k-Means++

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

k-Means++ ist eine randomisierte Regel für die anfänglichen Clusterzentren des Lloyd-Algorithmus: Der erste ist ein gleichverteilt zufällig gezogener Datenpunkt, jeder weitere ein Datenpunkt, der mit einer Wahrscheinlichkeit proportional zu seiner quadrierten euklidischer Abstand zum nächstgelegenen bereits gewählten Clusterzentrum gezogen wird. Die anfänglichen Clusterzentren werden dadurch über den Datensatz verteilt, sodass der Lloyd-Algorithmus selten mit mehreren von ihnen im selben Cluster startet. Der erwartete Clusteringfehler der anfänglichen Clusterzentren beträgt höchstens das $8(\ln k + 2)$-Fache des kleinsten mit k Clusterzentren erreichbaren Clusteringfehlers, für jeden Datensatz und vor einer einzigen Iteration des Lloyd-Algorithmus; eine gleichverteilte Initialisierung bietet keine solche Garantie. Da der Lloyd-Algorithmus den Clusteringfehler niemals vergrößert, gilt die Schranke auch für seine Ausgabe. Die Initialisierung ändert weder die Fixpunktiteration des Lloyd-Algorithmus noch seine Fixpunkte; sie wählt den Startpunkt und damit, welcher Fixpunkt erreicht wird. Auf den Bildausschnitten einer Fotografie mit drei Clustern erreicht der Lloyd-Algorithmus das beste gefundene Clustering ausgehend von 64% der k-Means++-Startpunkte und von 48% der gleichverteilten Startpunkte, und er benötigt weniger Iterationen.

Definition

Der Lloyd-Algorithmus für k-Means benötigt vor seiner ersten Iteration k anfängliche Clusterzentren, und das Clustering, das er liefert, hängt von ihnen ab. Betrachtet sei eine Fotografie des Ötschermassivs, die in 972 quadratische Bildausschnitte zerschnitten ist, von denen jeder ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$ ist, der seinen durchschnittlichen Grün- und Blauanteil enthält. Vegetation, Berg und Dunst sowie Himmel bilden drei Cluster von Bildausschnitten, und die besten für $k = 3$ gefundenen Clusterzentren haben den Clusteringfehler 0.018. Gestartet von drei gleichverteilt zufällig gezogenen Bildausschnitten, endet der Lloyd-Algorithmus in 516 von 1000 Durchläufen an einem anderen Fixpunkt mit dem Clusteringfehler 0.025: Zwei Clusterzentren teilen sich die Vegetation, und die Bergausschnitte werden zwischen dem Himmels-Clusterzentrum und einem der Vegetations-Clusterzentren aufgeteilt. k-Means++ ist eine randomisierte Regel für die anfänglichen Clusterzentren,

die diese über den Datensatz verteilt (Arthur and Vassilvitskii, 2007). Gestartet von ihren Startpunkten, endet der Lloyd-Algorithmus in 359 von 1000 Durchläufen am schlechteren Fixpunkt.

Der Clusteringfehler ist keine konvexe Funktion der Clusterzentren. Der Lloyd-Algorithmus vergrößert ihn nie und hält an einem Fixpunkt an, der ein lokales Minimum anstelle eines globalen sein kann; welcher Fixpunkt erreicht wird, hängt von den anfänglichen Clusterzentren ab (Bishop, 2006, Abschn. 9.1; Hastie et al., 2009, Abschn. 14.3.6). Die übliche Abhilfe besteht darin, den Lloyd-Algorithmus von mehreren anfänglichen Wahlen aus zu starten und den Durchlauf mit dem kleinsten Clusteringfehler zu behalten (Hastie et al., 2009, Abschn. 14.3.6), wobei die einfachste anfängliche Wahl aus k gleichverteilt zufällig gezogenen Datenpunkte besteht (Bishop, 2006, Abschn. 9.1). Eine gleichverteilte Initialisierung neigt dazu, mehrere anfängliche Clusterzentren im größten Cluster zu platzieren: 566 der 972 Bildausschnitte sind Vegetation, sodass alle drei gleichverteilten Startpunkte mit einer Wahrscheinlichkeit von 0.20 Vegetationsausschnitte sind. k-Means++ zieht die Startpunkte nacheinander und macht die Datenpunkte, die weit von den bisher gewählten Startpunkten entfernt liegen, zu den wahrscheinlichen nächsten (Algorithmus 1). Der erste Startpunkt $\mathbf{w}^{(1)}$ ist ein gleichverteilt zufällig aus dem Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ gezogener Datenpunkt. Für $c = 2, \ldots, k$ ist der Startpunkt $\mathbf{w}^{(c)}$ ein Datenpunkt, der mit einer Wahrscheinlichkeit proportional zu seiner quadrierten euklidischer Abstand zum nächstgelegenen bereits gewählten Startpunkt gezogen wird: Er gleicht dem Datenpunkt $\mathbf{x}^{(r)}$, für $r = 1, \ldots, m$, mit der Wahrscheinlichkeit

