu_1 = 1.4009$ und $\mu_2 = 8.1440$.

Aufgabe 34 Expectation Maximization

Die Formel für die A-posteriori-Wahrscheinlichkeiten lautet (siehe Vorlesungsfolien)

$$\begin{aligned} & \forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : \\ p_{Y_j|\vec{X}_j}(i|\vec{x}_j; \mathbf{C}_k) &= \frac{f_{\vec{X}_j, Y_j}(\vec{x}_j, i; \mathbf{C}_k)}{f_{\vec{X}_j}(\vec{x}_j; \mathbf{C}_k)} = \frac{f_{\vec{X}_j|Y_j}(\vec{x}_j|i; \mathbf{C}_k) \cdot p_{Y_j}(i; \mathbf{C}_k)}{\sum_{l=1}^{c} f_{\vec{X}_j|Y_j}(\vec{x}_j|l; \mathbf{C}_k) \cdot p_{Y_j}(l; \mathbf{C}_k)}, \end{aligned}$$

wobei nach Aufgabenstellung $p_{Y_j}(i; \mathbf{C}_k) = \frac{1}{2}$ (feste A-priori-Wahrscheinlichkeit $\frac{1}{2}$, $i = 1, 2$). Außerdem ist

$$f_{\vec{X}_j|Y_j}(\vec{x}_j|i; \mathbf{C}_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\vec{x}_j - \vec{\mu}_i)^\top (\vec{x}_j - \vec{\mu}_i)\right)$$

(Mischung von Normalverteilungen mit fester Varianz $\sigma_i^2 = 1$, $i = 1, 2$). Damit ergeben sich für Clusterzentren bei 1 und 5 die folgenden A-posteriori-Wahrscheinlichkeiten:

j	x_j	u_{1j}	u_{2j}
1	1	1.000	0.000
2	3	0.500	0.500
3	4	0.018	0.982
4	5	0.000	1.000
5	8	0.000	1.000
6	10	0.000	1.000
7	11	0.000	1.000
8	12	0.000	1.000

Diese Wahrscheinlichkeiten werden als Fallgewichte eines vervollständigten Datensatzes verwendet, aus dem dann mit einfacher Maximum-Likelihood-Schätzung die neuen Clusterzentren ermittelt werden (mit den Fallgewichten berechneter Mittelwert der Datenpunkte). Damit ergeben sich als neue Zentren nach dem ersten Schritt $\mu_1 = 1.4009$ und $\mu_2 = 8.1440$. Die Konvergenzpunkte liegen bei $\mu_1 = 3.25$ und $\mu_2 = 10.25$.

Aufgabe 35 Agglomeratives Clustering

Im Unterschied zum Beispiel der Vorlesung führen bei diesem Datensatz alle Verfahren zu einem strukturell gleichen Dendrogramm, also zu der gleichen Clusterhierarchie. Man beachte jedoch, daß sich die Höhen der Brücken, die zu vereinigende Cluster verbinden, unterscheiden, da sich die Hypermetriken (zur Abstandsberechnung zwischen Clustern) unterscheiden.