Cluster

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Ein Cluster ist eine Teilmenge der Datenpunkte eines Datensatz, die einander ähnlicher sind als den Datenpunkte außerhalb der Teilmenge. Das measure der Ähnlichkeit ist eine Entwurfsentscheidung, wie die euklidischer Abstand zwischen Merkmalsvektoren oder die Kantengewichte eines similarity graph. Bei einem hartes Clustering gehört jeder Datenpunkt zu genau einem Cluster; bei einem soft clustering gehört er zu jedem Cluster mit einer Zugehörigkeitsgrad, und die Zugehörigkeitsgrade eines Datenpunkt summieren sich zu eins. Eine Clustering-Methode repräsentiert ein Cluster durch ein Objekt, das die Zugehörigkeit jedes Datenpunkt bestimmt. Drei Repräsentationen sind in Gebrauch: ein Clusterzentrum, wie bei k-Means; eine Wahrscheinlichkeitsverteilung, wie die multivariate Normalverteilung einer Komponente eines Gaußsches Mischmodell (GMM); oder eine dicht verbundene Teilmenge der Knoten eines similarity graph, wie beim graph clustering. Die Anzahl und die Form der Cluster hängen von der Methode und ihren Hyperparameter ab: k-Means liefert konvex Cluster, ein GMM Ellipsoide, und density-based spatial clustering of applications with noise (DBSCAN) jede Form, die eine Kette benachbarter Datenpunkte nachzeichnen kann.

Definition

Man betrachte die 365 Tage eines Jahres in Krems an der Donau (Österreich), jeder ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$, der die Lufttemperatur um 06:00 und die Temperatur um 15:00 (UTC) des Tages enthält. Ein Tag im Januar, mit einem 06:00-Messwert nahe 0 und einem Maximum nahe 5 Grad, ähnelt den anderen Januartagen mehr als jedem Tag im Juli, dessen Temperaturen etwa 20 Grad höher liegen. Ein Cluster ist eine Teilmenge der Datenpunkte eines Datensatz, die einander ähnlicher sind als den Datenpunkte außerhalb der Teilmenge (Bishop, 2006, Sect. 9.1; Hastie et al., 2009, Sect. 14.3). Abb. 1(a) zeigt die zwei Cluster, die k-Means unter den Tagen des Jahres 2024 findet: 197 kalte Tage und 168 warme Tage. Jeder Tag im Januar und Februar ist ein kalter Tag, und jeder Tag im Juli und August ist ein warmer Tag. Gemessen an der euklidischer Abstand zwischen ihren Merkmalsvektoren, liegen zwei Tage desselben Clusters im Durchschnitt 8.2 Grad auseinander, ein kalter

und ein warmer Tag 21.3 Grad, und für 362 der 365 Tage gehört der ähnlichste andere Tag zum selben Cluster.

Das measure der Ähnlichkeit zwischen Datenpunkte ist eine Entwurfsentscheidung (Hastie et al., 2009, Sect. 14.3). Für Datenpunkte mit numerischen Merkmalsvektoren $\mathbf{x} \in \mathbb{R}^d$ sind zwei Datenpunkte ähnlich, wenn die euklidischer Abstand zwischen ihren Merkmalsvektoren klein ist. Datenpunkte ohne numerische Merkmalsvektoren, wie die Mitglieder eines sozialen Netzwerks oder die Artikel eines Zitationsnetzwerks, sind die Knoten eines similarity graph, dessen Kantengewichte ihre Ähnlichkeit direkt messen.

Formal betrachte man ein Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$. Ein Cluster ist eine Teilmenge $\mathcal{C}^{(c)} \subseteq \{1, \dots, m\}$ der Indizes der Datenpunkte. Ein hartes Clustering in k Cluster ist eine Partition der Indizes,

$$\mathcal{C}^{(1)} \cup \cdots \cup \mathcal{C}^{(k)} = \{1, \dots, m\} \quad \text{und} \quad \mathcal{C}^{(c)} \cap \mathcal{C}^{(c')} = \emptyset \text{ für } c \neq c', \tag{1}$$

sodass jeder Datenpunkt zu genau einem Cluster gehört, und das hartes Clustering durch den Index dieses Clusters in $\{1, \dots, k\}$ aufgezeichnet wird (Jain et al., 1999, Sect. 2). Die zwei Cluster von Abb. 1(a) bilden eine Partition (1) der 365 Tage. Ein soft clustering ersetzt den Index durch eine Zugehörigkeitsgrad $u_c^{(r)}$ des Datenpunkt r zu jedem Cluster $c = 1, \dots, k$,

