Woche 09: Wie war das nochmal? – Überblick und Ausblick

Skript

Erarbeitet von
Dr. Maike Mayer

Lernziele 1
Inhalt 1
Einstieg 1
Was schaust und shoppst du gern? 2
Drin, was draufsteht? 3
Auf geht's! 4
Quellen 4
Disclaimer 5

Lernziele

- Erinnern der Definition von Unsupervised Learning
- Nachvollziehen des Grundgedankens der Normalisierung
- Nachvollziehen der Zusammenhänge zwischen den Inhalten

Inhalt

Einstieg

Vor einiger Zeit haben wir schon mal über Algorithmen gesprochen, die Serien vorschlagen. Aber hast du dich auch schon mal gefragt, warum du beispielsweise andere Serien vorgeschlagen bekommst als eine Freundin oder ein Freund? Oder warum du andere Online-Werbung siehst? Das könnte daran liegen, dass dich ein Algorithmus in eine andere Kund*innengruppe einsortiert hat, also in eine andere Klasse.

Projekt „KI für alle"
HeiCAD – Heine Center for Artificial Intelligence and Data Science

CC BY

Einblendung Illustrationen

Ah Moment … Die Sache mit den Klassen sollte dir ein Stück weit bekannt vorkommen. Du kennst bereits den k-nearest neighbours Algorithmus, der Beobachtungen anhand der Klassenzugehörigkeit einer bestimmten Anzahl besonders nah gelegener Nachbarn einsortiert. Und in der letzten Woche hast du Entscheidungsbäume und Entscheidungswälder kennengelernt, mit deren Hilfe man ebenfalls neue Beobachtungen klassifizieren kann.

Einblendung Stichwörter

Und in dieser Woche werden wir uns das Clustering näher anschauen. Eine Methode, die gerne bei der Marktsegmentierung oder der Einteilung von Kund*innengruppen eingesetzt wird …

Was schaust und shoppst du gern?

Das Ziel von Clustering ist es, Daten in vorher unbekannte Gruppen – auch Cluster genannt – aufzuteilen. Dabei sollen die Beobachtungen, die innerhalb eines Clusters liegen, sich möglichst ähnlich sein, während sich die Cluster untereinander möglichst gut voneinander unterscheiden sollen. Die grundlegende Idee erinnert vielleicht ein bisschen an den k-nearest neighbours Ansatz. Allerdings werden die Nachbarschaften hier – also beim Clustering – erst noch ermittelt, wenn man so will. Nachbarn mit ähnlichen Eigenschaften werden zu einer Nachbarschaft zusammengefasst und die Nachbarschaften wiederum sollen dann möglichst verschieden sein.

Einblendung Illustrationen

Wie, die Nachbarschaften werden ermittelt? Ja, richtig gehört. Die Gruppen oder Cluster, in die die Datenmenge eingeteilt wird, stehen am Anfang nicht fest. Das Ziel des Clustering ist es, neue Gruppen bzw. Strukturen in den Daten zu finden. Es geht hier also nicht um Vorhersagen oder Prognosen wie beispielsweise bei der Regression, sondern darum, neues Wissen über vorhandene Daten zu gewinnen. Das klingt nicht so richtig nach Supervised Learning, oder? Stimmt, denn beim Clustering haben wir es mit einer Methode des Unsupervised Learnings zu tun. Zur Erinnerung: Beim Unsupervised Learning – im Gegensatz zum Supervised Learning – gibt der Mensch nicht vor, was gelernt werden soll. Stattdessen soll das System selbst Strukturen und Zusammenhänge (hier eben Gruppen) in den Daten finden.

Einblendung Illustrationen/Stichwörter

Wir stellen dir diese Woche das k-means Clustering vor, das klassischerweise in der Markt- und Kund*innensegmentierung (Stichwort personalisierte Werbung beispielsweise) eingesetzt wird. Mit k wird die Anzahl an Clustern vorgegeben, die wir in den Daten vermuten, beispielsweise, nachdem wir die Daten visualisiert haben. Wie das geht, hast du ja schon in Woche 5 gelernt. Wie das k-means Clustering genau funktioniert, lernst du in dieser Woche.

