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

Clustering zerlegt einen Datensatz in Teilmengen, die Cluster genannt werden, sodass Datenpunkte innerhalb desselben Cluster einander ähnlicher sind als Datenpunkte in anderen Cluster. Clustering-Methoden unterscheiden sich im measure der Ähnlichkeit zwischen Datenpunkte, in der Repräsentation eines Cluster und darin, ob jeder Datenpunkt genau einem Cluster (hartes Clustering) oder mehreren mit variierenden Zugehörigkeitsgrade (soft clustering) zugeordnet wird. K-Means repräsentiert ein Cluster durch einen Clusterzentrum; seine Cluster sind die Bereiche des Merkmalsraum, die einem Clusterzentrum näher sind als jedem anderen, welche konvex sind. Ein Gaußsches Mischmodell (GMM) repräsentiert ein Cluster durch eine multivariate Normalverteilung; seine Cluster sind Ellipsoide, die Niveaumengen dieser Verteilung. Graph clustering repräsentiert die Datenpunkte als die Knoten eines similarity graph und verwendet dessen gewichtete Kanten, um die Cluster zu bestimmen.

Definition

Ein in der Wachau, dem Donautal westlich von Krems an der Donau (Österreich), aufgenommenes Foto zeigt Himmel, Wolken, bewaldete Hänge, sonnenbeschienene Weinberge und Felsen (Abb. 1(a)). Das Zerschneiden in 5×5 Pixel große Patches ergibt 12288 Datenpunkte, von denen jeder durch einen Merkmalsvektor $\mathbf{x} \in \mathbb{R}^3$ charakterisiert ist, der die mittleren Rot-Grün-Blau-Werte (RGB) des Patches enthält. Clustering ist die Aufgabe, einen solchen Datensatz in Cluster zu zerlegen, sodass Datenpunkte innerhalb desselben Cluster einander ähnlicher sind als Datenpunkte in anderen Cluster (Bishop, 2006, Sect. 9.1; Hastie et al., 2009, Sect. 14.3). Das Einfärben jedes Patches mit der mittleren Farbe seines Cluster teilt das Foto in Regionen, eine image segmentation. Abb. 1(b)–(f) zeigt die Segmentierungen, die von fünf Clustering-Methoden geliefert werden: k-Means, einem Gaußsches Mischmodell (GMM), density-based spatial clustering of applications with noise (DBSCAN), spectral clustering und convex clustering. Die Anzahl k der Cluster ist entweder eine Eingabe, wie bei k-Means, oder wird von der Methode aus ihren anderen Hyperparameter bestimmt, wie bei DBSCAN.

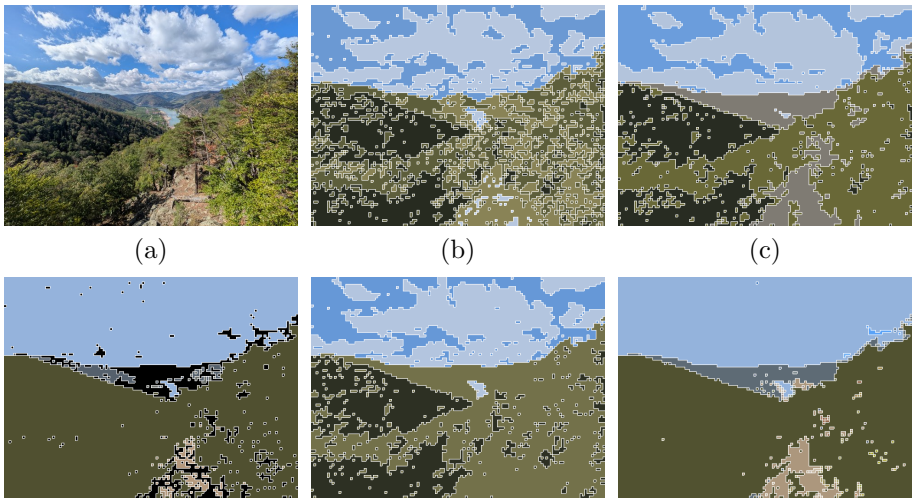


Abbildung 1: Image segmentation durch Clustering. (a) Ein in der Wachau (Österreich) aufgenommenes Foto, zerlegt in 96×128 Patches von 5×5 Pixeln; der Merkmalsvektor eines Patches enthält seine mittleren RGB-Werte. (b)–(f) Jeder Patch in der mittleren Farbe seines Cluster eingefärbt, mit weißen Linien zwischen den Cluster: (b) k-Means und (c) ein GMM, beide mit $k = 5$; (d) DBSCAN mit der Abstand $\varepsilon = 0.03$ und der minimalen Cluster- Größe $m^* = 32$, Rausch-Patches in Schwarz; (e) spectral clustering mit $k = 5$ und (f) convex clustering mit dem Regularisierungsparameter $\alpha \approx 18$, beide verwenden für jeden Patch die zehn Patches der ähnlichsten Farbe als seine Nachbarn. Daten generiert von `pythondemos/clustering.py`

