(Bild: © aga7ta - stock.adobe.com)

Definition

Was ist der k-Means-Algorithmus?

18.07.18 | Autor / Redakteur: Stefan Luber / Nico Litzel

Der k-Means-Algorithmus ist ein Rechenverfahren, das sich für die Gruppierung Objekten, die sogenannte Clusteranalyse, einsetzen lässt. Dank der effizienten I der Clusterzentren und dem geringen Speicherbedarf eignet sich der Algorithmu die Analyse großer Datenmengen, wie sie im Big-Data-Umfeld üblich sind.

Der k-Means-Algorithmus ist eines der am häufigsten verwendeten mathematisch zur Gruppierung von Objekten (Clusteranalyse). Der Algorithmus ist in der Lage, a Menge ähnlicher Objekte mit einer vorher bekannten Anzahl von Gruppen die jewe der Cluster zu ermitteln. Da es sich um ein sehr effizientes Verfahren handelt, das verschiedenen Datentypen zurecht kommt, und der Speicherbedarf gering ist, eig k-Means-Algorithmus für die Datenanalyse im Big-Data-Umfeld.

Die Laufzeit des Algorithmus ist linear zur Anzahl der vorhandenen Datenpunkte u der Schleifendurchläufe zur Ermittlung der Clusterzentren ist klein. Die Idee für de Algorithmus stammt ursprünglich von Hugo Steinhaus aus dem Jahr 1957. Erst 1 Algorithmus in einer Informatik-Zeitschrift veröffentlicht. Es besteht eine große Ä zum sogenannten Expectation-Maximization-Algorithmus. Für den k-Means-Algor

existieren verschiedene Erweiterungen wie k-Means++ oder der k-Median-Algorith

Ziele der Clusteranalyse mit dem k-Means-Algorithmus

Die Clusteranalyse ist ein Verfahren, mit dem sich eine bestimmte Anzahl von Ob homogene Gruppen einteilen lässt. Ziel ist es, dass die verschiedenen Objekte in Gruppe sich nach der erfolgten Einteilung möglichst ähnlich sind. Die Eigenschaf Objekte sind in Dimensionen eingeteilt. Die Gruppen nennen sich Cluster. Der k-M Algorithmus lässt sich für mehrdimensionale Objekte anwenden und nähert sich ständig wiederholende Neuberechnungen den jeweiligen Clusterzentren an, bis ke signifikante Veränderung mehr stattfindet.

Typischer Ablauf des k-Means-Algorithmus

Der k-Means-Algorithmus findet Anwendung auf Daten oder Objekte in einem n-di Raum. Das Verfahren verläuft in folgenden Schritten:

1. Wahl von k-Punkten als Anfangszentren der Berechnung
2. Zuordnung der Datenpunkte zu den verschiedenen Clustern auf Basis des Abst Zentren
3. Neuberechnung der Clusterzentren
4. Wiederholung ab Schritt 2 – bis sich die Lage der Zentren nicht mehr ändert

In den ersten Durchläufen des Algorithmus treten noch große Änderungen der Ze zunehmenden Schleifendurchläufen werden die Veränderungen immer kleiner. We einen effizienten Ablauf des Algorithmus ist die Wahl der Anfangszentren.

Probleme bei der Anwendung des k-Means-Verfahrens

Ein Problem bei der Anwendung des k-Means-Algorithmus stellt die Vorgabe der Cluster und die Wahl der Anfangszentren dar. Die gefundene Lösung ist stark von gewählten Startpunkten abhängig. Zu Wahl der Startpunkte ist die Kenntnis einer Clusterstruktur notwendig, die aus Voranalysen der Daten stammen könnte. In vie erfolgt der Durchlauf des Algorithmus mit unterschiedlichen Startwerten. Die ver Ergebnisse liefern Hinweise auf eine möglichst plausible Struktur der Cluster. Ver eine ungeeignete Anzahl von Clusterzentren als Startwerte, können sich unter Um komplett andere Lösungen oder ungeeignete Clustereinteilungen ergeben.

Auch problematisch für den k-Means-Algorithmus sind Datenmengen, die sich üb in Teilen nahtlos ineinander übergehen. In diesen Fällen ist das k-Means-Verfahre Lage, die verschiedenen Gruppen zuverlässig voneinander zu trennen. Daten mit Clusterstrukturen werden ebenfalls nicht unterstützt. Sind Ausreißer in der Daten

vorhanden, können diese das Ergebnis stark verfälschen, da k-Means keine Ausre
und jedes Objekt einem Cluster zuordnet. In diesen Fällen ist vor der Auswertung
k-Means-Algorithmus eine Bereinigung der Daten (Noisereduktion) durchzuführer

Die Erweiterung k-Means++

Für den k-Means-Algorithmus existiert die Erweiterung k-Means++. Sie beschreib
Verfahren, mit dem die Clusterzentren als Startwerte nicht mehr zufällig, sondern
Vorschrift gewählt werden. Sind die Clusterzentren als Startwerte bestimmt, konv
anschließend ausgeführte k-Means-Algorithmus sehr schnell und in wenigen
Schleifendurchläufen. Typischerweise lässt sich eine Verdopplung der Geschwinc
erreichen.

Anwendungen des k-Means-Algorithmus

Der k-Means-Algorithmus findet in vielen Bereichen Anwendung. Aufgrund seiner
des geringen Speicher- und Rechenbedarfs eignet er sich für die Datenanalyse gr
Datenmengen im Big-Data-Umfeld. In der Bildverarbeitung wird k-Means häufig zu
Segmentierung der Bilddaten eingesetzt. Mit den Ergebnissen des Algorithmus is
beispielsweise die Trennung von Vorder- und Hintergrund oder das Erkennen von
Objekten möglich.