Einblendung Illustrationen/Stichwort

Wir erläutern dir zunächst die Theorie und dann erfährst du selbstverständlich auch, wie du Clustering Modelle in Python erstellen und damit einen Datensatz in Cluster aufteilen kannst. Du ahnst es bestimmt schon, auch dafür benötigen wir das Modul Scikit-learn. Und dir begegnen auch die Pinguine aus Woche 7 wieder, die wir zur Veranschaulichung des Clustering verwenden. In diesem Zusammenhang erfährst du auch, warum es sinnvoll sein kann, die Daten vor dem Clustering zu normalisieren.

Einblendung Illustrationen/Stichwort

Was Normalisieren ist? Normalisieren hilft dabei, Daten vergleichbar zu machen. Mal ein ganz einfaches Alltagsbeispiel: Nehmen wir an, du möchtest eine bestimmte Reissorte im Laden deines Vertrauens kaufen. Dort wird die Reissorte von drei verschiedenen Marken angeboten und du möchtest gern die günstigste Variante kaufen. Allerdings bietet eine Marke 250 g Reis an, die Nächste 400 g und die Dritte 375 g. Die Menge wirkt sich natürlich auch auf den Preis aus … Die Frage ist, welche Marke ist nun also insgesamt die Günstigste? Dazu ein Tipp: Im Supermarkt werden auf den Preisschildern in der Regel die Preise für 1 kg (oder manchmal auch für 100 g) des Produkts ausgewiesen. Hier kannst du nachschauen, bei welcher Marke 1 kg Reis am günstigsten ist und weißt so, welche Marke du nehmen willst. Im Prinzip hat die Supermarktleitung die Normalisierung für dich übernommen, damit du die Preise trotz verschiedener Mengen des Endproduktes vergleichen kannst.

Einblendung Illustrationen

Gut, nicht immer ist in einer Verpackung oder einem Produkt das drin, was draufsteht, aber das ist eine andere Geschichte …

Drin, was draufsteht?

Ob das drin ist, was draufsteht, ist auch bei KI-Anwendungen unter ethischen Gesichtspunkten relevant und kann problematisch werden, wenn wir es mit Ethics Washing zu tun haben. Ethics Washing bezieht sich auf das Phänomen, dass KI-Unternehmen oder KI-Projekte beispielsweise, sich oft selbst regulieren und eigene Ethik-Leitlinien aufsetzen.

Quelle [1]

Ethik-Leitlinien waren am Rande schon mal Thema, als wir über Transparenz gesprochen haben. Zur Erinnerung: Transparenz bezieht sich auf die Nachvollziehbarkeit der Funktionsweise von KI-Anwendungen und wird oft als Grundsatz in Ethik-Leitlinien zum Thema Künstliche Intelligenz genannt.

Quelle [2,3]

Einblendung Stichwörter/Illustrationen

Das Problem an Leitlinien im Kontext des Ethics Washings, um das es diese Woche geht, ist allerdings, dass sich die Unternehmen nicht immer wirklich an ihre eigenen Leitlinien halten und eine abstrakte Selbstregulierung hier nicht ausreichend ist, sondern konkrete Leitlinien erforderlich sind, die auch umgesetzt werden.

Auf geht's!

In dieser Woche lernst du also mit dem k-means Clustering eine Methode des Unsupervised Learnings kennen. Wir erklären dir, wie diese Methode funktioniert und zeigen dir dann auch, wie du Clustering Modelle in Python umsetzen und so auf Cluster-Suche gehen kannst. Im Bereich der Ethik schauen wir uns in dieser Woche das sogenannte Ethics Washing näher an und zum Abschluss von Woche 9 haben wir ein spannendes Experteninterview zu Anwendungsbeispielen des Clusterings in den Geisteswissenschaften für dich vorbereitet.

Einblendung Videotitel

Nun viel Spaß auf der Cluster-Jagd und in dieser Woche!

Quellen

Quelle [1] Wagner, B. (2019). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens, & M. Hildebrandt (Eds.), *Being profiled* (pp. 84–89). Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Quelle [2] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Quelle [3] Diakopoulos, N. (2020). Transparency. In *The Oxford handbook of ethics of AI* (pp. 197–213). Oxford University Press.

Disclaimer

Transkript zu dem Video „Woche 09: Wie war das nochmal? – Überblick und Ausblick“, Dr. Maike Mayer.

Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz CC-BY 4.0 veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.