Clustering-Methoden unterscheiden sich in drei Entwurfsentscheidungen (Jain et al., 1999):

1) das measure der Ähnlichkeit zwischen Datenpunkte (Hastie et al., 2009, Sect. 14.3), z.B. die euklidischer Abstand zwischen ihren Merkmalsvektoren, die den Merkmalsraum zu einem metric space macht, oder die Kantengewichte eines Graph, dessen Knoten die Datenpunkte sind;

2) die Repräsentation eines Cluster, z.B. ein Clusterzentrum, ein Punkt des Merkmalsraum, der für das Cluster steht, eine multivariate Normalverteilung, eine Menge von Datenpunkte, die durch kurze Sprünge verbunden sind, oder eine gut verbundene Teilmenge der Knoten eines Graph; und

3) ob jeder Datenpunkt genau einem Cluster (hartes Clustering) oder mehreren mit variierenden Zugehörigkeitsgrade (soft clustering) zugeordnet wird.

Diese Entscheidungen ergeben drei Familien von Clustering-Methoden. Hartes Clustering, wie bei k-Means, verwendet die euklidischer Abstand (Wahl 1), einen

Clusterzentrum (Wahl 2) und ein Cluster pro Datenpunkt (Wahl 3). Soft clustering, wie bei einem GMM, behält die erste Wahl bei, repräsentiert ein Cluster durch eine multivariate Normalverteilung und ordnet jeden Datenpunkt jedem Cluster mit einem Zugehörigkeitsgrad zu. Graph clustering ersetzt die von k-Means und einem GMM verwendete euklidische Metrik durch einen diskreten metric space: einen ungerichteten gewichteten Graph, dessen Knoten die Datenpunkte sind und dessen Kantengewichte ihre Ähnlichkeit ausdrücken. Ein Cluster ist dann eine dicht verbundene Teilmenge seiner Knoten, und die Methoden sind hartes Clusterings, wobei jeder Knoten einer Teilmenge beitritt.

K-Means repräsentiert ein Cluster durch einen Punkt des Merkmalsraum, der als Clusterzentrum bezeichnet wird. Jeder Datenpunkt wird dem Cluster zugeordnet, dessen Clusterzentrum in euklidischer Abstand am nächsten liegt, sodass ein Cluster der Bereich des Merkmalsraum ist, der seinem Clusterzentrum näher ist als jedem anderen, eine konvex Menge. Die Clusterzentren werden so gewählt, dass die durchschnittliche quadrierte euklidischer Abstand zwischen einem Datenpunkt und seinem nächsten Clusterzentrum minimiert wird. Eine weit verbreitete Methode zur Lösung dieses Optimierungsproblem ist der Lloyd's algorithm (Lloyd, 1982). Diese iterative Optimierungsverfahren wechselt zwischen einem Zuordnungsschritt und einem Mittelungsschritt: Jeder Schritt senkt den Durchschnitt, und die Iterationen konvergieren zu einem lokalen minimum, nicht notwendigerweise dem globalen (Bishop, 2006, Sect. 9.1). Jeder Datenpunkt gehört zu genau einem Cluster, daher ist k-Means ein hartes Clustering (Abb. 2(a)).

