Zugehörigkeitsgrad

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 Zugehörigkeitsgrad ist eine Zahl, die das Ausmaß angibt, in dem ein Datenpunkt zu einem Cluster gehört, jeweils einer für jedes Paar aus einem Datenpunkt und einem Cluster einer Clustering. In der gebräuchlichsten Formulierung sind die Zugehörigkeitsgrade eines Datenpunkt nicht-negativ und summieren sich zu eins, sodass sie einen Vektor der probability simplex und eine probability mass function (pmf) über den Cluster-Indizes bilden. Die hartes Clustering ist der Spezialfall, in dem dieser Vektor einen Eintrag gleich eins und alle anderen gleich null hat, wie er von k-Means geliefert wird. Eine prinzipielle Konstruktion von Zugehörigkeitsgraden streng zwischen null und eins ist die A-posteriori-Verteilung des Cluster-Index unter einem Gaußsches Mischmodell (GMM), die der EM-Algorithmus (EM) Algorithmus in seinem E-Schritt berechnet. Die Zugehörigkeitsgrade bestimmen wiederum die Cluster: Der Mittelwert eines Cluster ist der Durchschnitt der Merkmalsvektoren, gewichtet mit ihnen, und mit Zugehörigkeitsgraden von null oder eins ist dieser gewichtete Durchschnitt der Mittelungsschritt des Lloyd's algorithm. Wenn die gemeinsame Varianz der Cluster eines GMM gegen null schrumpft, werden die Zugehörigkeitsgrade zu null oder eins, und die soft clustering wird zur hartes Clustering.

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 des Tages um 06:00 (UTC) und um 15:00 (UTC) enthält. Eine Clustering-Methode zerlegt diese Tage in ein kaltes Cluster und ein warmes Cluster. Ein Tag im Januar ist ein kalter Tag und ein Tag im Juli ein warmer Tag, aber der 22. Mai 2024, mit einer Messung von 16.0 Grad um 06:00 und 20.7 Grad um 15:00, liegt zwischen den beiden. Ein Zugehörigkeitsgrad $u_c^{(r)}$ ist eine Zahl, die das Ausmaß angibt, in dem Datenpunkt r zu Cluster c gehört, für $c = 1, \ldots, k$ und $r = 1, \ldots, m$. Für den 22. Mai 2024 liefert das Gaußsches Mischmodell (GMM) aus Abb. 2 die Zugehörigkeitsgrade 0.50 zum kalten und 0.50 zum warmen Cluster, während ein Tag im Januar 1.00 und 0.00 erhält. In der gebräuchlichsten Formulierung sind die Zugehörigkeitsgrade eines Datenpunkt nicht-negativ und

summieren sich zu eins (Bezdek, 1981; Ruspini, 1969),

$$u_c^{(r)} \geq 0 \quad \text{für } c = 1, \ldots, k \qquad \text{und} \qquad \sum_{c=1}^{k} u_c^{(r)} = 1. \tag{1}$$

Der Vektor $\mathbf{u}^{(r)} := \left(u_1^{(r)}, \ldots, u_k^{(r)}\right)^\top$ liegt dann in der probability simplex Δ^{k-1}. Es ist die probability mass function (pmf) einer Zufallsvariable (ZV) mit Werten in $\{1, \ldots, k\}$, dem Cluster-Index des Datenpunkt. Die possibilistische Clustering lässt die Summenbedingung der Simplex-Bedingung (1) fallen und lässt jeden Zugehörigkeitsgrad eigenständig in $[0, 1]$ variieren (Krishnapuram and Keller, 1993). Abb. 1 skizziert drei Datenpunkte und ihre Zugehörigkeitsgrade zu zwei überlappenden Cluster.

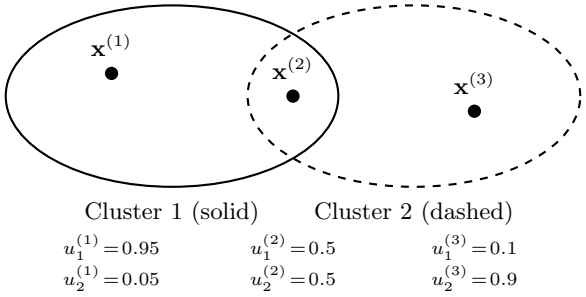


Abbildung 1: Drei Datenpunkte mit unterschiedlichen Zugehörigkeitsgraden zu zwei überlappenden Cluster, Cluster 1 (durchgezogene Ellipse) und Cluster 2 (gestrichelte Ellipse). Der Merkmalsvektor $\mathbf{x}^{(1)}$ liegt tief innerhalb von Cluster 1 und $\mathbf{x}^{(3)}$ tief innerhalb von Cluster 2, während $\mathbf{x}^{(2)}$ im Überlappungsbereich mit gleichen Zugehörigkeitsgraden liegt. In jedem Fall ist $u_1^{(r)} + u_2^{(r)} = 1$, die Simplex-Bedingung (1)