$$u_c^{(r)} \geq 0 \quad \text{und} \quad \sum_{c=1}^{k} u_c^{(r)} = 1 \quad \text{für } r = 1, \dots, m, \tag{2}$$

sodass die Zugehörigkeitsgrade eines Datenpunkt einen Vektor in der probability simplex Δ^{k-1} bilden. Ein Cluster ist dann keine Teilmenge mehr, sondern die Abbildung $r \mapsto u_c^{(r)}$, die die Zugehörigkeit jedes Datenpunkt bewertet. Ein hartes Clustering ist der Spezialfall, in dem jede Zugehörigkeitsgrad 0 oder 1 ist: das Cluster $\mathcal{C}^{(c)}$ besteht dann aus den Datenpunkte mit $u_c^{(r)} = 1$, und diese Cluster bilden eine Partition (1).

Eine Clustering-Methode speichert ein Cluster nicht als Liste von Indizes. Sie repräsentiert das Cluster durch ein Objekt, aus dem die Zugehörigkeit jedes Datenpunkt folgt, einschließlich eines Datenpunkt, der nicht zum Datensatz gehört. Dieses Objekt ist die Hypothese, die die Methode lernt, und die Generalisierung auf neue Datenpunkte ist ihr Zweck. Drei Repräsentationen sind in Gebrauch (Abb. 1).

Erstens ist ein Clusterzentrum $\mathbf{w}^{(c)} \in \mathbb{R}^d$ ein einzelner Punkt des Merkmalsraum (Bishop, 2006, Sect. 9.1). Es repräsentiert ein Cluster als die Menge der Datenpunkte, die ihm näher sind als jedem anderen Clusterzentrum,

$$\mathcal{C}^{(c)} = \left\{r : \left\|\mathbf{x}^{(r)} - \mathbf{w}^{(c)}\right\|_2 \leq \left\|\mathbf{x}^{(r)} - \mathbf{w}^{(c')}\right\|_2 \text{ für alle } c' \neq c\right\}. \tag{3}$$

Dies ist das hartes Clustering, das von k-Means geliefert wird, dessen Clusterzentren den clustering error minimieren, eine Instanz von empirische Risikominimierung (ERM). Der Bereich des Merkmalsraum, der die Ungleichungen im Cluster (3) erfüllt, ist ein Schnittpunkt von Halbräume, einer für jeden anderen

Clusterzentrum, und daher konvex. Für die Tage in Krems und $k = 2$ sind die beiden Clusterzentren $(3.9, 9.0)$ und $(17.5, 24.9)$, und die Grenze zwischen den beiden Clustern ist die gerade Linie durch ihren Mittelpunkt im rechten Winkel zu ihrem Verbindungssegment (Abb. 1(a)).

Zweitens repräsentiert eine Wahrscheinlichkeitsverteilung ein Cluster. Ein Gaußsches Mischmodell (GMM) repräsentiert Cluster c durch eine multivariate Normalverteilung mit Mittelwert $\boldsymbol{\mu}^{(c)}$ und Kovarianzmatrix $\mathbf{C}^{(c)}$, gezogen mit Wahrscheinlichkeit p_c. Die Zugehörigkeitsgrad eines Datenpunkt zu Cluster c ist die A-posteriori-Verteilung seines Cluster-Index gegeben seinen Merkmalsvektor, ausgewertet bei c (Bishop, 2006, Sect. 9.2),

$$u_c^{(r)} = \frac{p_c \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}. \tag{4}$$

Die Zugehörigkeitsgrade (4) erfüllen die Bedingungen (2), sodass das Ergebnis ein soft clustering ist. Die Niveaumengen der Wahrscheinlichkeitsdichte einer multivariate Normalverteilung sind Ellipsoide, deren Achsen die Eigenvektoren der Kovarianzmatrix sind (Bishop, 2006, Sect. 2.3), sodass ein auf diese Weise repräsentiertes Cluster ein Ellipsoid ist und kein Schnittpunkt von Halbräume. Für die Tage in Krems haben die vom EM-Algorithmus (EM) Algorithmus gelernten Kovarianzmatrizen eine lange Achse, die 3.2-fach bzw. 2.2-fach so lang ist wie die kurze, geneigt um 48 und 52 Grad, sodass jedes Cluster entlang der Richtung gestreckt ist, in der beide Temperaturen gemeinsam ansteigen (Abb. 1(b)). Die größere der beiden Zugehörigkeitsgrade benennt das k-Means-Cluster für 315 der 365 Tage, und 60 Tage haben eine Zugehörigkeitsgrad zwischen 0.2 und 0.8 zu beiden Clustern. Abb. 1(d) vergleicht die beiden Zuordnungen über das Jahr: der k-Means-Index springt zwischen 0 und 1, während die Zugehörigkeitsgrad zum warmen Cluster an jedem Tag von November bis Februar 0 ist und an Tagen von März bis Oktober Zwischenwerte durchläuft.