$$\mathbb{P}\left(\mathbf{w}^{(c)} = \mathbf{x}^{(r)}\right) = \frac{\min_{c'=1,\ldots,c-1} \left\|\mathbf{x}^{(r)} - \mathbf{w}^{(c')}\right\|_2^2}{\sum_{r'=1}^{m} \min_{c'=1,\ldots,c-1} \left\|\mathbf{x}^{(r')} - \mathbf{w}^{(c')}\right\|_2^2}. \tag{1}$$

Ein Datenpunkt, der bereits ein Startpunkt ist, hat unter der Initialisierungsregel (1) die Wahrscheinlichkeit 0, und ein Datenpunkt, der doppelt so weit von den Startpunkten entfernt ist wie ein anderer, wird viermal so wahrscheinlich gezogen. Ist ein Vegetationsausschnitt als erster Startpunkt gegeben, so ist der zweite Startpunkt ein Himmelsausschnitt mit einer Wahrscheinlichkeit von 0.74 unter der Initialisierungsregel (1) und 0.21 unter gleichverteilter Initialisierung, dem Anteil der Himmelsausschnitte (Abb. 2(a)). Er ist ein weiterer Vegetationsausschnitt mit einer Wahrscheinlichkeit von 0.06 gegenüber 0.58. Die Startpunkte sind die Eingabe des Lloyd-Algorithmus, und die Kombination der Initialisierung mit dem Lloyd-Algorithmus ist das, was Arthur and Vassilvitskii (2007) k-Means++ nannten. Ostrovsky et al. (2006) schlugen dieselbe Initialisierung unabhängig davon vor.

Algorithm 1 k-Means++-Initialisierung (Arthur and Vassilvitskii, 2007, Abschn. 2.2)

Input: Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, Cluster count k

draw $\mathbf{w}^{(1)}$ uniformly at random from $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$

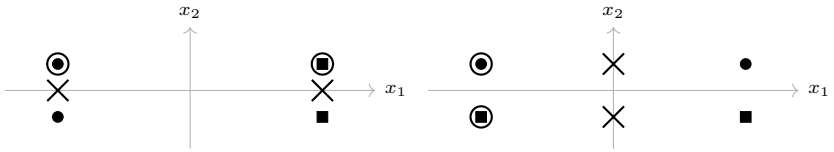
for $c = 2, \ldots, k$ **do**

    draw $\mathbf{w}^{(c)}$ from $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ with the Wahrscheinlichkeiten (1)

end for

Output: initial Clusterzentren $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}$ for Lloyd-Algorithmus

Abb. 1 zeigt die Auswirkung der Startpunkte auf vier Datenpunkte mit den Merkmalsvektoren $(-5, 1)^\top$, $(-5, -1)^\top$, $(5, 1)^\top$ und $(5, -1)^\top$ sowie $k = 2$. Zwei Startpunkte auf gegenüberliegenden Seiten führen den Lloyd-Algorithmus zur Links-Rechts-Aufteilung mit den Clusterzentren $(\pm 5, 0)^\top$ und dem Clusteringfehler 1, dem kleinstmöglichen. Zwei Startpunkte auf derselben Seite weisen jeden rechten Datenpunkt dem Startpunkt auf gleicher Höhe zu, sodass der erste Mittelungsschritt die Clusterzentren $(0, \pm 1)^\top$ liefert. Diese sind ein Fixpunkt des Lloyd-Algorithmus, und die durch sie definierte Oben-Unten-Aufteilung hat den Clusteringfehler 25. Gleichverteilte Initialisierung zieht zwei Startpunkte auf derselben Seite mit der Wahrscheinlichkeit 1/3, zwei der sechs Paare. Unter k-Means++ ist der zweite Startpunkt gemäß der Initialisierungsregel (1) mit der Wahrscheinlichkeit $4/(4{+}100{+}104) = 1/52$ der andere Datenpunkt auf der Seite des ersten. Der erwartete Clusteringfehler nach dem Lloyd-Algorithmus beträgt daher 9 unter gleichverteilter Initialisierung und 1.46 unter k-Means++.



