hartes Clustering

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

Hartes Clustering ist die Aufgabe, einen Datensatz in einige wenige Cluster zu zerlegen, sodass jeder Datenpunkt zu genau einem Cluster gehört und die Cluster somit eine Partition des Datensatz bilden. Seine Ausgabe ist eine Clusterzuweisung, ein Index in $\{1, \ldots, k\}$ für jeden Datenpunkt, die eine Label-ähnliche Größe ist, die allein aus den Merkmalsvektoren berechnet wird und über die Partition hinaus keine Bedeutung hat. Als One-Hot-Vektor geschrieben, ist die Zuweisung ein Zugehörigkeitsgrad, der entweder 0 oder 1 ist, sodass hartes Clustering der Spezialfall von soft clustering ist, bei dem jeder Vektor von Zugehörigkeitsgrade ein Eckpunkt des probability simplex ist. K-Means liefert ein hartes Clustering, dessen Cluster konvexe Regionen um Clusterzentren sind, density-based spatial clustering of applications with noise (DBSCAN) eines, dessen Cluster Ketten von dicht umgebenen Datenpunkte sind, und spectral clustering eines, das aus Eigenvektoren der Laplacian matrix eines similarity graph berechnet wird. Eine Methode, die ein Cluster durch einen Clusterzentrum repräsentiert, lernt eine Hypothese, die jeden Merkmalsvektor einem Cluster zuweist, und die Minimierung ihres clustering error ist empirische Risikominimierung (ERM). Ein hartes Clustering wird intern durch den clustering error und extern gegen eine Referenzpartition bewertet, durch die Genauigkeit nach der besten Umnummerierung der Cluster oder durch den Rand-Index, den Anteil der Paare von Datenpunkte, die beide Partitionen zusammen oder getrennt platzieren.

Definition

Man betrachte die 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), wobei jeder Tag ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$ ist, der den Messwert um 06:00 Uhr und die Temperatur um 15:00 Uhr (UTC) des Tages enthält. Ein Heizversorger, der seine Brennstoffkäufe nach Wetterregime plant, benötigt jeden Tag in genau einem Regime. Hartes Clustering ist die Clustering-Aufgabe, die dies liefert: Es zerlegt den Datensatz in k Cluster, sodass jeder Datenpunkt zu genau einem von ihnen gehört (Hastie et al., 2009, Sect. 14.3.5; Jain et al., 1999, Sect. 2; Shalev-Shwartz and Ben-David, 2014, Ch. 22). Mit $k = 2$ teilt k-Means die Tage in 198 kalte Tage und 167

warme Tage auf (Abb. 1(a)); jeder Tag ist ein kalter Tag oder ein warmer Tag, niemals beides und niemals keines von beiden.

Formal besteht ein hartes Clustering eines Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ aus k Cluster $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(k)} \subseteq \mathcal{D}$, die disjunkt sind und zusammen den Datensatz abdecken,

$$\mathcal{C}^{(c)} \cap \mathcal{C}^{(c')} = \emptyset \text{ für } c \neq c' \quad \text{und} \quad \bigcup_{c=1}^{k} \mathcal{C}^{(c)} = \mathcal{D}, \tag{1}$$

d.h. eine Partition des Datensatz. Die Partition (1) ist dasselbe Objekt wie die Clusterzuweisung $\widehat{c}^{(r)} \in \{1, \dots, k\}$ jedes Datenpunkt $r = 1, \dots, m$, da $\mathcal{C}^{(c)} = \{\mathbf{x}^{(r)} : \widehat{c}^{(r)} = c\}$. Die Zuweisung ist eine Label-ähnliche Ausgabe. Wie das Label eines Datenpunkt in der Klassifikation nimmt es einen von endlich vielen Werten an. Im Gegensatz zu einem Label ist es nicht mit den Daten gegeben, sondern wird allein aus den Merkmalsvektoren berechnet, was hartes Clustering zu einer Form des unüberwachtes Lernen macht. Der Index c eines Cluster hat über die Partition hinaus keine Bedeutung: Eine Umnummerierung der Cluster ergibt dasselbe harte Clustering (Hastie et al., 2009, Sect. 14.3.5).

