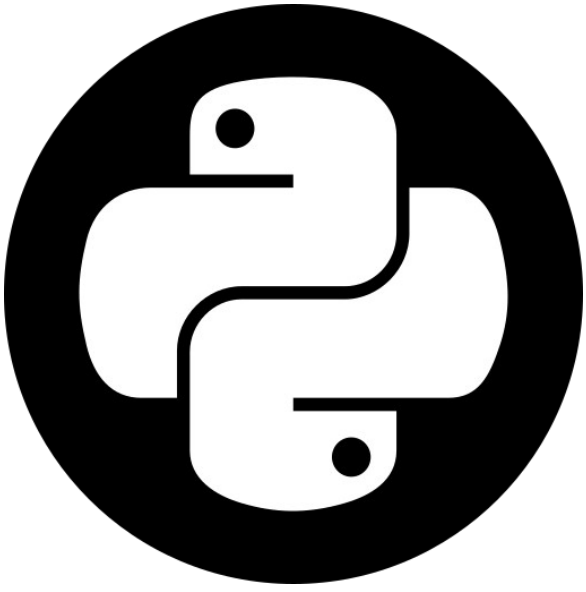
DATA SCIENCE MIT PYTHON

GLOSSAR



www.**STATISTICAL-THINKING**.de

WILLKOMMEN

Herzlich willkommen zum Workshop Data Science mit Python!

Data Science ist ein Bereich der Informatik, der sich mit der Verarbeitung, Analyse und Interpretation großer Datenmengen beschäftigt. In Zeiten von Big Data ist dieser Bereich besonders wichtig, um aus den gesammelten Daten wertvolle Erkenntnisse zu gewinnen und Entscheidungen zu treffen.

In diesem Workshop werden wir uns mit den Grundlagen von Data Science beschäftigen. Data Science ist ein interdisziplinäres Feld, das statistische Methoden, maschinelles Lernen, Datenanalyse und andere Technologien nutzt, um aus Daten Wissen zu ziehen und Einblicke zu gewinnen. Wir werden uns mit den verschiedenen Phasen des Data-Science-Prozesses auseinandersetzen, von der Datenaufbereitung bis zur Modellierung und Interpretation der Ergebnisse.

Darüber hinaus werden verschiedene Tools und Methoden vorgestellt, die in der Datenanalyse verwendet werden. Dazu werden wir praktische Übungen durchführen, um das Gelernte direkt anzuwenden und so ein besseres Verständnis zu erlangen.

Data Science ist insgesamt ein schnell wachsender Bereich, der eine Vielzahl von Karrieremöglichkeiten bietet, zu denen dieser Workshop einen Einstieg ermöglichen soll.

Viel Spaß und Erfolg

Dennis Klinkhammer

DATA SCIENCE

Data Science ist ein multidisziplinärer Ansatz zur Gewinnung von Erkenntnissen und Wissen aus Daten. Es umfasst verschiedene Technologien und Methoden zur Datenerfassung, -verarbeitung, -analyse und -visualisierung, um wertvolle Informationen zu extrahieren, die zur Unterstützung von Entscheidungen und zur Entwicklung von Strategien verwendet werden können.

Data Science umfasst eine Vielzahl von Fähigkeiten und Disziplinen, darunter Statistik, Mathematik, Informatik, Datenbanken, Machine Learning, künstliche Intelligenz, Data Mining, Visualisierung und Big Data, die verschiedene Branchen und Anwendungen bedient, wie beispielsweise das Gesundheitswesen, Finanzen, Marketing, Wissenschaft und Technologie.

In der Praxis beinhaltet Data Science typischerweise den Prozess der Datenerfassung und -bereinigung, die Anwendung von statistischen Methoden und Machine-Learning-Techniken, um Muster und Trends zu identifizieren, und schließlich die Kommunikation der Ergebnisse in Form von Berichten, Visualisierungen und anderen Formen von Erkenntnissen.

Data Science hat einen erheblichen Einfluss auf die Entscheidungsfindung in verschiedenen Bereichen und wird in vielen Unternehmen und Organisationen eingesetzt, um bessere Entscheidungen zu treffen, Prozesse zu optimieren, Risiken zu minimieren und Wettbewerbsvorteile zu erzielen.

BIG DATA

Big Data bezieht sich auf große und komplexe Datensätze, die von Unternehmen oder Organisationen gesammelt werden, um Muster, Trends und Erkenntnisse zu identifizieren. Diese Daten können strukturiert oder unstrukturiert sein, aus verschiedenen Quellen stammen und in unterschiedlichen Formaten vorliegen.

Beispiele für Big Data in der Versicherungsbranche können sein:

- Kundendaten: Versicherungsunternehmen sammeln große Mengen an Kundendaten, die von der Anzahl der Versicherungsverträge und -prämien bis hin zu Schadensmeldungen und Vertragskündigungen reichen. Mit diesen Daten können Unternehmen Verhaltensmuster und Trends identifizieren, um personalisierte Angebote und Dienstleistungen zu entwickeln.
- Risikoanalyse: Versicherungsunternehmen nutzen Big Data, um Risikomodelle zu entwickeln, die die Wahrscheinlichkeit von Schäden und Verlusten vorhersagen und so die Tarifgestaltung und die Unternehmensstrategie optimieren können.

Beispiele für Big Data im Gesundheitswesen können sein:

- Elektronische Patientenakten: Krankenhäuser und Arztpraxen speichern große Mengen an Patientendaten, einschließlich medizinischer Diagnosen, Symptome, Behandlungen und Arzneimittel. Mit diesen Daten können Ärzte und Forscher Krankheitsmuster und Behandlungsoptionen identifizieren, um bessere Ergebnisse für Patienten zu erzielen.
- Genomik: Die Genomsequenzierung setzt auf Big Data, da jeder Mensch etwa drei Milliarden Basenpaare in seinem Genom hat. Diese Daten können genutzt werden, um genetische Anomalien zu identifizieren, die mit bestimmten Krankheiten oder Erkrankungen zusammenhängen, und personalisierte Behandlungen zu entwickeln.

In beiden Branchen können Big Data-Technologien und Methoden, wie bspw. das maschinelle Lernen, eingesetzt werden, um diese Daten zu analysieren und nützliche Einblicke und Erkenntnisse zu gewinnen.

