Online-Test 5: Clustering und Outlier Detection

Test und Assessment – Druckansicht

Online-Test 5: Clustering und Outlier Detection

Datum: Tue Nov 2 07:59:45 2021 Maximale Punktezahl: 10

Frage 1 - Arten von Clustering-Algorithmen (1 Punkt) [ID: 932138]

Ordnen Sie die Clustering-Verfahren ihrer jeweiligen Kategorie zu!

Partitionierend	passt zu	CLARANS	(0.25 Punkte)
Hierarchisch	passt zu	DIANA	(0.25 Punkte)
Dichte-basiert	passt zu	OPTICS	(0.25 Punkte)
Probabilistisch	passt zu	Mixture Models	(0.25 Punkte)

Siehe Vorlesung.

Frage 2 - Auswahl von Clustering-Verfahren (1 Punkt) [ID: 932322]

Angenommen, Sie haben einen zweidimensionalen Datensatz, in welchem die Cluster die Form der Buchstaben "K", "I" und "T" annehmen. Gehen Sie davon aus, dass die Buchstaben direkt aufeinander folgen (wie in der Zeichenkette "KIT"), mit etwas Abstand zwischen den Buchstaben und keinen Ausreißern. Welche Clustering-Verfahren sind hier prinzipiell geeignet? (Das Finden von geeigneten Hyper-Parametern für die Verfahren ignorieren wir hier.)

- ☐ k-means *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*
- ☐ hierarchisch mit complete linkage *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*
- ☑ hierarchisch mit single linkage *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☑ DBSCAN *(Ausgewählt = 0.2 Punkte, Nicht ausgewählt = 0 Punkte)*
- ☐ Gaussian Mixture Models *(Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.2 Punkte)*

k-means ist auf sphärische Cluster ausgelegt. Gaussian Mixture Models sind ein wenig flexibler und lassen sich auch an elliptische Cluster anpassen, was in unserem Fall aber auch nicht ausreicht. complete linkage führt ebenfalls eher zu kompakten Clustern. Für das Finden von Clustern beliebiger Form sind DBSCAN und hierarchisch mit single linkage gut geeignet, vorausgesetzt, die sonstigen Rahmenbedingungen stimmen (z.B. ähnlich dichte Cluster für DBSCAN und keine Kette von Ausreißern zwischen den Clustern für single linkage).

Frage 3 - k-means (1 Punkt) [ID: 932143]

Bringen Sie die Bestandteile des k-means-Algorithmus in die richtige Reihenfolge!

- Wähle Cluster-Zentren
- [BEGINN] Schleife
- Ordne Punkte den Clustern zu
- Berechne Cluster-Zentren
- [ENDE] Schleife

Siehe Vorlesung. Für die Wahl der initialen Cluster-Zentren gibt es verschiedene Strategien. Im einfachsten Fall werden sie zufällig gewählt.

Frage 4 - Jaccard-Koeffzient (1 Punkt) [ID: 932140]

Gegeben seien die Mengen {D, A, B} und {G, B, E, F, D}. Berechnen Sie den Jaccard-Koeffizienten!

Runden Sie Ihre Lösung auf zwei Nachkommastellen falls notwendig. Verwenden sie einen Punkt als

Dezimaltrennzeichen.

Der Wert muss zwischen 0.33 und 0.33 liegen

Die Formel für den Jaccard-Koeffizienten zweier Mengen A und B lautet $\frac{|A \cap B|}{|A \cup B|}$. Im gegebenen Beispiel gibt es sechs unterschiedliche Elemente, von denen zwei in beiden Mengen enthalten sind.

Frage 5 - DBSCAN (1 Punkt) [ID: 932145]

Bringen Sie die Bestandteile des DBSCAN-Algorithmus in die richtige Reihenfolge!

- [BEGINN] Schleife: Für jeden unverarbeiteten Punkt x
- Markiere x als verarbeitet
- N := Nachbarn(x, epsilon)
- Wenn |N| >= minPts:
- C := neues Cluster
- Füge x zu C hinzu
- Erweitere C rekursiv
- Sonst:
- Markiere x als "Noise"
- [ENDE] Schleife

Siehe Vorlesung.

Frage 6 - Statistische Ausreißererkennung (1 Punkt) [ID: 932320]

Welcher Anteil der Daten hat in einer Normalverteilung eine Distanz von mindestens zwei Standardabweichungen zum Mittelwert? Sie können dies mit R berechnen. Verwenden Sie beim Angeben der Lösung einen Punkt als Dezimaltrennzeichen und runden Sie die Lösung auf drei Nachkommastellen.