Die Zuweisung des Datenpunkt r kann auch als Vektor von Zugehörigkeitsgrade $u_c^{(r)}$ geschrieben werden, einer für jedes Cluster,

$$u_c^{(r)} := \begin{cases} 1 & \text{falls } \widehat{c}^{(r)} = c, \\ 0 & \text{sonst,} \end{cases} \qquad c = 1, \dots, k, \tag{2}$$

ein One-Hot-Vektor mit einem einzigen Eintrag 1 (Bishop, 2006, Sect. 9.1). Eine soft clustering-Methode weist jedem Datenpunkt einen Vektor $\left(u_1^{(r)}, \dots, u_k^{(r)}\right)^\top$ von nicht-negativen Zugehörigkeitsgrade zu, die sich zu eins summieren, d.h. einen Punkt des probability simplex Δ^{k-1}; ein Gaußsches Mischmodell (GMM) verwendet beispielsweise die A-posteriori-Verteilung des Cluster-Index. Hartes Clustering ist der Spezialfall, in dem jeder solche Vektor ein Eckpunkt des probability simplex ist, sodass jeder Zugehörigkeitsgrad (2) 0 oder 1 ist. Ein soft clustering wird in ein hartes Clustering umgewandelt, indem jeder Datenpunkt dem Cluster mit dem größten Zugehörigkeitsgrad zugewiesen wird (Jain et al., 1999, Sect. 5), und ein GMM, dessen Kovarianzmatrizen alle $\sigma^2\mathbf{I}$ sind, liefert, wenn $\sigma^2 \to 0$, das harte Clustering von k-Means (Bishop, 2006, Sect. 9.3.2).

Harte Clustering-Methoden unterscheiden sich darin, wie sie ein Cluster repräsentieren. K-Means repräsentiert Cluster c durch einen Clusterzentrum $\mathbf{w}^{(c)} \in \mathbb{R}^d$ und weist jeden Datenpunkt dem nächsten zu, $\widehat{c}^{(r)} = \arg\min_{c=1,\dots,k} \left\|\mathbf{x}^{(r)} - \mathbf{w}^{(c)}\right\|_2$ (Lloyd, 1982; MacQueen, 1967). Die den $\mathbf{w}^{(c)}$ zugewiesenen Merkmalsvektoren bilden eine konvexe Region des Merkmalsraum, die durch Hyperebenen begrenzt ist, eine für jeden anderen Clusterzentrum, nämlich die Mittelsenkrechte des Segments zwischen den beiden Clusterzentren. Für die beiden Clusterzentren in Krems, $(3.9, 9.0)$ und $(17.5, 24.9)$, ist diese Grenze die gestrichelte Linie in Abb. 1(a); der ihr nächste Tag, der 28. September, liegt 0.06 Grad davon entfernt. Density-based spatial clustering of applications with noise (DBSCAN) repräsentiert ein Cluster

als eine maximale Menge von Datenpunkte, die durch eine Kette von dicht umgebenen Datenpunkte verbunden sind, sodass ein Cluster jede beliebige Form haben kann und die Zahl k eine Ausgabe und kein Input ist (Ester et al., 1996). Datenpunkte, die zu keiner solchen Kette gehören, sind Rauschen, und die Cluster von DBSCAN partitionieren den Datensatz erst, nachdem die Rauschpunkte entfernt oder in einer eigenen Menge gesammelt wurden. Spectral clustering repräsentiert die Datenpunkte als die Knoten eines similarity graph, berechnet aus k Eigenvektoren seiner Laplacian matrix einen neuen Merkmalsvektor in $\mathbb{R}^k$ für jeden Knoten und führt k-Means auf diesen neuen Merkmalsvektoren aus (von Luxburg, 2007, Sect. 4). Die resultierenden Cluster sind konvexe Regionen im neuen Merkmalsraum, müssen aber nicht konvex im ursprünglichen sein.