Die hartes Clustering ist der Spezialfall, in dem jeder Datenpunkt zu genau einem Cluster gehört. Mit $\widehat{c}^{(r)}$ als Index des Cluster, dem Datenpunkt r zugewiesen wird, sind seine Zugehörigkeitsgrade

$$u_c^{(r)} = \begin{cases} 1 & \text{für } c = \widehat{c}^{(r)}, \\ 0 & \text{sonst,} \end{cases} \tag{2}$$

sodass der Vektor $\mathbf{u}^{(r)}$ ein Eckpunkt der probability simplex Δ^{k-1} (Bishop, 2006, Sect. 9.1). K-Means liefert eine solche hartes Clustering: $\widehat{c}^{(r)}$ ist der Index des Clusterzentrum, der $\mathbf{x}^{(r)}$ in euklidischer Abstand am nächsten liegt. Für die Tage in Krems und zwei Clusterzentren weist k-Means 197 Tage dem kalten Cluster mit Clusterzentrum $(3.9, 9.0)$ und 168 Tage dem warmen Cluster mit Clusterzentrum $(17.5, 24.9)$ zu. Die gestrichelte Kurve in Abb. 2(b) zeigt diese harten Zuweisungen über das Jahr: Sie springen zwischen 0 und 1 von einem Frühlingstag zum nächsten.

Soft clustering-Methoden weisen einen Datenpunkt zu mehreren Cluster mit Zugehörigkeitsgraden streng zwischen 0 und 1 zu. Eine prinzipielle Konstruktion verwendet ein Wahrscheinlichkeitsmodell für die Datenpunkte, das GMM. Dabei ist der Cluster-Index $I^{(r)} \in \{1, \dots, k\}$ des Datenpunkt r eine unbeobachtete ZV mit $\mathbb{P}\left(I^{(r)} = c\right) = p_c$. Gegeben $I^{(r)} = c$, wird der Merkmalsvektor $\mathbf{x}^{(r)}$ aus einer multivariate Normalverteilung mit Mittelwert $\boldsymbol{\mu}^{(c)}$ und Kovarianzmatrix $\mathbf{C}^{(c)}$ gezogen. Der Zugehörigkeitsgrad ist dann die A-posteriori-Verteilung des Cluster-Index gegeben den beobachteten Merkmalsvektor,

$$u_c^{(r)} := \mathbb{P}\left(I^{(r)} = c \mid \mathbf{x}^{(r)}\right) = \frac{p_c \, \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \, \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}, \tag{3}$$

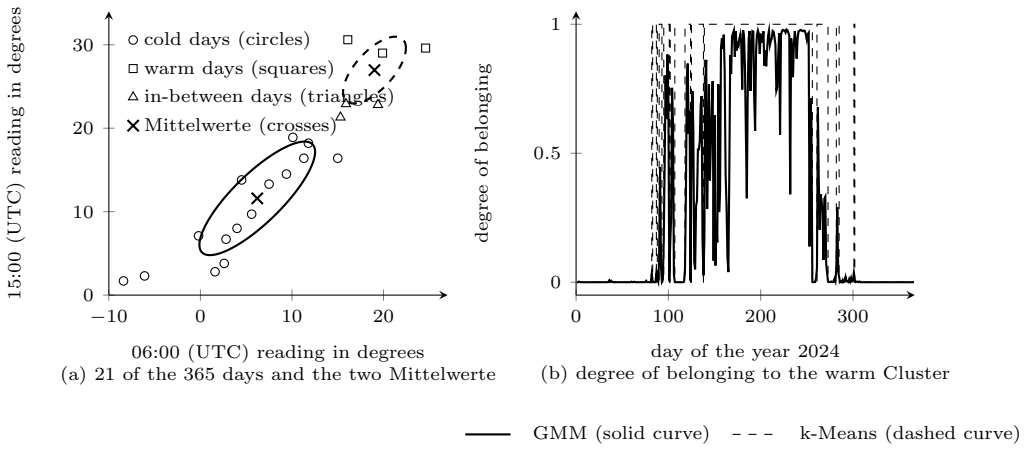
die Verantwortung von Cluster c für den Datenpunkt (Bishop, 2006, Sect. 9.2; Murphy, 2012, Sect. 11.2.3). Die posterioren Zugehörigkeitsgrade (3) sind die Wahrscheinlichkeiten der k Werte einer ZV, sodass sie die Simplex-Bedingung (1) per Konstruktion erfüllen. Die Modellparameter p_c, $\boldsymbol{\mu}^{(c)}$ und $\mathbf{C}^{(c)}$ des GMM werden mittels maximum likelihood gelernt. Dies ist ein Beispiel für empirische Risikominimierung (ERM): Der Verlust eines Datenpunkt ist der negative Logarithmus der Wahrscheinlichkeitsdichte des GMM an seinem Merkmalsvektor, und der clustering error ist der Durchschnitt dieses Verlust über den Datensatz. Für die Tage in Krems hat ein Zweikomponenten-GMM, das mit dem EM-Algorithmus (EM) Algorithmus gelernt wurde, die Mittelwerte $(6.2, 11.6)$ und $(19.0, 27.0)$ sowie die Cluster-Wahrscheinlichkeiten 0.69 und 0.31. Dabei haben 60 der 365 Tage keinen Zugehörigkeitsgrad über 0.8, davon 22 im Mai und keiner zwischen November und Februar (Abb. 2).