PYTHON

Python ist eine interpretierte, objektorientierte und höhere Programmiersprache, die für eine Vielzahl von Anwendungen wie Webentwicklung, wissenschaftliche Berechnungen, künstliche Intelligenz und Datenanalyse verwendet wird. Im Folgenden sind die Funktionsweise und einige Vorteile von Python beschrieben:

Funktionsweise von Python:

- Einfach zu lesen und zu schreiben: Python ist eine benutzerfreundliche Programmiersprache, die eine klare und gut strukturierte Syntax hat. Dies erleichtert das Lesen und Schreiben von Code und ermöglicht es Entwicklern, schnell und effizient zu arbeiten.
- Interpreter-basiert: Python verwendet einen Interpreter, um den Code direkt in ausführbaren Code umzuwandeln. Dadurch entfällt der komplizierte Kompilierungs-prozess und es wird Zeit und Mühe gespart.
- Plattformunabhängig: Python kann auf verschiedenen Betriebssystemen wie Windows, Linux und Mac ausgeführt werden, was es Entwicklern ermöglicht, plattformunabhängige Anwendungen zu erstellen.
- Umfangreiche Standardbibliothek: Python verfügt über eine umfangreiche Standardbibliothek, die eine Vielzahl von Funktionen und Modulen enthält, die Entwickler in ihren Projekten nutzen können. Dadurch können Entwickler vorhandene Funktionen einfach anpassen und wiederverwenden.

Vorteile von Python:

- Einfach zu erlernen und zu verwenden: Python ist eine einfach zu erlernende Programmiersprache, die auch für Anfänger zugänglich ist. Die klare Syntax und die umfangreiche Dokumentation erleichtern den Einstieg.
- Große Entwicklergemeinschaft: Python hat eine große und aktive Entwicklergemeinschaft, die kontinuierlich neue Bibliotheken und Frameworks entwickelt und pflegt. Dies erleichtert die Arbeit von Entwicklern, da sie auf vorhandene Lösungen zurückgreifen können.
- Vielseitigkeit: Python kann für eine Vielzahl von Anwendungen verwendet werden, einschließlich Webentwicklung, Datenanalyse, künstliche Intelligenz und Spieleentwicklung. Dadurch wird Python zu einer universellen Programmiersprache.
- Skripting-Fähigkeit: Python kann als Skriptsprache verwendet werden, um die Automatisierung von Aufgaben und schnelle Prototypenentwicklung zu erleichtern.

DATENSTRUKTUREN IN PYTHON

In Python gibt es verschiedene Datenstrukturen, die für die Speicherung und Verarbeitung von Daten verwendet werden können. Hier sind einige der häufigsten Datenstrukturen in Python:

- Liste: Eine Liste ist eine geordnete Sammlung von Elementen, die in eckigen Klammern [] notiert werden. Listen können Elemente desselben oder unterschiedlichen Typs enthalten und sind veränderlich (d.h. ihre Elemente können hinzugefügt, gelöscht oder geändert werden).
- Tupel: Ein Tupel ist ähnlich wie eine Liste, jedoch unveränderlich (d.h. ihre Elemente können nicht verändert werden). Es wird in runden Klammern () notiert.
- Dictionaries: Ein Dictionary ist eine ungeordnete Sammlung von Schlüssel-Wert-Paaren, die in geschweiften Klammern {} notiert werden. Dictionaries ermöglichen eine schnelle Suche nach einem bestimmten Schlüssel und sind veränderlich.
- Sets: Ein Set ist eine ungeordnete Sammlung von einzigartigen Elementen, die in geschweiften Klammern {} notiert werden. Sets bieten schnelle Mitgliedschaftstests und sind veränderlich.
- Arrays: Ein Array ist eine Sammlung von Elementen desselben Typs, die in einer Numpy-Bibliothek implementiert sind. Arrays ermöglichen eine effiziente Berechnung von mathematischen Operationen.
- DataFrames: Ein DataFrame (DF) ist eine tabellarische Datenstruktur, die in der Pandas-Bibliothek implementiert ist. DataFrames ermöglichen die Speicherung von Daten in einer Spalten- und Zeilenstruktur und bieten zahlreiche Funktionen zur Datenaufbereitung und -analyse.

JUPYTER NOTEBOOKS

Jupyter Notebooks sind eine webbasierte interaktive Entwicklungsumgebung für Data Science und Datenanalyse. Sie bieten eine flexible und intuitive Möglichkeit, Code, Text und Visualisierungen in einem Dokument zu kombinieren.

Die wichtigsten Vorteile von Jupyter Notebooks sind:

1. Interaktive Datenanalyse: Jupyter Notebooks ermöglichen es Benutzern, Code-Zellen auszuführen und Ergebnisse sofort zu sehen, was ein effektives Mittel für interaktive Datenanalyse und -exploration darstellt.
2. Einfache Integration: Jupyter Notebooks können mit vielen verschiedenen Programmiersprachen wie Python, R, Julia und anderen verwendet werden und bieten eine nahtlose Integration mit anderen Bibliotheken und Frameworks, die in diesen Sprachen verfügbar sind.
3. Dokumentation und Kommunikation: Jupyter Notebooks sind eine effektive Möglichkeit, komplexe Datenanalysen zu dokumentieren und zu kommunizieren. Sie ermöglichen es Benutzern, Code, Text, Diagramme und Visualisierungen in einem Dokument zu kombinieren, das für andere Benutzer leicht zugänglich ist.
4. Wiederverwendbarkeit: Jupyter Notebooks können leicht wiederverwendet werden, indem sie als Vorlagen für ähnliche Aufgaben oder Analysen dienen.
5. Open Source: Jupyter Notebooks sind kostenlos und Open Source, was bedeutet, dass sie von einer großen Community von Entwicklern unterstützt werden und ständig weiterentwickelt und verbessert werden.

Insgesamt bieten Jupyter Notebooks eine effektive Möglichkeit, Daten zu analysieren, zu dokumentieren und zu kommunizieren, was sie zu einem wichtigen Werkzeug für Data Science und Datenanalyse macht.

DOCKER

Docker ist eine Open-Source-Plattform, die es ermöglicht, Anwendungen und Dienste in Containern zu verpacken und bereitzustellen. Im Kontext der Programmiersprache Python kann Docker verwendet werden, um Python-basierte Anwendungen in isolierten Umgebungen auszuführen.

Ein Vorteil von Docker im Python-Kontext besteht darin, dass Entwickler eine vollständige Entwicklungs- und Testumgebung erstellen können, die alle erforderlichen Abhängigkeiten und Bibliotheken enthält. Dadurch können Entwickler vermeiden, dass sie lokal Abhängigkeiten installieren müssen, was oft zu Konflikten zwischen verschiedenen Versionen von Bibliotheken führen kann.

Darüber hinaus kann Docker genutzt werden, um Python-basierte Anwendungen in einer konsistenten Umgebung bereitzustellen und skalieren zu können. Entwickler können eine Docker-Image-Datei erstellen, die die Anwendung und alle benötigten Abhängigkeiten enthält, und diese Image-Datei auf verschiedenen Servern oder in der Cloud ausführen.

Docker kann auch zur Integration von Python-Anwendungen mit anderen Anwendungen und Diensten verwendet werden, indem es ermöglicht, dass diese in separate Container verpackt und bereitgestellt werden. Dies kann die Entwicklung und Bereitstellung von Anwendungen vereinfachen und beschleunigen, insbesondere in einer Cloud-basierten Umgebung.