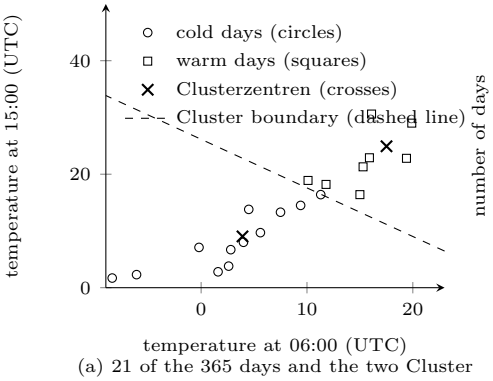
Eine harte Clustering-Methode, die ein Cluster durch einen Clusterzentrum repräsentiert, lernt eine Hypothese: die Abbildung $h(\mathbf{x}) = \mathbf{w}^{(c^\star(\mathbf{x}))}$, die jeden Merkmalsvektor zu seinem nächsten Clusterzentrum sendet, mit $c^\star(\mathbf{x}) = \arg\min_{c=1,\dots,k} \left\|\mathbf{x} - \mathbf{w}^{(c)}\right\|_2$. Die Partition (1) des Datensatz ist die Einschränkung dieser Abbildung auf den Datensatz, und ein Datenpunkt außerhalb des Datensatz, beispielsweise ein Tag des Jahres 2025, wird dem Cluster seines nächsten Clusterzentrum zugewiesen, ohne die Partition neu zu berechnen. Die Wahl der Clusterzentren, die den clustering error, den durchschnittlichen quadrierten euklidischer Abstand zwischen einem Datenpunkt und seinem Clusterzentrum, minimieren, ist empirische Risikominimierung (ERM) über den Hypothesenraum aller solcher Abbildungen. Der clustering error ist daher der Trainingsfehler der gelernten Hypothese, und er ist das interne Maß eines harten Clusterings, das allein aus dem Datensatz berechnet wird. Er sinkt, wenn k wächst, und verschwindet, wenn jeder Datenpunkt sein eigenes Cluster ist, sodass er k nicht wählen kann.

Ein externes Maß vergleicht das harte Clustering mit einer Referenzpartition desselben Datensatz, hier das kalte Halbjahr von Oktober bis März (182 Tage) und das warme Halbjahr von April bis September (183 Tage). Da die Cluster-Indizes keine Bedeutung haben, darf der Vergleich nicht von ihrer Nummerierung abhängen. Eine Möglichkeit besteht darin, die Cluster durch die Bijektion auf die Referenzindizes umzubenennen, die den Anteil der Datenpunkte maximiert, bei denen die beiden Partitionen übereinstimmen. Dieser Anteil ist die Genauigkeit des umnummerierten harten Clusterings, eins minus seines durchschnittlichen 0/1 loss gegenüber den Referenzindizes, und die dahinterstehende Häufigkeitstabelle ist eine Konfusionsmatrix. In Krems fallen 174 der 198 kalten Tage in das kalte Halbjahr und 159 der 167 warmen Tage in das warme Halbjahr, eine Genauigkeit von 91.2 Prozent. Die Uneinigkeit konzentriert sich auf vier Monate (Abb. 1(b)): Januar, Februar, November und Dezember liegen vollständig im kalten Cluster, Mai bis August vollständig im warmen, und die Tage der Monate März, April, September und Oktober sind aufgeteilt.
Die andere Methode zählt Paare von Datenpunkte. Sei a die Anzahl der Paare, die beide Partitionen in einem Cluster platzieren, und e die Anzahl derer, die