Drittens, für Datenpunkte, die die Knoten $\mathcal{V} = \{1, \ldots, m\}$ eines similarity graph mit Kantengewichte $A_{r,r'} \geq 0$ sind, ist ein Cluster eine Teilmenge $\mathcal{C} \subseteq \mathcal{V}$ von Knoten, die untereinander dicht verbunden und mit dem Rest spärlich verbunden sind. Methoden des graph clustering messen dies durch die Summe der Kantengewichte der Kanten, die die Teilmenge verlassen, geteilt durch ihre Größe,

$$\frac{1}{|\mathcal{C}|} \sum_{r \in \mathcal{C}} \sum_{r' \notin \mathcal{C}} A_{r,r'}, \tag{5}$$

und ein hartes Clustering in k Cluster ist gut, wenn die Summe der Verhältnisse (5) über seine Cluster, der Ratio Cut, klein ist (von Luxburg, 2007, Sect. 5, Eq. (1)). Spectral clustering minimiert den Ratio Cut annähernd. Für die Tage in Krems hat der k-nearest-neighbor graph (k-NN graph), der jeden Tag mit seinen 10 nächsten Nachbarn verbindet, mit Einheits-Kantengewichte, 2214 Kanten; 2170 davon verbinden zwei Tage desselben k-Means-Clusters und 44 verbinden einen kalten mit einem warmen Tag. Der Ratio Cut der beiden Cluster beträgt 0.49, gegenüber 12.50 für die Aufteilung der Tage in solche mit ungeradem und solche mit geradem Index (Abb. 1(c)).

Welche Teilmengen eines Datensatz Cluster sind, hängt daher von der Methode und ihren Hyperparameter ab, nicht allein vom Datensatz. Die Anzahl k der Cluster ist ein Input für k-Means und für ein GMM, während density-based spatial clustering of applications with noise (DBSCAN) sie aus einem Abstand-Schwellenwert und einer minimalen Clustergröße bestimmt. Die Form eines Clusters ergibt sich aus seiner Repräsentation: ein Schnittpunkt von Halbräume für k-Means, eine Ellipsoid für ein GMM, und jede Form, die eine Kette benachbarter Datenpunkte für DBSCAN und für graph clustering nachzeichnen kann. Dieselben 365 Tage werden in Abb. 1(a) durch eine gerade Linie geteilt, während in Abb. 1(b) jeder Tag eine Zugehörigkeitsgrad zu jeder von zwei geneigten Ellipsen erhält. Die Wahl der Repräsentation ist eine Wahl des Hypothesenraum, und die Anzahl der Cluster wird wie jeder andere Hyperparameter gewählt, da der clustering error allein sie nicht wählen kann.

Cluster dienen als Kategorien, wo kein Label sie bereitstellt, dem Setting des unüberwachtes Lernen. Ein Energieversorger kann die kalten Tage in Abb. 1(a) als Heizperiode behandeln und einen neuen Tag nach der Seite der Grenzlinie klassifizieren, auf die sein Merkmalsvektor fällt. Bei der image segmentation ist ein Cluster von Pixeln mit ähnlichen Farben eine Region des Bildes (Bishop, 2006, Sect. 9.1.1), und bei der Kundensegmentierung ist ein Cluster von Kunden mit ähnlichen Kaufhistorien ein Marktsegment (Wedel and Kamakura, 2000).