KUBERNETES

Kubernetes ist ebenfalls ein Open-Source-System zur Automatisierung der Bereitstellung, Skalierung und Verwaltung von Container-Anwendungen. Im Kontext der Programmiersprache Python kann Kubernetes verwendet werden, um Python-basierte Anwendungen zu orchestrieren und in einer Container-basierten Umgebung bereitzustellen.

Kubernetes bietet mehrere Vorteile im Python-Kontext. Zunächst einmal ermöglicht es Kubernetes Entwicklern, Python-basierte Anwendungen effektiv zu skalieren und zu verwalten. Wenn die Anforderungen an die Anwendung steigen, können Entwickler die Anzahl der Container, die die Anwendung ausführen, automatisch erhöhen, um die Last zu bewältigen.

Darüber hinaus bietet Kubernetes eine Möglichkeit zur Integration von Python-basierten Anwendungen mit anderen Anwendungen und Diensten. Entwickler können beispielsweise einen Kubernetes-Cluster erstellen, der Python-basierte Anwendungen zusammen mit anderen Anwendungen in separaten Containern ausführt. Dadurch können Entwickler eine effektive Multi-Service-Architektur erstellen, die einfach zu skalieren und zu verwalten ist.

Ein weiterer Vorteil von Kubernetes im Python-Kontext besteht darin, dass es Entwicklern eine effektive Möglichkeit zur Verwaltung von Anwendungsabhängigkeiten und Anwendungskonfigurationen bietet. Entwickler können Kubernetes-Objekte wie ConfigMaps und Secrets verwenden, um Konfigurationsdaten und geheime Informationen zu verwalten, die von Python-basierten Anwendungen verwendet werden.

GIT(HUB)

Git ist ein Versionskontrollsystem, das von Entwicklern verwendet wird, um Code-Änderungen zu verwalten und gemeinsam an einem Projekt zu arbeiten. Die Arbeit mit Git beinhaltet in der Regel die folgenden Schritte:

1. Einrichten von Git: Zunächst muss Git auf dem lokalen Computer installiert und konfiguriert werden. Dazu gehört das Erstellen eines Git-Repositorys für das Projekt.
2. Änderungen vornehmen: Sobald das Repository eingerichtet ist, können Entwickler Änderungen am Code vornehmen. Diese Änderungen können entweder lokal oder auf einem Remote-Repository erfolgen.
3. Committen von Änderungen: Wenn Entwickler Änderungen am Code vorgenommen haben, müssen sie diese committen. Dabei werden die Änderungen im lokalen Repository gespeichert und mit einer kurzen Beschreibung versehen.
4. Pushen von Änderungen: Wenn Entwickler Änderungen an einem Remote-Repository vornehmen, müssen sie diese pushen, damit sie für andere Entwickler verfügbar sind.
5. Pullen von Änderungen: Wenn andere Entwickler Änderungen am Remote-Repository vorgenommen haben, müssen Entwickler diese Änderungen pullen, um ihre lokale Kopie des Repositorys auf den neuesten Stand zu bringen.

Einige Vorteile der Arbeit mit Git sind:

1. Versionskontrolle: Git bietet eine zuverlässige Methode zur Versionskontrolle. Entwickler können den Code-Verlauf verfolgen, Änderungen nachverfolgen und zu früheren Versionen zurückkehren.
2. Zusammenarbeit: Git ermöglicht es Entwicklern, gemeinsam an einem Projekt zu arbeiten, ohne sich in die Quere zu kommen. Es stellt sicher, dass Änderungen von verschiedenen Entwicklern korrekt zusammengeführt werden.
3. Backup und Wiederherstellung: Git bietet eine effektive Methode zur Sicherung des Codes und zur Wiederherstellung des Codes im Falle eines Systemfehlers.
4. Verzweigungen: Git ermöglicht es Entwicklern, Verzweigungen im Code zu erstellen, um verschiedene Versionen des Codes zu testen und zu entwickeln.

GitHub ist die onlinebasierte Variante des Versionskontrollsystems und wird von vielen Entwicklern weltweit verwendet.

C R I S P – D M

CRISP-DM steht für Cross-Industry Standard Process for Data Mining und ist ein bewährtes Framework für den Data-Mining-Prozess. Es ist eine strukturierte Methode, die Data-Mining-Projekte in sechs Schritte gliedert, die von der Planung bis zur Implementierung reichen. Die sechs Schritte sind:

1. Business-Verständnis: In diesem Schritt wird das Problem definiert, das gelöst werden soll, und das Ziel des Data-Mining-Projekts wird festgelegt. Es wird auch eine erste Bewertung der Kosten, Ressourcen und Zeitpläne für das Projekt durchgeführt.
2. Datenverständnis: In diesem Schritt werden die Daten, die für das Projekt benötigt werden, gesammelt und untersucht. Es wird eine erste Datenbewertung durchgeführt, um zu bestimmen, ob die Daten für das Projekt geeignet sind, und um ein besseres Verständnis für die Daten zu gewinnen.
3. Datenvorbereitung: In diesem Schritt werden die Daten für die Analyse vorbereitet, indem sie gereinigt, integriert und transformiert werden. Die Daten werden in einem Format präsentiert, das für die Analyse geeignet ist.
4. Modellierung: In diesem Schritt werden verschiedene Modelle erstellt, um das Problem zu lösen. Es wird eine Modellauswahl durchgeführt, um das beste Modell für das Problem zu finden. Das Modell wird dann auf den Daten trainiert, um Vorhersagen zu treffen.
5. Bewertung: In diesem Schritt wird das Modell auf einer unabhängigen Stichprobe von Daten getestet, um die Genauigkeit und die Leistung des Modells zu bewerten. Es wird auch eine Bewertung des Modells gegen das Ziel des Projekts durchgeführt.
6. Implementierung: In diesem Schritt wird das Modell in die Produktionsumgebung integriert, um Vorhersagen für neue Daten zu treffen. Es wird auch eine Überwachung des Modells durchgeführt, um sicherzustellen, dass es weiterhin richtig funktioniert.

CRISP-DM ist ein iterativer Prozess, der es dem Team ermöglicht, durch die Schritte zu gehen und bei Bedarf Schritte zu wiederholen oder zurückzukehren. Das Framework hilft, ein strukturiertes Vorgehen für Data-Mining-Projekte bereitzustellen und sicherzustellen, dass sie erfolgreich durchgeführt werden.

N U M P Y

NumPy ist eine Python-Bibliothek für numerische Berechnungen, die grundlegende Funktionen für die Verarbeitung von Arrays und Matrizen bereitstellt.

Numpy bietet eine Vielzahl von Funktionen und Operationen für mathematische Operationen auf Arrays und Matrizen, einschließlich lineare Algebra, Fourier-Transformation, Zufallszahlengenerierung und mehr. Mit NumPy können komplexe numerische Berechnungen in Python durchgeführt werden, was besonders nützlich ist, wenn man mit großen Datenmengen arbeitet.