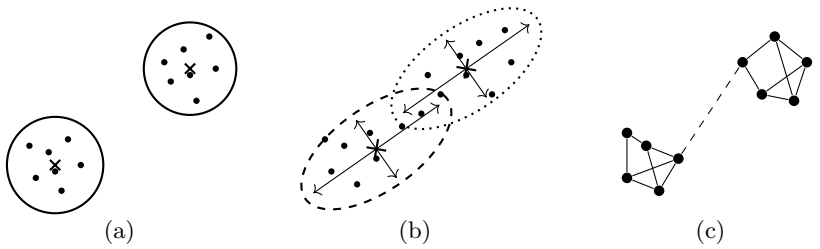


Abbildung 2: Drei Familien von Clustering-Methoden, skizziert für Datenpunkte mit Merkmalsvektoren in $\mathbb{R}^2$. (a) Hartes Clustering wie bei k-Means: Jeder Datenpunkt gehört zu genau einem Cluster, dem mit dem nächsten Clusterzentrum (Kreuze). (b) Soft clustering wie bei einem GMM: zwei überlappende Ellipsen (gestrichelte und gepunktete Ellipsen), Kreuze an ihren Mittelwerten und dünne Doppelpfeile entlang ihrer Hauptrichtungen; ein Datenpunkt in der Überlappung gehört zu beiden Cluster mit einem Zugehörigkeitsgrad zwischen 0 und 1. (c) Graph clustering: Die Datenpunkte sind die Knoten eines Graph, dessen Kanten ähnliche Datenpunkte verbinden. Zwei Cluster mit jeweils fünf Knoten: Die durchgezogenen Kanten verlaufen innerhalb eines Cluster, und die einzelne gestrichelte Kante ist die einzige zwischen den beiden, sodass jedes Cluster eine dicht verbundene Teilmenge von Knoten mit wenigen Kanten zum Rest

Auf dem Foto in Abb. 1(a) sind die fünf Clusterzentren von k-Means mit $k = 5$ fünf repräsentative Farben. Jedes Cluster besteht aus den Patches, deren Farbe einem von ihnen am nächsten ist: Himmel, Wolke, schattiger Wald, sonnenbeschienene Hänge und Felsen (Abb. 1(b)). Das Einfärben jedes Patches mit der Farbe seines Clusterzentrum komprimiert das Foto auch auf fünf Farben. Ein aus einem Foto gelerntes Clustering wird danach beurteilt, wie es neue Patches zuordnet: Bei k-Means tritt ein neuer Datenpunkt dem Cluster des nächsten Clusterzentrum bei, sodass die Clusterzentren die gelernten Hypothese sind.

Für einige Anwendungen ist die Repräsentation eines Cluster allein durch einen Clusterzentrum unzureichend, da die Cluster unterschiedliche Formen haben könnten, z.B. in verschiedene Richtungen des Merkmalsraum gestreckt sein. Ein Beispiel hierfür sind die beiden Cluster in Abb. 2(b), die entlang unterschiedlicher Richtungen gestreckt sind. Hier ist es nützlich, jedes Cluster, indiziert $c = 1, \ldots, k$, durch eine multivariate Normalverteilung $\mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)$ zu repräsentieren. Diese multivariate Normalverteilungen repräsentieren überlappende Cluster, deren relative Größen durch Cluster Wahrscheinlichkeiten p_c für $c = 1, \ldots, k$ bestimmt werden. Zusammen mit diesen Cluster Wahrscheinlichkeiten definieren die multivariate Normalverteilungen ein GMM (Bishop, 2006, Sect. 9.2).

Eine soft clustering-Methode wird erhalten, indem die GMM Modellparameter $p_c, \boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}$ für $c = 1, \ldots, k$ über einen approximativen maximum likelihood-Schätzer wie den EM-Algorithmus (EM) Algorithmus (Bishop, 2006, Sect. 9.2) gelernt werden. Für jeden Datenpunkt ist der Zugehörigkeitsgrad zum c-ten Cluster die A-posteriori-Verteilung seines Cluster-Index. Im Merkmalsraum können die resultierenden Cluster durch Ellipsoide, die Niveaumengen einer multivariate Normalverteilung, dargestellt werden (Bishop, 2006, Sect. 2.3). Hartes Clustering ist der Extremfall einer soft clustering-Methode, die erzwingt, dass jeder Zugehörigkeitsgrad entweder 0 oder 1 ist. Dieser Extremfall entspricht dem Grenzwert eines GMM, bei dem jede Komponente dieselbe Kovarianzmatrix $\epsilon\mathbf{I}$ hat und $\epsilon \to 0$ (Bishop, 2006, Sect. 9.3.2).

