Data Mining / Intelligente Datenanalyse WS 2017/2018

PD Dr.-Ing. habil. Christian Borgelt

6. Übungsblatt

Aufgabe 29 k-Means-Clustering

Gegeben sei der folgende zweidimensionale Datensatz:

x	1	6	8	3	2	2	6	6	7	7	8	8
y	5	2	1	5	4	6	1	8	3	6	3	7

Bearbeiten Sie diesen Datensatz mit k-Means-Clustering mit $k = 3$ (d.h., bilden Sie 3 Cluster)! Verwenden Sie die ersten drei Datentupel als Anfangszentren der Cluster und verfolgen Sie die Wanderung der Zentren.

Aufgabe 30 k-Means-Clustering

Betrachten Sie den folgenden einfachen Datensatz:

x	3	3	4	4	5	6	7	7	8	9	1	2	2	3	4	5	5	6	7	7
y	1	2	2	3	3	4	4	6	5	7	3	4	5	6	6	7	8	8	8	9

a) Welches Problem zeigt sich bei der Bearbeitung dieses Datensatzes mit dem k-Means-Algorithmus mit $k = 2$ (d.h., 2 Cluster)?
   Hinweis: Wie sieht das gewünschte Ergebnis aus? Was liefert stattdessen der k-Means-Algorithmus? (Sie brauchen das exakte Ergebnis des Algorithmus nicht auszurechnen, eine qualitative Beschreibung reicht.)

b) Wie kann man dieses Problem ggf. beheben?
   Hinweis: Erinnern Sie sich daran, wie sich ein voller von einem naiven Bayes-Klassifikator unterscheidet.

Aufgabe 31 Lagrange-Theorie

Wenn man Konservendosen (idealisiert: Zylinder) mit einem bestimmten Volumen V (z.B. 1 Liter) unter Verwendung von möglichst wenig Blech herstellen möchte, welche Abmessungen (Durchmesser d und Höhe h) muß dann eine Konservendose haben? Lösen Sie das Problem mit Hilfe der Methode der Lagrange-Multiplikatoren!

Aufgabe 32 Fuzzy Clustering

Betrachten Sie die Zielfunktion des Fuzzy Clustering für einen Fuzzifier $w = 1$, also

$$J(\mathbf{X}, \mathbf{B}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\, d^2(\beta_i, \vec{x}_j),$$

die unter der Nebenbedingung

$$\forall j \in \{1, \dots, n\} : \quad \sum_{i=1}^{c} u_{ij} = 1$$

zu minimieren ist. Zeigen Sie, daß sich auch dann, wenn man Zugehörigkeitsgrade u_{ij} aus dem Intervall $[0, 1]$ zuläßt, eine scharfe Partitionierung der Datenpunkte ergibt, also für das Minimum der Funktion J gilt: $\forall i \in \{1, \ldots, c\} : \forall j \in \{1, \ldots, n\} : u_{ij} \in \{0, 1\}$. (Hinweis: Vielleicht finden Sie es einfacher, zunächst den Fall $c = 2$ (zwei Cluster) zu betrachten und die Summanden für einen Datenpunkt $\vec{x}_j$ zu untersuchen. Verallgemeinern Sie dann das Ergebnis.)

Aufgabe 33 Fuzzy Clustering

Betrachten Sie den eindimensionalen Datensatz

$$1, 3, 4, 5, 8, 10, 11, 12.$$

Dieser Datensatz soll mit Fuzzy-c-Means-Clustering mit $c = 2$ (zwei Cluster) und einem Fuzzifier von $w = 2$ bearbeitet werden. Die Clusterzentren seien zu 1 und 5 initialisiert. Führen Sie einen Schritt der alternierenden Optimierung des Fuzzy Clustering aus, d.h.:

a) Berechnen Sie die Zugehörigkeitsgrade der Datenpunkte für die initialen Clusterzentren!

b) Berechnen Sie aus den so bestimmten Zugehörigkeitsgraden neue Clusterzentren!

Aufgabe 34 Expectation Maximization

Betrachten Sie erneut den eindimensionalen Datensatz aus Aufgabe 33, der in dieser Aufgabe mit dem Expectation-Maximization-Algorithmus zur Schätzung der Parameter einer Mischung von 2 Normalverteilungen bearbeitet werden soll. Die A-priori-Wahrscheinlichkeiten der beiden Cluster seien auf $\theta_i = \frac{1}{2}$, die Varianz der beiden Cluster auf $\sigma_i^2 = 1$, $i = 1, 2$, fixiert. (Es sollen also nur die Erwartungswerte der Normalverteilungen (Clusterzentren) angepaßt werden.) Benutzen Sie als Startpunkte für die Cluster die gleichen Werte wie in Aufgabe 40, also 1 und 5. Berechnen Sie einen Expectation- und einen Maximization-Schritt, d.h.:

a) Berechnen Sie die A-posteriori-Wahrscheinlichkeiten der Datenpunkte für die initialen Clusterzentren!

b) Schätzen Sie aus den so bestimmten Fallgewichten neue Clusterzentren!

Aufgabe 35 Agglomeratives Clustering

Gegeben sei der folgende eindimensionale Datensatz:

$$2, 5, 11, 12, 17, 21, 32.$$

Bearbeiten Sie diesen Datensatz mit agglomerativem Clustering unter Verwendung

a) der Zentroidmethode,

b) der Single-Linkage-Methode,

c) der Complete-Linkage-Methode!

Zeichnen Sie jeweils ein Dendrogramm!