Zusammen mit anderen Bibliotheken wie Pandas, Scikit-learn und Matplotlib ist NumPy ein wesentlicher Bestandteil des wissenschaftlichen Python-Stacks und wird in vielen Bereichen wie der Datenanalyse, maschinellen Lernens, Bildverarbeitung, Signalverarbeitung und Simulationen verwendet.

PANDAS

Pandas ist eine leistungsstarke und flexible Python-Bibliothek zur Datenmanipulation und -analyse. Die Funktionsweise von Pandas basiert auf zwei grundlegenden Datenstrukturen: Series und DataFrames.

Eine Series ist eine eindimensionale Datenstruktur, die eine Liste von Werten und einem zugehörigen Label-Array enthält. Ein DataFrame ist eine zweidimensionale Datenstruktur, die mehrere Series enthält, die zusammen eine Tabelle darstellen.

Pandas bietet eine Vielzahl von Funktionen, um Daten zu importieren, exportieren, transformieren, filtern und aggregieren. Hier sind einige der wichtigsten Funktionen:

- Datenimport: Pandas kann Daten aus verschiedenen Quellen wie CSV-, Excel-, SQL-, JSON-, HTML- und Textdateien importieren.
- Datenmanipulation: Pandas bietet Funktionen, um Daten zu filtern, sortieren, zusammenzuführen, transformieren und umzubenennen. Die Funktionen sind ähnlich wie in SQL.
- Datenanalyse: Pandas bietet Funktionen, um Statistiken wie Mittelwerte, Varianzen, Korrelationen und Regressionen zu berechnen. Auch gibt es Funktionen um fehlende Werte zu handhaben, wie sie imputiert werden können und wie man Daten visuell darstellt.
- Datenexport: Pandas kann Daten in verschiedenen Formaten exportieren, wie CSV-, Excel-, SQL-, JSON-, HTML- und Textdateien.

Pandas ist eine äußerst nützliche Bibliothek für die Datenanalyse, insbesondere für diejenigen, die mit tabellarischen Daten arbeiten. Es ist ein wichtiger Bestandteil des wissenschaftlichen Python-Stacks.

SCIKIT–LEARN

Scikit-learn ist eine Python-Bibliothek für maschinelles Lernen, die eine Vielzahl von Funktionen und Algorithmen für die Analyse und Vorhersage von Daten bereitstellt. Die Bibliothek ist in Python geschrieben und verwendet andere Bibliotheken wie NumPy und Pandas, um effizientere Berechnungen durchzuführen.

Scikit-learn enthält viele wichtige Funktionen für die Datenanalyse und das maschinelle Lernen, einschließlich:

- Datenvorverarbeitung: Scikit-learn bietet eine Vielzahl von Funktionen für die Vorverarbeitung von Daten, einschließlich der Skalierung, Normalisierung, Reduzierung von Dimensionen und der Handhabung von fehlenden Werten.
- Modellauswahl: Scikit-learn bietet Funktionen für die Auswahl des besten Modells für die gegebene Datenmenge. Dazu gehören Funktionen wie die Kreuzvalidierung und die Grid-Suche.
- Überwachtes Lernen: Scikit-learn bietet eine Vielzahl von Algorithmen für das überwachte Lernen, einschließlich der Klassifikation, Regressions- und Entscheidungsbaum-Modelle.
- Unüberwachtes Lernen: Scikit-learn bietet auch Algorithmen für das unüberwachte Lernen, einschließlich der Clustering- und Dimensionsreduzierungsmethoden.
- Modellbewertung: Scikit-learn bietet Funktionen zur Bewertung der Leistung von Modellen anhand von Metriken wie Genauigkeit, Präzision, Recall und F1-Score.

Scikit-learn ist eine der am häufigsten verwendeten Bibliotheken für maschinelles Lernen in Python.

MATPLOTLIB

Matplotlib ist eine Python-Bibliothek zur Erstellung von grafischen Darstellungen und Diagrammen. Die Bibliothek bietet eine Vielzahl von Funktionen und Tools zur Erstellung von hochwertigen visuellen Darstellungen von Daten.

Matplotlib kann verwendet werden, um verschiedene Arten von Diagrammen wie Linien-, Balken-, Flächen-, Punktwolken- und Konturdiagramme zu erstellen. Die Bibliothek kann auch verwendet werden, um mehrere Diagramme in einem einzigen Diagramm anzuzeigen und um benutzerdefinierte Farben, Achsenbeschriftungen, Titel und Legenden hinzuzufügen.

Die Funktionsweise von Matplotlib basiert auf der Verwendung von Python-Objekten zur Darstellung von Diagrammen. Matplotlib verwendet dabei ein "Layering"-Modell: Ein Diagramm wird als eine Reihe von Ebenen erstellt, von der Hintergrundebene bis zur Vordergrundebene. Jede Ebene kann Elemente wie Achsen, Titel, Diagrammlinien, Marker, Legenden usw. enthalten.

Matplotlib kann auch mit anderen Bibliotheken wie NumPy und Pandas verwendet werden, um Daten zu visualisieren und zu analysieren. Matplotlib bietet eine hohe Flexibilität bei der Erstellung von Diagrammen, sodass Benutzer eine Vielzahl von Anpassungen vornehmen und das Aussehen und Verhalten von Diagrammen genau steuern können.

Insgesamt ist Matplotlib eine unverzichtbare Bibliothek für jeden, der mit Datenvisualisierung in Python arbeitet, und es ist eine der am häufigsten verwendeten Bibliotheken für wissenschaftliche Grafiken.

DATA MINING UND DATA CRAWLING

Data Crawling und Data Mining sind unterschiedliche Methoden und beziehen sich auf die Gewinnung von Daten aus verschiedenen Quellen und die Extraktion von Information aus diesen Daten.

Data Crawling bezieht sich auf den Prozess der automatischen Extraktion von Daten aus dem Internet oder anderen öffentlich zugänglichen Quellen. Dies wird oft von Suchmaschinen durchgeführt, um Inhalte auf Websites zu indizieren und Suchergebnisse bereitzustellen. Beim Crawling werden Webseiten durchsucht und Daten wie Texte, Bilder und Videos extrahiert. Das Ziel des Crawling besteht darin, möglichst viele Daten zu sammeln, ohne bestimmte Parameter oder Merkmale im Voraus festzulegen.

Data Mining hingegen bezieht sich auf den Prozess der Analyse von Daten, um Muster, Zusammenhänge und Erkenntnisse zu identifizieren. Die Daten stammen oft aus verschiedenen Quellen wie Datenbanken, Tabellenkalkulationen oder Textdateien. Beim Data Mining werden spezielle Algorithmen verwendet, um die Daten zu analysieren und Muster zu finden. Das Ziel des Data Mining besteht bspw. darin, Erkenntnisse und Trends zu identifizieren, die in den Daten verborgen sind, und diese für Geschäftsentscheidungen oder andere Zwecke zu nutzen.

Zusammenfassend lässt sich sagen, dass Data Crawling das Sammeln von Daten aus dem Internet oder anderen öffentlich zugänglichen Quellen bezeichnet, während Data Mining die Analyse von Daten zur Identifizierung von Mustern und Trends umfasst.