Der Wert muss zwischen 0.046 und 0.046 liegen

Die Lösung lässt sich in R unter Nutzung der Verteilungsfunktion der Normalverteilung berechnen: *round(2 * (1 - pnorm(2, mean = 0, sd = 1)), digits = 3)*

Wir berechnen hierfür die Verteilungsfunktion an der Stelle x = 2, invertieren die Wahrscheinlichkeit (da uns der Bereich außerhalb interessiert) und multiplizieren mit 2 (da die Verteilung symmetrisch ist).

Frage 7 - Sparsity (1 Punkt) [ID: 932318]

Gegeben sei ein Datensatz mit 2 Milliarden Datenobjekten und 8 Attributen. Zwecks Ausreißererkennung wollen Sie jedes Attribut gleichmäßig in 10 Partitionen unterteilen. Wie viele Datenobjekte kann man in einer Zelle des Raumes nach der Unterteilung erwarten?

Der Wert muss zwischen 20 und 20 liegen

$2 * 10^9 * (\frac{1}{10})^8 = 20$

Frage 8 - Subspace Search (1 Punkt) [ID: 932314]

Gegeben sei ein Datensatz mit 50 Attributen. Wie viele Teilräume mit 5 Dimensionen gibt es?

Der Wert muss zwischen 2118760 und 2118760 liegen

Wir haben ein einfaches kombinatorisches Problem vorliegen, bei dem 5 aus 50 Elementen ausgewählt werden sollen. Die Reihenfolge der Elemente spielt keine Rolle. Dementsprechend lautet die Lösung $\frac{50!}{(50-5)!*5!}$.

Frage 9 - Strong und Weak Outlier (1 Punkt) [ID: 932306]

Stellen sie sich vor, ein Lehrstuhl wertet eine Klausur mit vier Aufgaben aus und erstellt eine Statistik. Die

meisten Punktzahlen befinden sich im mittleren Bereich. Es sollen Ausreißer auf dem gesamten (vierdimensionalen) Datenraum gefunden werden.

Szenario 1: Jemand erreicht mit Abstand die höchste Gesamtpunktzahl. In den einzelnen Aufgaben gibt es aber jeweils auch eine gewisse Menge an Studierenden mit ähnlicher Punktzahl, und zum Teil auch einzelne Studierende mit sehr hoher oder niedriger Punktzahl.

Szenario 2: Jemand erreicht in allen vier Aufgaben volle Punktzahl. Das hat sonst niemand geschafft, auch nicht in einzelnen Aufgaben.

Bewerten Sie folgende Aussagen!

- ☐ In Szenario 1 liegt ein strong outlier vor.
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☑ In Szenario 1 liegt ein weak outlier vor.
  (Ausgewählt = 0.25 Punkte, Nicht ausgewählt = 0 Punkte)
- ☐ In Szenario 2 liegt ein strong outlier vor.
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)
- ☐ In Szenario 2 liegt ein weak outlier vor.
  (Ausgewählt = 0 Punkte, Nicht ausgewählt = 0.25 Punkte)

Szenario 1: Da die Punktzahl nur insgesamt auffällig ist, nicht aber in den Teilräumen, liegt ein nicht-trivialer Ausreißer vor. Er ist aber nicht strong, da es laut Beschreibung weitere Ausreißer in den Teilräumen geben kann.

Szenario 2: Der Ausreißer ist trivial, da er bereits in den Teilräumen auftritt. Somit ist er weder strong noch weak.

Frage 10 - Projected Clustering (1 Punkt) [ID: 932328]

Bewerten Sie folgende Aussagen zu Projected Clustering!

Für jede Aussage muss entschieden werden: [richtig] oder [falsch]

richtig	falsch	
○	◉	Jedes Cluster bezieht sich auf dieselben Attribute.
○	◉	Jedes Cluster bezieht sich auf dieselbe Anzahl von Attributen.
○	◉	Jedes Cluster enthält dieselben Datenobjekte.
○	◉	Jedes Cluster enthält dieselbe Anzahl von Datenobjekten.

Die Datenobjekte sollten sich natürlich zwischen den Clustern unterscheiden (wozu clustert man sonst?) und auch die Anzahl der Datenobjekte pro Cluster ist nicht festgelegt. Weiterhin wird für jedes Cluster eine Teilmenge an Attributen ausgewählt, die sich von Cluster zu Cluster unterscheiden kann.