Auf dem Foto in Abb. 1(a) wird jeder Oberflächentyp zu einer multivariate Normalverteilung über Farben. Jeder Patch erhält fünf Zugehörigkeitsgrade, einen pro Cluster, nicht-negativ und sich zu eins summierend; sein Zugehörigkeitsgrad zum Himmels-Cluster ist die Wahrscheinlichkeit, dass seine Farbe aus der multivariate Normalverteilung des Himmels gezogen wurde. Ein Patch, dessen größter Zugehörigkeitsgrad 0.9 beträgt, wird fast vollständig durch ein Cluster erklärt; ein Patch, dessen größter Zugehörigkeitsgrad 0.6 beträgt, wird geteilt, wobei die restlichen 0.4 auf die anderen Cluster verteilt sind, wie ein Datenpunkt in der Überlappung von Abb. 2(b). Das Einfärben jedes Patches mit dem Cluster mit seinem größten Zugehörigkeitsgrad ergibt das hartes Clustering von Abb. 1(c); die Überlappung selbst wird nicht gezeichnet. Zwölf Prozent der Patches, einer von acht, hauptsächlich entlang des Horizonts und der Wolkenränder, haben keinen Zugehörigkeitsgrad über 0.8.

DBSCAN definiert ein Cluster durch zwei Hyperparameter, eine Abstand ε und eine Mindestgröße m^* eines Cluster (Ester et al., 1996). Ein Datenpunkt ist ein Kernpunkt, wenn mindestens m^* Datenpunkte innerhalb der euklidischer

Abstand ε von ihm liegen. Ein Cluster ist eine maximale Menge von Datenpunkte, die durch eine Kette von Kernpunkten verbunden sind, die höchstens ε voneinander entfernt sind. Datenpunkte, die weder Kernpunkte noch innerhalb von ε von einem liegen, werden als *Rauschen* deklariert. Ein Cluster kann daher jede Form annehmen, die eine Kette benachbarter Datenpunkte nachzeichnen kann, und die Anzahl der Cluster ist eine Ausgabe, keine Eingabe. Auf dem Foto in Abb. 1(a) ist ε eine Abstand zwischen Farben. Mit den beiden Hyperparameter von Abb. 1(d) werden das Blau des Himmels und das Weiß der Wolken durch eine Kette von Zwischenfarben verbunden, sodass Himmel und Wolke ein Cluster und das Land ein anderes bilden. Die Patches entlang des Horizonts, deren Mischfarben selten sind, sind Rauschen (Abb. 1(d)).
Graph clustering-Methoden, darunter spectral clustering, flow-based clustering und convex clustering, stellen den Datensatz als die Knoten $r = 1, \ldots, m$ eines Graph dar, dessen Kantengewichte die Ähnlichkeit zwischen Datenpunkte ausdrücken. Wenn es beispielsweise eine Vorstellung von Abstand zwischen Datenpunkte gibt, verbindet ein k-nearest-neighbor graph (k-NN graph) jeden Datenpunkt mit seinen nächsten Nachbarn (von Luxburg, 2007, Sect. 2). Die Cluster werden durch die Konnektivität der Knoten bestimmt: Ein Cluster ist eine Teilmenge von Knoten, deren gesamtes Kantengewicht zum Rest des Graph, geteilt durch die Anzahl der Knoten in der Teilmenge, klein ist (Abb. 2(c)). Die Summe dieses Verhältnisses über die Cluster ist der Ratio Cut, den spectral clustering annähernd minimiert (von Luxburg, 2007, Sect. 5). Spectral clustering und flow-based clustering verwenden nur den Graph; convex clustering kombiniert die Konnektivität der Knoten mit den knotenweisen Merkmalsvektoren $\mathbf{x}^{(r)}$, $r = 1, \ldots, m$ (Hocking et al., 2011) (Abb. 1(f)).

Spectral clustering berechnet die Cluster aus Eigenvektoren der Laplacian matrix des Graph (von Luxburg, 2007; Ng et al., 2001) und flow-based clustering aus Netzwerkflüssen zwischen ausgewählten Knoten (SarcheshmehPour et al., 2021). Diese beiden Methoden, die nur den Graph verwenden, eignen sich für Datenpunkte ohne offensichtliche Wahl für numerische Merkmalsvektoren: wissenschaftliche Arbeiten, die durch Zitate verbunden sind, Benutzer eines sozialen Netzwerks, die durch Freundschaften verbunden sind, oder Webseiten, die durch Links verbunden sind. Auf dem Foto in Abb. 1(a) sind die Knoten die Patches, und jeder Patch ist mit den zehn Patches der ähnlichsten Farbe verbunden. Spectral clustering trennt dann die fünf Oberflächentypen (Abb. 1(e)), da die Patches eines Oberflächentyps farblich Nachbarn sind. Abb. 3 ordnet die Methoden nach der Darstellung der Datenpunkte, als euklidische Vektoren oder als die Knoten eines similarity graph, und eines Cluster an.