STICHPROBE UND GRUNDGESAMTHEIT

In der Statistik wird zwischen einer Stichprobe und einer Grundgesamtheit unterschieden.

Die Grundgesamtheit ist die Gesamtheit aller Elemente, die eine bestimmte Eigenschaft oder Merkmal haben und die für die Fragestellung relevant sind. Beispiele für Grundgesamtheiten können sein: alle Menschen in Deutschland, alle Autos eines bestimmten Herstellers oder alle Bäume in einem bestimmten Wald.

Eine Stichprobe hingegen ist eine Teilmenge der Grundgesamtheit, die ausgewählt wurde, um Schlussfolgerungen über die Grundgesamtheit zu ziehen. Die Stichprobe sollte repräsentativ für die Grundgesamtheit sein, d.h. sie sollte die gleichen Eigenschaften oder Merkmale wie die Grundgesamtheit aufweisen. Eine repräsentative Stichprobe ermöglicht es, genaue Aussagen über die Grundgesamtheit zu machen, ohne alle Elemente der Grundgesamtheit untersuchen zu müssen.

Es gibt verschiedene Arten von Stichproben, wie zum Beispiel die Zufallsstichprobe, bei der jedes Element der Population die gleiche Chance hat, in die Stichprobe aufgenommen zu werden.

Das Konzept von Stichprobe und Grundgesamtheit ist für die Statistik und die empirische Forschung von zentraler Bedeutung, da es ermöglicht, Schlüsse auf die Grundgesamtheit zu ziehen, ohne jedes Element der Grundgesamtheit untersuchen zu müssen.

KONFIDENZNIVEAU UND KONFIDENZINTERVALL

Das Konfidenzniveau ist ein Begriff aus der Statistik und gibt an, mit welcher Wahrscheinlichkeit ein geschätzter Parameter (bspw. das arithmetische Mittel einer Grundgesamtheit) in einem bestimmten Intervall liegt.

Ein Konfidenzniveau wird oft als Prozentsatz angegeben und liegt typischerweise zwischen 90% und 99%. Ein Konfidenzniveau von 95% bedeutet beispielsweise, dass bei wiederholtem Durchführen (bspw. eines Experiments) mit der gleichen Stichprobe in 95% der Fälle das Intervall den wahren Wert des Parameters enthält.

Das Konfidenzniveau hängt eng mit dem Konfidenzintervall zusammen, das das Intervall um den geschätzten Parameter herum angibt. Ein Konfidenzintervall mit einem Konfidenzniveau von 95% bedeutet, dass das Intervall in 95% der Fälle den wahren Wert des Parameters enthalten sollte.

Das Konfidenzniveau ist ein wichtiger Faktor bei der Interpretation von statistischen Ergebnissen und der Beurteilung der Zuverlässigkeit von Schätzungen. Ein höheres Konfidenzniveau bedeutet in der Regel ein breiteres Konfidenzintervall, da die Wahrscheinlichkeit, dass das Intervall den wahren Wert des Parameters enthält, höher sein muss.

FEHLERMARGE

Die Fehlermarge, auch als Fehlergrenze oder Konfidenzintervallbreite bezeichnet, ist ein Maß dafür, wie weit die Schätzungen eines Parameters in einer Stichprobe vom wahren Wert des Parameters in der Grundgesamtheit abweichen können.

Die Fehlermarge wird oft in Verbindung mit einem Konfidenzintervall angegeben, das angibt, wie sicher man sich sein kann, dass der wahre Wert des Parameters in einem bestimmten Intervall liegt. Die Fehlermarge ist dann die Differenz zwischen den oberen und unteren Grenzen des Konfidenzintervalls.

Eine breitere Fehlermarge bedeutet, dass die Schätzungen in der Stichprobe ungenauer sind und der wahre Wert des Parameters in einem größeren Bereich liegen kann. Eine schmalere Fehlermarge bedeutet hingegen, dass die Schätzungen genauer sind und der wahre Wert des Parameters mit höherer Wahrscheinlichkeit in einem kleineren Bereich liegt.

Die Fehlermarge hängt von verschiedenen Faktoren ab, wie bspw. der Größe der Stichprobe, dem Konfidenzniveau und der Standardabweichung der Daten. Eine größere Stichprobe oder ein höheres Konfidenzniveau führen zu einer kleineren Fehlermarge, während eine größere Standardabweichung zu einer größeren Fehlermarge führt.

Die Fehlermarge ist ein wichtiger Faktor bei der Interpretation von statistischen Ergebnissen und der Beurteilung der Genauigkeit von Schätzungen. Eine Angabe der Fehlermarge zusammen mit einer Schätzung gibt an, wie weit die Schätzungen in der Stichprobe vom wahren Wert des Parameters abweichen können und hilft bei der Beurteilung der Zuverlässigkeit der Schätzungen.

MODUS, MEDIAN UND ARITHMETISCHES MITTEL

Die Lagemaße Modus, Median und arithmetisches Mittel werden in der Statistik verwendet, um verschiedene Aspekte der Verteilung einer Stichprobe oder einer Population zu beschreiben.

Der Modus ist der Wert, der in einer Stichprobe oder einer Population am häufigsten vorkommt. Mit anderen Worten, es ist der Wert, der am meisten "modisch" ist. Wenn es mehrere Werte gibt, die gleich häufig vorkommen, dann hat die Verteilung mehrere Modi. Der Modus ist besonders nützlich, wenn man mit nominalen Daten arbeitet, bei denen es um die Häufigkeit des Auftretens von Kategorien geht.

Der Median ist der Wert, der in der Mitte einer sortierten Stichprobe oder einer Population liegt. Das bedeutet, dass 50% der Daten kleiner als der Median und 50% der Daten größer als der Median sind. Der Median ist besonders nützlich, wenn man mit ordinalen oder intervallskalierten Daten arbeitet, bei denen die Rangfolge der Daten wichtig ist.

Das arithmetische Mittel oder der Durchschnitt ist der Summenwert aller Daten in einer Stichprobe oder Population, geteilt durch die Anzahl der Datenpunkte. Es gibt einen Hinweis darauf, wie die Daten insgesamt verteilt sind. Das arithmetische Mittel ist besonders nützlich, wenn man mit intervall- oder ratioskalierten Daten arbeitet, bei denen die Differenz zwischen den Datenpunkten wichtig ist.

Es ist wichtig zu beachten, dass jedes dieser Lagemaße Vor- und Nachteile hat und dass die Wahl des geeigneten Lagemaßes von der Art der Daten und der Forschungsfrage abhängt.

VARIANZ UND KOVARIANZ

Die Varianz ist ein statistisches Maß für die Streuung von Datenpunkten um den Mittelwert. Sie gibt an, wie weit die einzelnen Werte von ihrem Durchschnitt entfernt sind und wie stark die Daten um den Mittelwert herum variieren. Je größer die Varianz ist, desto weiter entfernt sind die einzelnen Werte vom Mittelwert und desto heterogener sind die Daten insgesamt.