Die Zugehörigkeitsgrade bestimmen wiederum die Cluster. Im M-Schritt des EM Algorithmus ist der Mittelwert von Cluster c der Durchschnitt der Merkmalsvektoren, gewichtet mit den Zugehörigkeitsgraden. Die Cluster-Wahrscheinlichkeit ist der Durchschnitt der Zugehörigkeitsgrade selbst. In Formeln,

$$\boldsymbol{\mu}^{(c)} = \frac{\sum_{r=1}^{m} u_c^{(r)} \, \mathbf{x}^{(r)}}{\sum_{r=1}^{m} u_c^{(r)}} \qquad \text{und} \qquad p_c = \frac{1}{m} \sum_{r=1}^{m} u_c^{(r)} \quad \text{für } c = 1, \dots, k. \tag{4}$$

Die Kovarianzmatrix $\mathbf{C}^{(c)}$ ist die Kovarianzmatrix der Merkmalsvektoren unter derselben Gewichtung (Bishop, 2006, Sect. 9.2.2). Ein Tag mit den Zugehörigkeitsgraden 0.50 und 0.50 zählt zur Hälfte zu jedem Mittelwert. Mit den Zugehörigkeitsgraden (2) einer hartes Clustering ist der gewichtete Durchschnitt (4) der Mittelwert der Datenpunkte, die Cluster c zugewiesen sind, was der Mittelungsschritt des Lloyd's algorithm für k-Means ist, und p_c ist der Anteil der Datenpunkte in diesem Cluster. Der EM Algorithmus wechselt die posteriore (3) mit den gewichteten Durchschnitten (4) ab, sodass die Zugehörigkeitsgrade in jeder Iteration neu berechnet werden: Sie sind keine Modellparameter, sondern eine Funktion der Modellparameter und des Merkmalsvektor. Am Fixpunkt des EM Algorithmus für die Tage in Krems entspricht jeder Mittelwert dem Durchschnitt der Tage, gewichtet mit ihren