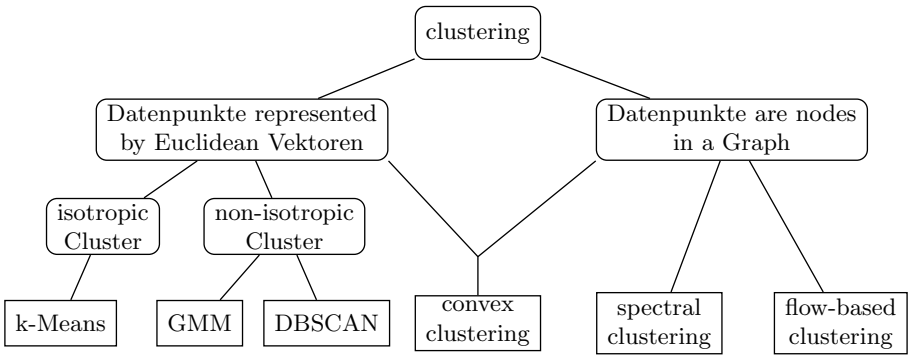


Abbildung 3: Clustering-Methoden, angeordnet nach der Darstellung der Datenpunkte: als euklidische Vektoren oder als die Knoten eines Graph mit gewichteten Kanten. Methoden auf dem linken Zweig unterscheiden sich in den Cluster-Formen, die sie darstellen können: k-Means liefert konvexe Cluster, die Regionen des Merkmalsraum, die einem Clusterzentrum näher sind als jedem anderen; ein GMM liefert ellipsenförmige Cluster und DBSCAN jede Form, die eine Kette nahegelegener Datenpunkte nachzeichnen kann. Methoden auf dem rechten Zweig bestimmen die Cluster aus der Konnektivität der Knoten: spectral clustering aus Eigenvektoren der Laplacian matrix, flow-based clustering aus Netzwerkflüssen zwischen ausgewählten Knoten. Convex clustering verwendet beide Darstellungen

Convex clustering weist jedem Datenpunkt $\mathbf{x}^{(r)}$ einen eigenen Clusterzentrum $\mathbf{w}^{(r)}$ zu, was die Nicht-Konvexität des clustering error vermeidet. Seine konvexe Zielfunktion addiert zu den quadrierten euklidische Abstände zwischen den Datenpunkte und ihren Clusterzentren einen Strafterm, gewichtet mit einem Regularisierung- Parameter $\alpha > 0$ und mit den Kantengewichte $A_{r,r'}$, der die Clusterzentren benachbarter Datenpunkte verschmilzt (Hocking et al., 2011; Sun et al., 2021, Gl. (1); Lindsten et al., 2011). Die Datenpunkte, die einen Clusterzentrum am eindeutigen Minimierer teilen, bilden ein Cluster, und die Anzahl der Cluster wird indirekt über α festgelegt: Bei $\alpha = 0$ ist jeder Datenpunkt sein eigenes Cluster, und für große α ist eine zusammenhängende Komponente des Graph ein Cluster. Auf dem Foto, mit den Kantengewichte des Farb-Graph der Patches und dem α aus Abb. 1(f), lassen die verschmolzenen Clusterzentren 65 Cluster übrig, vier davon mit mindestens 80 Patches; wie DBSCAN verschmilzt die Methode Himmel mit Wolken entlang des allmählichen Farbübergangs (Abb. 1(f)).

Die Wahl einer Methode folgt den drei Entwurfsentscheidungen. K-Means ist die Standardwahl, wenn die Anzahl der Cluster bekannt ist und isolierte, kompakte Cluster erwartet werden (Jain et al., 1999, Sect. 5.2.1). Ein GMM ist die Wahl, wenn der Zugehörigkeitsgrad wichtig ist, wie bei den gemischten Farben entlang des Horizonts (Bishop, 2006, Sect. 9.3.2). DBSCAN ist die Wahl, wenn ein Cluster jede beliebige Form haben kann, die Anzahl der Cluster unbekannt ist und isolierte Datenpunkte als Rauschen markiert werden sollen (Ester et al.,