Sie wird berechnet, indem man für jeden Datenpunkt die Abweichung vom Mittelwert quadriert, alle quadrierten Abweichungen aufsummiert und diese Summe dann durch die Anzahl der Datenpunkte teilt.

Insgesamt ist die Varianz ein wichtiges Konzept in der Statistik und wird in vielen Bereichen eingesetzt, um die Streuung von Datenpunkten zu messen. Sie wird oft zusammen mit dem Mittelwert verwendet, um ein umfassendes Verständnis von Daten und deren Verteilung zu erhalten. Je nach Anwendungsbereich gibt es verschiedene Varianten der Varianz, wie beispielsweise die Stichprobenvarianz oder die Populationsvarianz.

Die Kovarianz ermöglicht es dabei den Zusammenhang zwischen Variablen zu untersuchen. Es ist jedoch wichtig zu beachten, dass die Kovarianz nicht standardisiert ist und somit schwer zu vergleichen ist. Eine Möglichkeit, den Zusammenhang zwischen Variablen zu vergleichen, ist die Verwendung des Korrelationskoeffizienten, der die Kovarianz durch das Produkt der Standardabweichungen der Variablen dividiert.

S T A N D A R D A B W E I C H U N G

Die Standardabweichung ist ein statistisches Maß für die Streuung von Datenpunkten um den Mittelwert. Sie gibt an, wie weit die einzelnen Werte von ihrem Durchschnitt entfernt sind und wie stark die Daten um den Mittelwert herum variieren. Die Standardabweichung ist einfach die Quadratwurzel der Varianz, und sie wird oft als eine Maßeinheit verwendet, um zu beschreiben, wie eng oder breit eine Verteilung ist. Dazu wird einfach die Wurzel aus der Varianz gezogen.

Es handelt sich um ein für viele Formeln und Algorithmen wichtiges Maß in der Statistik, um die Streuung von Datenpunkten zu messen. Sie wird oft zusammen mit dem Mittelwert verwendet, um ein umfassendes Verständnis von Daten und deren Verteilung zu erhalten. Je nach Anwendungsbereich gibt es verschiedene Varianten der Standardabweichung, wie beispielsweise die Stichprobenstandardabweichung oder die Populationsstandardabweichung.

KORRELATION

Korrelation ist ein statistisches Konzept, das misst, wie eng zwei Variablen miteinander zusammenhängen. Eine Korrelation von + 1.00 bedeutet, dass die beiden Variablen perfekt positiv korreliert sind, während eine Korrelation von - 1.00 bedeutet, dass sie perfekt negativ korreliert sind. Eine Korrelation von 0.00 bedeutet, dass es keinen linearen Zusammenhang zwischen den Variablen gibt.

Die Funktionsweise der Korrelation kann wie folgt beschrieben werden:

1. Datenerfassung: Zunächst werden Daten für die beiden Variablen erfasst, die untersucht werden sollen. Die Daten können durch Umfragen, Experimente, Beobachtungen oder andere Methoden gewonnen werden.
2. Berechnung des Korrelationskoeffizienten: Der Korrelationskoeffizient ist ein statistisches Maß für die Stärke und Richtung des Zusammenhangs zwischen den Variablen. Der am häufigsten verwendete Korrelationskoeffizient ist der Pearson-Korrelationskoeffizient, der für kontinuierliche Variablen verwendet wird. Der Pearson-Korrelationskoeffizient wird berechnet, indem die Kovarianz der beiden Variablen durch das Produkt ihrer Standardabweichungen dividiert wird. Andere Korrelationskoeffizienten wie der Spearman-Rangkorrelationskoeffizient oder der Kendall-Rankkorrelationskoeffizient werden für nicht kontinuierliche Variablen verwendet.
3. Überprüfung der Signifikanz: Es ist wichtig zu überprüfen, ob die Korrelation statistisch signifikant ist. Dies kann durch Berechnung eines p-Wertes erfolgen, der angibt, wie wahrscheinlich es ist, dass die beobachtete Korrelation aufgrund des Zufalls auftritt. Wenn der p-Wert klein genug ist (in der Regel kleiner als 0.05), kann die Korrelation als statistisch signifikant angesehen werden.

LINEARE REGRESSION

Lineare Regression ist eine statistische Methode, die verwendet wird, um die Beziehung zwischen einer abhängigen Variablen (Y) und einer oder mehreren unabhängigen Variablen (X) zu modellieren. Die grundlegende Idee der linearen Regression besteht darin, eine Linie durch die Punkte in einem Streudiagramm zu ziehen, die die Beziehung zwischen den Variablen am besten beschreibt. Diese Linie wird als Regressionsgerade bezeichnet und kann verwendet werden, um Vorhersagen über die abhängige Variable zu treffen, wenn die unabhängigen Variablen bekannt sind.

Die Funktionsweise der linearen Regression lässt sich wie folgt beschreiben:

1. Zunächst werden die Daten gesammelt und in einem Streudiagramm dargestellt. Die abhängige Variable wird auf der vertikalen Achse (y-Achse) und die unabhängigen Variablen auf der horizontalen Achse (x-Achse) dargestellt.
2. Eine Linie wird durch die Punkte im Streudiagramm gezogen, die den bestmöglichen Anpassung an die Daten bietet. Dies bedeutet, dass die Summe der Abweichungen der Datenpunkte von der Regressionsgerade minimiert wird.
3. Die Regressionsgerade wird durch eine Gleichung dargestellt, die die Form y = a + b * x hat, wobei Y die abhängige Variable, X die unabhängige Variable, b die Steigung der Linie und a der y-Achsenabschnitt ist. Die Steigung b gibt an, um wie viele Einheiten die abhängige Variable sich ändert, wenn sich die unabhängige Variable um eine Einheit ändert, während der y-Achsenabschnitt a angibt, wo die Linie die y-Achse schneidet.
4. Die Güte der Anpassung der Regressionsgeraden an die Daten kann anhand des Bestimmtheitsmaßes R^2 bewertet werden. R^2 gibt an, wie viel Prozent der Variation in der abhängigen Variable durch die unabhängige Variable erklärt werden kann. Ein höherer R^2-Wert zeigt an, dass die Regressionsgerade besser an die Daten angepasst ist.
5. Wenn die Regressionsgerade erstellt wurde, kann sie verwendet werden, um Vorhersagen über die abhängige Variable zu treffen, wenn die unabhängigen Variablen bekannt sind. Dazu wird einfach der Wert der unabhängigen Variable in die Gleichung eingesetzt, um den Wert der abhängigen Variable zu berechnen.

Insgesamt bietet die lineare Regression daher eine einfache Möglichkeit, die Beziehung zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen zu modellieren und Vorhersagen über die abhängige Variable zu treffen.