Zugehörigkeitsgraden, und unterscheidet sich vom k-Means Clusterzentrum desselben Cluster um höchstens 2.6 Grad.
Die hartes Clustering ist auch ein Grenzwert der posterioren Zugehörigkeitsgrade (3). Man fixiert jede Cluster Wahrscheinlichkeit bei $1/k$ und jede Kovarianzmatrix bei $\sigma^2\mathbf{I}$, ein Vielfaches der EinheitsMatrix $\mathbf{I}$ mit einer gemeinsamen Varianz σ^2. Die posteriore (3) hängt dann von $\mathbf{x}^{(r)}$ nur über ihre quadrierten euklidische Abstände zu den Mittelwerte ab. Wenn σ^2 gegen 0 schrumpft, konzentriert sie sich auf den nächsten Mittelwert: jeder Zugehörigkeitsgrad tendiert gegen 0 oder 1, und der EM Algorithmus reduziert sich auf den Lloyd's algorithm (Bishop, 2006, Sect. 9.3.2; Hastie et al., 2009, Sect. 14.3.7). Für die Tage in Krems, mit den beiden k-Means Clusterzentren als Mittelwerte und $\sigma^2 = 1$, hat kein Tag beide Zugehörigkeitsgrade über 0.2, und 364 der 365 Tage haben jeden Zugehörigkeitsgrad innerhalb von 0.05 ihres k-Means-Wertes (2). Die Ausnahme ist der 28. September 2024, dessen quadrierte euklidische Abstände zu den beiden Clusterzentren sich um 0.07 unterscheiden: ein Tag an der k-Means-Grenze behält den Grad 0.51, egal wie klein σ^2 ist, sodass der Grenzwert an jedem Datenpunkt außerhalb dieser Grenze gilt und nicht gleichmäßig. Mit $\sigma^2 = 50$ haben 13 Prozent der Tage beide Grade über 0.2. Mit $\sigma^2 = 1000$ hat jeder Tag beide über 0.2 und kein Zugehörigkeitsgrad überschreitet 0.65. Die Zugehörigkeitsgrade eines Datenpunkt sind der Wert einer Hypothese an seinem Merkmalsvektor. Die posteriore (3) ist eine Abbildung $h : \mathbb{R}^d \to \Delta^{k-1}$, die durch die Modellparameter des GMM bestimmt wird, und die Zuordnungsregel von k-Means ist eine Abbildung von $\mathbb{R}^d$ zu den Eckpunkten von Δ^{k-1}, die durch die Clusterzentren bestimmt wird. Jede Abbildung gilt für einen Tag, der nicht zum Datensatz gehörte, aus dem die Cluster gelernt wurden, was die Art und Weise ist, wie eine Clustering generalisiert. Eine hartes Clustering wird aus Zugehörigkeitsgraden abgeleitet, indem jeder Datenpunkt dem Cluster mit dem größten zugeordnet wird, und $1 - \max_{c=1,\ldots,k} u_c^{(r)}$ misst, wie unsicher diese Zuordnung ist (Murphy, 2012, Sect. 11.2.3). Für die Tage in Krems benennt der größte posteriore Zugehörigkeitsgrad das k-Means Cluster an 86.3 Prozent der Tage. An jedem der 50 Tage, an denen die beiden sich unterscheiden, alle zwischen März und Oktober, sagt k-Means warm und das GMM sagt kalt: die kalte Komponente hat die größere Kovarianzmatrix und die größere Cluster Wahrscheinlichkeit, daher beansprucht sie die milden Tage, die näher am warmen Clusterzentrum liegen. In der medizinischen image segmentation lassen die Zugehörigkeitsgrade eines Pixels zu den Gewebetypen ein Pixel an der Grenze zwischen zwei Geweben teilweise zu beiden gehören, während eine hartes Clustering eines auswählen muss (Pham et al., 2000).



(a) 21 of the 365 days and the two Mittelwerte (b) degree of belonging to the warm Cluster

Abbildung 2: (a) Jeder 18. der 365 Tage des Jahres 2024 in Krems, 21 Tage, dargestellt nach ihren posterioren Zugehörigkeitsgraden (3) unter dem Zweikomponenten-GMM: kalte Tage (Kreise) und warme Tage (Quadrate) haben einen Zugehörigkeitsgrad über 0.8 zu einem Cluster, Zwischentage (Dreiecke) haben beide unter 0.8. Die Kreuze sind die beiden Mittelwerte, und die Ellipsen sind die Niveaumengen der beiden multivariate Normalverteilungen bei einer Standardabweichung: die kalte (durchgezogene Ellipse) und die warme (gestrichelte Ellipse). (b) Der Zugehörigkeitsgrad zum warmen Cluster jedes Tages des Jahres: das Posteriori unter dem GMM (durchgezogene Kurve) und die 0 oder 1 von k-Means (gestrichelte Kurve). Daten generiert von `pythondemos/dob.py`

Synonyme: membership degree, responsibility. Siehe auch: Cluster, Clustering, soft clustering, hartes Clustering, Clusterzentrum, GMM, EM, k-Means, Lloyd's algorithm, A-posteriori-Verteilung, probability simplex, pmf, clustering error.

Querverweise

Literatur

1. Bezdek (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, NY, USA.

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

3. Krishnapuram and Keller (1993). A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1(2), pp. 98–110.

4. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

5. Pham et al. (2000). Current methods in medical image segmentation. Annual Review of Biomedical Engineering, 2, pp. 315–337.

6. Ruspini (1969). A new approach to clustering. Information and Control, 15(1), pp. 22–32.

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