(a) seeds on opposite sides: left/right split (b) seeds on the same side: top/bottom split

Abbildung 1: Vier Datenpunkte und zwei Startpunkte (Ringe) für den Lloyd-Algorithmus mit $k = 2$; die Kreuze sind die Clusterzentren am erreichten Fixpunkt, und ein Datenpunkt ist je nach dem Cluster, in dem er landet, als ausgefüllter Kreis oder als ausgefülltes Quadrat dargestellt. (a) Startpunkte auf gegenüberliegenden Seiten, die von k-Means++ mit der Wahrscheinlichkeit 51/52 gezogen werden, führen zu der Links/Rechts-Aufteilung mit dem Clusteringfehler 1. (b) Startpunkte auf derselben Seite, die durch gleichverteilte Wahl mit der Wahrscheinlichkeit 1/3 gezogen werden, führen zu den Clusterzentrenn $(0, \pm 1)^\top$, einem Fixpunkt des Lloyd-Algorithmus, dessen Oben/Unten-Aufteilung den Clusteringfehler 25 aufweist.

Schon die Startpunkte allein bringen eine Garantie mit sich. Es bezeichne $E^\star$ den kleinsten Clusteringfehler, den beliebige k Clusterzentren auf dem Datensatz erzielen. Die Startpunkte $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}$ aus Algorithmus 1 sind Zufallsvariablen

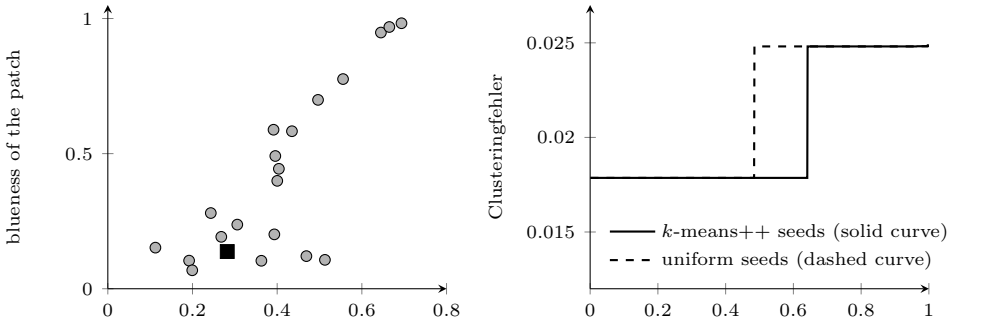
(ZV), und ihr erwarteter Clusteringfehler erfüllt für jeden Datensatz und vor einer einzigen Iteration des Lloyd-Algorithmus (Arthur and Vassilvitskii, 2007, Thm. 1.1)

$$\mathbb{E}\left\{\frac{1}{m}\sum_{r=1}^{m}\min_{c=1,\dots,k}\left\|\mathbf{x}^{(r)}-\mathbf{w}^{(c)}\right\|_2^2\right\}\leq 8\left(\ln k+2\right)E^\star. \tag{2}$$

Da der Lloyd-Algorithmus den Clusteringfehler niemals vergrößert, erfüllt seine von den Startpunkten ausgehende Ausgabe die Schranke (2) ebenfalls. Der Logarithmus lässt sich nicht beseitigen: Es gibt Datensätze, auf denen der erwartete Clusteringfehler der Startpunkte mindestens ein konstantes Vielfaches von $\ln k$ mal $E^\star$ beträgt (Arthur and Vassilvitskii, 2007, Abschn. 4). Gleichverteilte Initialisierung besitzt keine Schranke dieser Art: Auf den vier Datenpunkte aus Abb. 1 beträgt der erwartete Clusteringfehler zweier gleichverteilter Startpunkte 18, gegenüber 2.92 für k-Means++ und der Schranke $8\left(\ln 2+2\right)\cdot 1 = 21.55$ der Garantie (2). Bewegt man die beiden Paare auseinander auf $(\pm 50, \pm 1)^\top$, bleibt $E^\star = 1$ erhalten, der Wert bei Gleichverteilung steigt jedoch auf 1668, während der k-Means++-Wert 3.00 beträgt. Auf den Bildausschnitten liegt der mittlere Clusteringfehler von 1000 Startpunktmengen bei 0.038 für k-Means++ und 0.086 für gleichverteilte Initialisierung, dem 2.1- bzw. 4.8-Fachen des besten gefundenen Wertes; die schlechteste k-Means++-Startpunktmenge hat den Clusteringfehler 0.149, die schlechteste gleichverteilte Startpunktmenge 0.551.