1996, Sect. 1). Spectral clustering ist die Wahl, wenn die Ähnlichkeit als Graph gegeben ist, mit der Anzahl der Cluster als Eingabe; es wird nichts über die Form eines Cluster angenommen, und die Wahl des Graph entscheidet über das Ergebnis (von Luxburg, 2007, Sect. 9). Convex clustering ist die Wahl, wenn die Datenpunkte sowohl Merkmalsvektoren als auch einen Graph tragen, der die Cluster formen soll, wie z.B. Benutzer mit Profilen und Freundschaften, und wenn das Ergebnis nicht vom Start abhängen darf: Seine Zielfunktion ist konvex, sodass jeder Solver das gleiche Clustering erreicht, während der Lloyd's algorithm an einem lokalen minimum anhalten kann (Sun et al., 2021, Sect. 1). Auf dem Foto teilen sich die Methoden entlang einer Linie auf. DBSCAN und convex clustering folgen Ketten ähnlicher Datenpunkte, wie es Single Linkage tut (Hastie et al., 2009, Sect. 14.3.12), und verschmelzen daher Himmel mit Wolken über den allmählichen Farbübergang hinweg (Abb. 1(d), (f)), während k-Means und spectral clustering diesen Übergang durchschneiden, erstere, weil sie die Abstand zu einem Clusterzentrum misst, letztere, weil sie einen ausgewogenen Schnitt minimiert (von Luxburg, 2007, Sect. 5) (Abb. 1(b), (e)).

Über die image segmentation hinaus wird Clustering für die Kundensegmentierung (Wedel and Kamakura, 2000), für die Gruppierung von Dokumenten nach Themen in natural language processing (NLP) (Manning et al., 2008, Sect. 16.1), für die Erkennung von Gemeinschaften in sozialen Netzwerken (Fortunato, 2010) und für die anomaly detection verwendet, wobei ein Datenpunkt, der in kein Cluster oder in ein winziges fällt, ein Ausreißer ist (Chandola et al., 2009). Als preprocessing-Schritt für überwachtes Lernen reduziert Clustering einen Datensatz auf seine Clusterzentren als Prototypen (Hastie et al., 2009, Sect. 13.2.1) oder fügt die Abstände eines Datenpunkt zu den Clusterzentren als neue Merkmale hinzu (Coates and Ng, 2012). Es kann auch den Datensatz aufteilen, sodass pro Cluster eine separate Hypothese gelernt wird, z.B. eine lineare Regression pro Cluster (Späth, 1979). Für einen großen Datensatz wird k-Means auf einem coreset ausgeführt: einer kleinen gewichteten Teilmenge der Datenpunkte, deren gewichteter durchschnittlicher quadrierter euklidischer Abstand zum nächsten Clusterzentrum den des gesamten Datensatz für jede Wahl der Clusterzentren annähert (Har-Peled and Mazumdar, 2004).

Synonyme: Segmentierung.

Siehe auch: Cluster, Clusterzentrum, clustering error, Zugehörigkeitsgrad, hartes Clustering, soft clustering, k-Means, GMM, EM, DBSCAN, spectral clustering, graph clustering, flow-based clustering, convex clustering, coreset, Hyperparameter, anomaly detection, image segmentation, similarity graph, Laplacian matrix, k-NN graph, unüberwachtes Lernen.

Querverweise

Literatur

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

2. Chandola et al. (2009). Anomaly Detection: A Survey. ACM Comput. Surv., 41(3), pp. 1–58.

3. Coates and Ng (2012). Learning Feature Representations with K-Means. In: Neural Networks: Tricks of the Trade, pp. 561–580. Springer.

4. 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.

5. SarcheshmehPour et al. (2021). Flow-Based Clustering and Spectral Clustering: A Comparison. In: 2021 55th Asilomar Conf. Signals, Syst., Comput., pp. 1292–1296.

6. Fortunato (2010). Community detection in graphs. Physics Reports, 486(3–5), pp. 75–174.

7. Har-Peled and Mazumdar (2004). On coresets for k-means and k-median clustering. In: Proc. 36th Annual ACM Symp. Theory of Computing (STOC), pp. 291–300.

8. Hocking et al. (2011). Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties. In: Proc. of the 28th International Conference on Machine Learning,.

9. Sun et al. (2021). Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm. J. Mach. Learn. Res., 22(9), pp. 1–32.

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

11. Lindsten et al. (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation. In: 2011 IEEE Statistical Signal Processing Workshop (SSP), pp. 201–204.

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

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

14. Manning et al. (2008). Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

15. Ng et al. (2001). On Spectral Clustering: Analysis and an Algorithm. In: Adv. Neural Inf. Process. Syst., pp. 849–856.

16. Späth (1979). Algorithm 39: Clusterwise linear regression. Computing, 22(4), pp. 367–373.

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

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