Ein weiterer wichtiger Anwendungsbereich sind das Marketing und die Analyse d
Kundenverhaltens. k-Means findet in vorliegenden Kundendaten verschiedene Gru
Kunden mit jeweils ähnlichem Kundenverhalten. Die von k-Means ermittelten hom
Gruppen lassen sich klar voneinander trennen. Mithilfe spezifischer Marketingma
die Gruppen effizienter und mit größerer Erfolgsaussicht ansprechbar.

Über den Autor

Stefan Luber

Definition

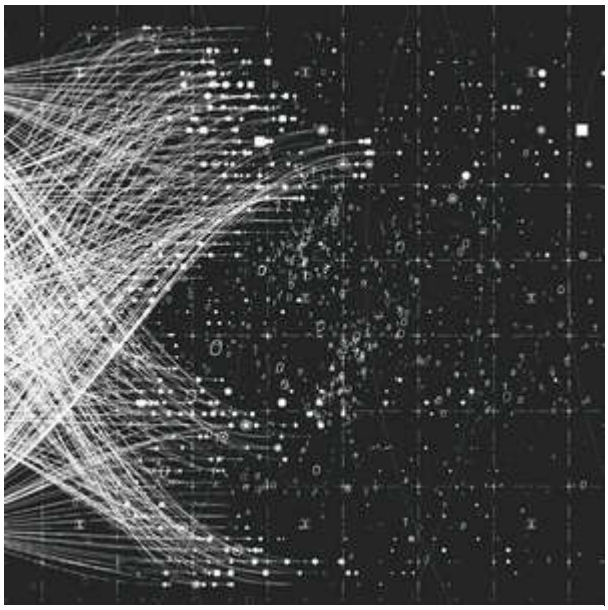
Was ist Customer Experience?

Z

Zur Wahrung unserer Interessen speichern wir zusätzlich zu den o.g. Informationen die IP-Adresse. Dies dient aussc
Zweck, dass Sie als Urheber des Kommentars identifiziert werden können. Rechtliche Grundlage ist die Wahrung ber
gem. Art 6 Abs 1 lit. f) DSGVO.

Kommentieren

AKTUELLE BEITRÄGE ZU DIESEM THEMA

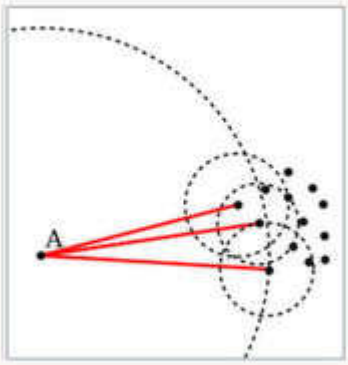


Algorithmen

Analytics-Methoden – von deskriptiven Analysen bis Machine-Learning-Algorithmen

Dem Datenanalysten stehen zahlreiche Methoden zur Verfügung. Der folgende Artikel erläutert einige Methoden – von statistischen, deskriptiven Methoden bis zu Supervised und Unsupervised Machine Learning.

lesen



Grundlagen Statistik & Algorithmen, Teil 7

So deckt der Local Outlier Factor Anomalien auf

Um Trends zu erkennen, wird oft die Clusteranalyse herangezogen. Der k-Means-Algorithmus etwa zeigt an, wo sich Analyseergebnisse in einer Normalverteilung ballen. Für manche Zwecke ist es aber aufschlussreicher, Ausreißer zu untersuchen, denn sie bilden die Antithese zum „Normalen“, etwa im Betrugswesen. Der Local-Outlier-Factor-Algorithmus (LOF) ist in der Lage, den Abstand von Ausreißern zu ihren Nachbarn zu berechnen und deckt so Anomalien auf.

lesen

Gru
Op
Se
Al
Der
sich
Clu
Ber
Spe
die
Um
in d
les

Machine Learning

So kommen Sie mit den passenden Algorithmen zum Ziel

Einfache Algorithmen sind das tägliche Brot vieler Programmierer. Wer etwa eine Software erstellt, die den Preis eines Hauses basierend auf der Größe ausgibt, schreibt normalerweise einen Algorithmus, der abhängig vom Input (Hausgröße) einen bestimmten Output (Preis) berechnet. Wenn es um Machine Learning geht, steigen die Anforderungen an die Programmierung und die Mathematik.

lesen

Kommentar von Lars Milde, Tableau

Top 10 der Business Intelligence Trends für das Jahr 2017

2016 lagen Self-Service-Analysen im Trend. Viele Unternehmen führten den modernen Business-Analytics-Ansatz ein, bei dem IT und Geschäftsbetrieb zusammenarbeiten, um die eigenen Daten optimal zu nutzen. Die IT begann, skalierbare und ausbaufähige Technologien zu nutzen, Geschäftsanwender teilten ihre Daten und arbeiteten gemeinsam daran.

lesen

Lei

Ta

Be

Vor

10.

als

unt

und

les

MEHR ZUM THEMA

Algorithmen

Analytics-Methoden – von deskriptiven Analysen bis Machine-Learning-Algorithmen

mehr...

Data Science

IoT-Basics – die technische Basis von Big Data

mehr...

Definition

Was ist BigTable?

mehr...

Grundlagen Statistik & Algorithmen, Teil 7

So deckt der Local Outlier Factor Anomalien auf

mehr...

PASSENDE WHITEPAPER & WEBCASTS

Diese aktuellen Themen bewegen die Big-Data-Welt

Das BEST OF BigData-Insider

mehr...

Technology-Update für IT-Manager

CIOBRIEFING 12/2018

mehr...

Alle Whitepaper

Alle Webcasts

Oder kontaktieren Sie uns direkt