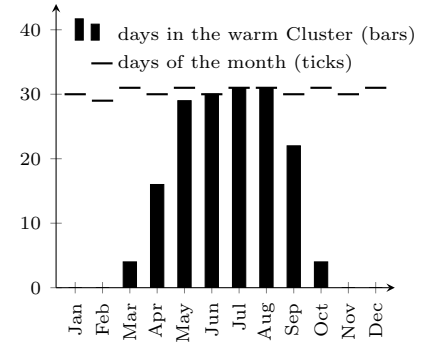
beide in verschiedenen Cluster platzieren. Der Rand-Index ist

$$\frac{a+e}{\binom{m}{2}}, \tag{3}$$

der Anteil der Paare, bei denen die Partitionen übereinstimmen (Murphy, 2012, Sect. 25.1.2; Rand, 1971). Er ist 1 für identische Partitionen und durch Umnummerierung unverändert, da er nie einen Index liest. Von den 66430 Paaren von Tagen in Krems sind 27916 in beiden Partitionen zusammen und 27858 getrennt, ein Rand-Index (3) von 0.840. Ein angepasster Rand-Index subtrahiert den Wert, der für zwei unabhängige Zufallspartitionen mit denselben Cluster-Größen erwartet wird (Hubert and Arabie, 1985). Eine Referenzpartition ist ein von der Anwendung festgelegter Standard, keine Eigenschaft der Merkmalsvektoren. Mit $k = 4$ Cluster gegenüber den vier Jahreszeiten sinkt die Genauigkeit nach der Umnummerierung auf 58.1 Prozent und der Rand-Index auf 0.730. Frühlings- und Herbsttage haben die gleichen Temperaturen: Das Cluster, das die meisten Frühlingstage enthält, besteht zu 53 Prozent aus Frühlingstagen und zu 21 Prozent aus Herbsttagen. Bei der image segmentation weist ein hartes Clustering der Pixel eines Fotos nach Farbe jedem Pixel genau eine Region zu, und seine Bewertung gegenüber einer handgezeichneten Segmentierung verwendet dieselben beiden Maße.



(a) 21 of the 365 days and the two Cluster



month of 2024

(b) the warm Cluster against the warm half-year

Abbildung 1: (a) Jeder 18 von 365 Tagen des Jahres 2024 in Krems, 21 Tage, dargestellt als kalte Tage (Kreise) oder warme Tage (Quadrate) durch das Cluster, dem k-Means mit $k = 2$ sie zuweist; die Kreuze sind die beiden Clusterzentren und die gestrichelte Linie ist die Grenze zwischen den beiden Cluster, die Mittelsenkrechte der Strecke zwischen den Clusterzentren. (b) Für jeden Monat des Jahres 2024 die Anzahl seiner Tage im warmen Cluster (Balken) und die Anzahl seiner Tage (horizontale Teilstriche); die Referenzpartition ordnet April bis September dem warmen Halbjahr zu. Daten generiert von `pythondemos/hardclustering.py`

Siehe auch: Clustering, Cluster, soft clustering, Zugehörigkeitsgrad, Clus-

terzentrum, clustering error, k-Means, Lloyd's algorithm, DBSCAN, spectral clustering, graph clustering, GMM, EM-Algorithmus (EM), probability simplex, Label, Klassifikation, unüberwachtes Lernen, ERM, Hypothese, Trainingsfehler, Genauigkeit, 0/1 loss, Konfusionsmatrix, image segmentation.

Querverweise

Literatur

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

2. Ester et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

3. Hubert and Arabie (1985). Comparing Partitions. Journal of Classification, 2(1), pp. 193–218.

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

5. Lloyd (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theory, 28(2), pp. 129–137.

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

7. MacQueen (1967). Some methods for classification and analysis of multivariate observations. In: Proc. Fifth Berkeley Symp. Math. Statist. Probab., pp. 281–297. University of California Press.

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

9. Rand (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336), pp. 846–850.

10. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

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