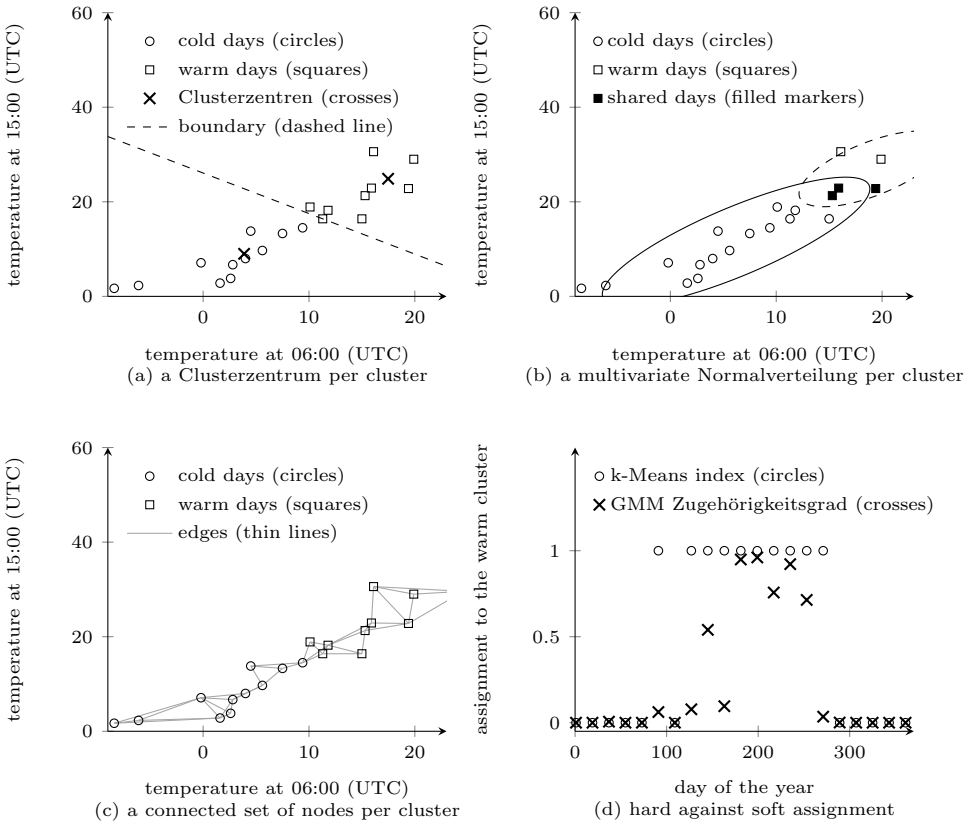


Abbildung 1: Jeder 18 der 365 Tage des Jahres 2024 in Krems, 21 Tage, unter den drei Darstellungen eines Clusters. (a) Die Tage, dargestellt als kalte Tage (Kreise) oder warme Tage (Quadrate) durch den näheren der beiden Clusterzentren von k-Means (Kreuze); die Grenze zwischen den beiden Clustern ist eine gerade Linie (gestrichelte Linie). (b) Die beiden multivariate Normalverteilungen eines GMM, der kalte Cluster (durchgezogene Ellipse) und der warme Cluster (gestrichelte Ellipse), jeweils dargestellt als Niveaumenge seiner Wahrscheinlichkeitsdichte beim doppelten der Standardabweichung entlang jeder Achse; ein Tag wird durch den Cluster mit der größeren Zugehörigkeitsgrad dargestellt, und ausgefüllt, wo diese Zugehörigkeitsgrad unter 0.8 liegt (ausgefüllte Marker). (c) Der 3-Nächste-Nachbarn-Graph der 21 dargestellten Tage, der jeden Tag mit den drei ihm ähnlichsten Tagen verbindet (dünne Linien); 38 der 41 Kanten verbinden zwei Tage desselben Clusters und drei kreuzen zwischen ihnen. (d) Dieselben 21 Tage nach ihrem Tag des Jahres: die k-Means-Zuordnung, 0 für einen kalten und 1 für einen warmen Tag (Kreise), und die Zugehörigkeitsgrad zum warmen Cluster unter dem GMM (Kreuze). Daten generiert von `pythondemos/cluster.py`

Siehe auch: Clustering, Clusterzentrum, clustering error, hartes Clustering,

soft clustering, Zugehörigkeitsgrad, k-Means, GMM, multivariate Normalverteilung, graph clustering, similarity graph, k-NN graph, Kantengewicht, spectral clustering, DBSCAN, Datenpunkt, Merkmalsvektor, euklidischer Abstand, Merkmalsraum, Hyperparameter, konvex, probability simplex, unüberwachtes Lernen.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Jain et al. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), pp. 264–323.

3. von Luxburg (2007). A tutorial on spectral clustering. Statist. Comput., 17(4), pp. 395–416.

4. Wedel and Kamakura (2000). Market Segmentation: Conceptual and Methodological Foundations. 2nd ed., Kluwer Academic Publishers, Boston, MA.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.