Die Initialisierung ändert nichts am Lloyd-Algorithmus selbst. Fasst man die Clusterzentren in einem Vektor $\mathbf{w} = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(k)})$ zusammen, ist der Lloyd-Algorithmus die Fixpunktiteration $\mathbf{w}^{(t+1)} = \mathcal{F}(\mathbf{w}^{(t)})$ für $t = 0, 1, \dots$. Der Operator $\mathcal{F}$ wendet einen Zuweisungsschritt und einen Mittelungsschritt an. Seine Fixpunkte sind Clusterzentren, die jeweils dem Mittelwert der ihnen zugewiesenen Datenpunkte gleichen, und seine Konvergenz folgt aus der monotonen Abnahme des Clusteringfehlers. k-Means++ wählt lediglich den Startpunkt $\mathbf{w}^{(0)}$, und damit den Fixpunkt, der erreicht wird, sowie die Anzahl der Iterationen bis dorthin: Auf den Bildausschnitten benötigt der Lloyd-Algorithmus im Durchschnitt 10.0 Iterationen ausgehend von k-Means++-Startpunkten und 12.7 ausgehend von gleichverteilten Startpunkten. Die Initialisierung selbst ist kein iteratives Verfahren, sondern besteht aus k Ziehungen, sodass keine Fixpunktbetrachtung auf sie anwendbar ist; ihre Ausgabe ist eine ZV, und die Garantie (2) ist eine Aussage über ihren Erwartungswert. Die Initialisierung schließt auch den schlechteren Fixpunkt nicht aus; sie macht ihn seltener (Abb. 2(b)). Auf den Bildausschnitten enden 35.9% der Durchläufe ausgehend von k-Means++-Startpunkten und 51.6% der Durchläufe ausgehend von gleichverteilten Startpunkten beim Clusteringfehler 0.025; bei fünf Neustarts, von denen der Durchlauf mit dem kleinsten Clusteringfehler behalten wird, verfehlen somit alle fünf das beste Clustering mit einer Wahrscheinlichkeit von 0.6% gegenüber 3.7%. Im Sinne der empirische Risikominimierung (ERM) minimiert k-Means den Trainingsfehler einer Hypothese, die einen Merkmalsvektor auf seinen nächstgelegenen Clusterzentrum abbildet; die Initialisierung

ändert weder den Hypothesenraum noch die Verlustfunktion, sondern nur das Optimierungsverfahren und damit das lokale Minimum des Trainingsfehlers, das es findet.



(a) 20 of the 972 patches and the first seed (b) Clusteringfehler after Lloyd-Algorithmus, sorted

Abbildung 2: (a) Jeder 49. der 972 Bildausschnitte (20 Ausschnitte), dargestellt als Kreise, deren Fläche proportional zur Wahrscheinlichkeit (1) ist, der zweite Startpunkt zu sein, wenn der erste Startpunkt der durch das ausgefüllte Quadrat markierte Vegetationsausschnitt ist; die drei größten Kreise sind Himmelsausschnitte. (b) Der Clusteringfehler nach dem Lloyd-Algorithmus in 1000 Durchläufen mit k-Means++-Startpunkten (durchgezogene Kurve) und 1000 Durchläufen mit gleichverteilten Startpunkten (gestrichelte Kurve), jeweils aufsteigend sortiert über dem Anteil der Durchläufe; der Sprung von 0.018 auf 0.025 liegt bei 64% der k-Means++-Durchläufe und bei 48% der gleichverteilten Durchläufe. Daten erzeugt mit `pythondemos/kmeanspp.py`.

Jeden Bildausschnitt der Fotografie in der Durchschnittsfarbe seines Clusters einzufärben, entspricht einer Bildsegmentierung, und die beiden Fixpunkte liefern zwei unterschiedliche Segmentierungen. Die bessere trennt Himmel, Berg und Vegetation; die schlechtere färbt die 204 Himmelsausschnitte und 91 Bergausschnitte in einer Farbe ein und teilt die Vegetation entzwei. k-Means++ ist die Standardinitialisierung von k-Means in scikit-learn (Pedregosa et al., 2011).

Siehe auch: k-Means, Lloyd-Algorithmus, Clusterzentrum, Clusteringfehler, Clustering, Cluster, hartes Clustering, Fixpunkt, Fixpunktiteration, euklidischer Abstand, Minimum, konvex, ERM, Bildsegmentierung.

Querverweise

Literatur

1. Arthur and Vassilvitskii (2007). k-means++: the advantages of careful seeding. In: Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035. Society for Industrial and Applied Mathematics Philadelphia.

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

3. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), pp. 2825–2830.

4. Ostrovsky et al. (2006). The effectiveness of Lloyd-type methods for the k-means problem. In: Proc. 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp. 165–176.

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.