LOGISTISCHE REGRESSION

Die logistische Regression ist ebenfalls eine statistische Methode, die verwendet wird, um die Beziehung zwischen einer binären abhängigen Variablen und einer oder mehreren unabhängigen Variablen zu modellieren. Die Methode wird häufig in der medizinischen Forschung, Marketing, Ökonometrie und anderen Bereichen verwendet, bei denen die Zielgröße nur zwei mögliche Ausgänge hat (bspw. ja / nein, erfolgreich / nicht erfolgreich, krank / gesund).

Die Funktionsweise der logistischen Regression lässt sich wie folgt beschreiben:

1. Zunächst werden die Daten gesammelt und in einer Tabelle oder einem Datensatz dargestellt, wobei die abhängige Variable (Y) als binäre Variable kodiert wird (bspw. 0 für "nicht erfolgreich" und 1 für "erfolgreich").
2. Eine logistische Regressionsgleichung wird erstellt, die die Wahrscheinlichkeit (p) für das Eintreten des Ereignisses als Funktion der unabhängigen Variablen (X) beschreibt. Die Gleichung hat die Form p = 1 / (1 + e^-z), wobei $z = b_0 + b_1 * x1 + b_2 * x_2 + ... + b_n * x_n$ ist und b_0, b_1, b_2, ..., b_n die Koeffizienten der Regressionsgleichung sind. Die Koeffizienten werden so berechnet, dass die logistische Regressionsgleichung die bestmögliche Anpassung an die Daten bietet.
3. Die logistische Regressionsgleichung wird verwendet, um Vorhersagen über die Wahrscheinlichkeit des Ereignisses (bspw. Erfolg) zu treffen, wenn die Werte der unabhängigen Variablen bekannt sind. Wenn die Wahrscheinlichkeit größer als 0,5 ist, wird das Ereignis als "ja" vorhergesagt, wenn sie kleiner als 0,5 ist, als "nein".

Insgesamt bietet die logistische Regression eine Möglichkeit, die Beziehung zwischen einer binären abhängigen Variablen und einer oder mehreren unabhängigen Variablen zu modellieren und Vorhersagen über das Eintreten des Ereignisses zu treffen. Die Methode ist besonders nützlich, wenn das Ziel darin besteht, die Wahrscheinlichkeit des Ereignisses als Funktion der unabhängigen Variablen zu verstehen und Vorhersagen über die Wahrscheinlichkeit des Ereignisses zu treffen.

INTERAKTIONSEFFEKT

In der Statistik bezeichnen Interaktionseffekte das Phänomen, dass der Effekt einer unabhängigen Variable auf die abhängige Variable von der Ausprägung einer oder mehrerer anderer unabhängiger Variablen abhängt.

Ein Beispiel für einen Interaktionseffekt könnte sein, dass der Zusammenhang zwischen der körperlichen Aktivität und dem Gewichtsverlust von der Ernährung abhängt. Das heißt, dass die Auswirkungen von körperlicher Aktivität auf das Gewicht davon abhängen, was eine Person isst. Wenn eine Person eine gesunde Ernährung hat, könnte körperliche Aktivität zu einem größeren Gewichtsverlust führen als bei einer Person mit einer ungesunden Ernährung.

Interaktionseffekte werden in der Statistik oft durch die Verwendung von Regressionsanalysen untersucht. Die Interaktionseffekte können in einer Regressionsanalyse durch Hinzufügen von Interaktionstermen oder Wechselwirkungstermen in das Modell berücksichtigt werden. Ein Interaktionsterm ist das Produkt zweier unabhängiger Variablen. Wenn ein Interaktionsterm in das Modell aufgenommen wird, kann der Effekt einer unabhängigen Variable auf die abhängige Variable variiert werden, je nachdem, welche Werte die anderen unabhängigen Variablen haben.

Es ist wichtig, Interaktionseffekte in der Statistik zu berücksichtigen, da sie dazu beitragen können, den wahren Zusammenhang zwischen unabhängigen und abhängigen Variablen zu verstehen. Ohne die Berücksichtigung von Interaktionseffekten können Ergebnisse falsch interpretiert werden, was zu falschen Schlussfolgerungen und Entscheidungen führen kann.

BOOTSTRAPPING

Bootstrapping ist eine statistische Methode, die verwendet wird, um Parameter und Schätzungen in einem Datensatz zu generieren. Die Methode basiert darauf, dass man aus einer vorhandenen Stichprobe (Sample) eine große Anzahl von künstlichen Stichproben erstellt, indem man zufällige Beobachtungen mit Zurücklegen aus der vorhandenen Stichprobe zieht.

Die Funktionsweise von Bootstrapping kann in folgenden Schritten beschrieben werden:

1. Ziehen einer Stichprobe: Zunächst wird eine Stichprobe aus einer Grundgesamtheit gezogen.
2. Erstellen von künstlichen Stichproben: Aus der vorhandenen Stichprobe werden mit Zurücklegen weitere Stichproben erstellt. Die Größe der künstlichen Stichproben ist dabei in der Regel genauso groß wie die der ursprünglichen Stichprobe.
3. Berechnung von Schätzungen: Für jede künstliche Stichprobe wird der Schätzwert für den Parameter berechnet, den man untersuchen möchte (bspw. Mittelwert, Standardabweichung, Korrelation, etc.). Dadurch erhält man eine Verteilung von Schätzungen.
4. Zusammenfassung der Ergebnisse: Die Ergebnisse der Schätzungen werden zusammengefasst, bspw. indem man den Mittelwert oder Median der Verteilung berechnet. Dabei erhält man ein robustes Schätzergebnis.

Bootstrapping ist besonders nützlich, wenn die Verteilung der Grundgesamtheit nicht bekannt ist oder wenn die Stichprobe nicht normalverteilt ist. Durch das Erstellen von künstlichen Stichproben wird die Varianz des Schätzwerts reduziert und die Ergebnisse werden robuster.

Bootstrapping wird oft in der Anwendung verwendet, um Vertrauensintervalle für eine Schätzung zu erstellen oder um die statistische Signifikanz eines Effekts zu testen. Es ist auch eine nützliche Methode beim Machine Learning, um Modelle zu trainieren und zu evaluieren, insbesondere wenn die Anzahl der Datenpunkte begrenzt ist.

S I G N I F I K A N Z

Statistische Signifikanz bezieht sich auf den Grad der Gewissheit oder Überzeugung, mit der man sagen kann, dass eine Beziehung zwischen Variablen in einem Datensatz nicht auf Zufall beruht, sondern tatsächlich vorhanden ist. Mit anderen Worten, es gibt eine signifikante Korrelation oder Unterschiede zwischen Gruppen in den Daten, die nicht auf natürliche Variation oder zufällige Stichprobenfehler zurückzuführen sind.

Die statistische Signifikanz wird oft durch die Anwendung von Hypothesentests beurteilt. Ein Hypothesentest prüft, ob die beobachteten Daten mit einer bestimmten Nullhypothese vereinbar sind oder nicht. Die Nullhypothese ist in der Regel die Annahme, dass es keine signifikanten Unterschiede oder Beziehungen zwischen den Variablen gibt. Wenn der p-Wert des Tests kleiner als das vorher festgelegte Signifikanzniveau ist, lehnt man die Nullhypothese ab und schließt daraus, dass es eine signifikante Beziehung zwischen den Variablen gibt.

Ein weiteres Konzept, das eng mit der statistischen Signifikanz verbunden ist, ist die Stichprobengröße. Je größer die Stichprobe, desto wahrscheinlicher ist es, dass man signifikante Ergebnisse erhält, da die Chance, zufällige Variation zu finden, sinkt. Daher ist es wichtig, die Stichprobengröße bei der Interpretation von statistischen Ergebnissen zu berücksichtigen.

Es ist jedoch wichtig zu betonen, dass statistische Signifikanz nicht gleichbedeutend mit praktischer Bedeutung oder Relevanz ist. Ein Unterschied oder eine Beziehung kann statistisch signifikant sein, aber dennoch eine geringe oder unbedeutende Auswirkung in der realen Welt haben. Daher sollte man bei der Interpretation von statistischen Ergebnissen immer die praktischen Auswirkungen und den Kontext berücksichtigen.

MASCHINELLES LERNEN

Das maschinelle Lernen (Machine Learning) ist ein Teilgebiet der künstlichen Intelligenz, bei dem Computer lernen, aus Daten Muster zu erkennen und Vorhersagen zu treffen, ohne explizit programmiert zu werden.

Die allgemeine Funktionsweise des maschinellen Lernens ist wie folgt:

1. Daten sammeln: Zunächst werden Daten aus verschiedenen Quellen gesammelt, bspw. aus Sensoren, Datenbanken, Social Media usw.
2. Daten vorbereiten: Die gesammelten Daten werden bereinigt, transformiert und aufbereitet, um sicherzustellen, dass sie für das Modell geeignet sind.
3. Modellauswahl: Ein passendes Modell wird ausgewählt, um die Daten zu analysieren. Dies kann ein klassisches Modell wie die logistische Regression oder ein komplexeres Modell wie neuronale Netze sein.
4. Trainieren des Modells: Das ausgewählte Modell wird auf den bereitgestellten Trainingsdaten trainiert, um Muster und Zusammenhänge zu erkennen.
5. Validierung: Das trainierte Modell wird auf validierenden Daten getestet, um zu überprüfen, wie gut es funktioniert. Wenn es nicht ausreichend genau ist, wird das Modell angepasst und erneut trainiert.
6. Vorhersage: Das trainierte und validierte Modell wird schließlich verwendet, um Vorhersagen über neue Daten zu treffen.
7. Optimierung: Das Modell wird optimiert und verbessert, indem verschiedene Techniken wie Hyperparameteroptimierung, Feature Engineering und Regularisierung angewendet werden.

Die allgemeine Funktionsweise des maschinellen Lernens ist ein iterativer Prozess, bei dem das Modell kontinuierlich trainiert, validiert und optimiert wird, um die besten Ergebnisse zu erzielen. Maschinelles Lernen findet Anwendung in verschiedenen Bereichen wie bspw. der Bild- und Spracherkennung.

TRAININGS– UND VALIDIERUNGSDATENSATZ

Beim Machine Learning geht es darum, aus Daten zu lernen und Vorhersagen zu treffen. Ein wichtiger Aspekt dabei ist die Aufteilung der Daten in einen Trainings- und einen Validierungsdatensatz.

Der Trainingsdatensatz ist die Menge an Daten, die verwendet wird, um das Modell zu trainieren. Das Modell lernt aus diesen Daten, welche Beziehungen zwischen den Eingabevariablen und der Ausgabevariablen bestehen, um Vorhersagen auf neuen Daten treffen zu können. Der Trainingsdatensatz sollte eine ausreichend große Stichprobe aus der gesamten Datenmenge sein, um das Modell ausreichend zu trainieren.

Der Validierungsdatensatz hingegen ist eine separate Menge von Daten, die verwendet wird, um die Leistung des Modells zu bewerten. Das Modell wird auf dem Validierungsdatensatz getestet, um zu sehen, wie gut es in der Lage ist, Vorhersagen auf neuen Daten zu treffen. Der Validierungsdatensatz sollte eine ausreichend große Stichprobe aus der gesamten Datenmenge sein, um die Leistung des Modells zuverlässig zu bewerten.

Die Aufteilung der Daten in Trainings- und Validierungsdatensatz ist wichtig, um sicherzustellen, dass das Modell nicht überangepasst wird. Überanpassung tritt auf, wenn das Modell zu stark auf die Trainingsdaten ausgerichtet ist und dadurch schlechte Vorhersagen auf neuen Daten trifft. Durch die Verwendung eines separaten Validierungsdatensatzes kann überprüft werden, ob das Modell tatsächlich gute Vorhersagen auf neuen Daten treffen kann.

Es ist daher wichtig, den Validierungsdatensatz nicht für das Training des Modells zu verwenden. Wenn der Validierungsdatensatz für das Training des Modells verwendet wird, besteht das Risiko, dass das Modell überangepasst wird und schlechte Vorhersagen auf neuen Daten trifft.

METRIKEN ZUR EVALUIERUNG

Folgende Metriken werden in der Evaluierung von Klassifikationsmodellen im Rahmen des maschinellen Lernens verwendet:

- Die Genauigkeit (Accuracy) ist die am häufigsten verwendete Metrik zur Bewertung von Klassifikationsmodellen. Sie gibt den Prozentsatz der korrekt vorhergesagten Beobachtungen im Verhältnis zu allen Beobachtungen an.
- Die Präzision (Precision) gibt an, wie viele der vorhergesagten positiven Ergebnisse tatsächlich korrekt sind. Sie gibt das Verhältnis von wahren positiven Ergebnissen zur Anzahl der positiven Vorhersagen insgesamt an.
- Der Recall (auch Sensitivität genannt) gibt an, wie viele der tatsächlich positiven Ergebnisse von einem Modell korrekt vorhergesagt wurden. Er gibt das Verhältnis von wahren positiven Ergebnissen zur Anzahl der tatsächlich positiven Ergebnisse insgesamt an.
- Der F1-Score ist eine gewichtete Mittelung von Präzision und Recall, die einen Wert zwischen 0 und 1 liefert. Er kombiniert Präzision und Recall zu einer einzigen Metrik, um die Gesamtleistung des Modells zu bewerten.

Diese Metriken sind alle wichtige Werkzeuge für die Evaluierung von Klassifikationsmodellen und helfen dabei, die Leistung des Modells in Bezug auf die Anzahl der korrekten und falschen Vorhersagen zu verstehen. Abhängig von den Anforderungen der spezifischen Anwendung kann eine oder mehrere dieser Metriken bevorzugt werden.

DEEP LEARNING

Deep Learning ist ein Teilgebiet des maschinellen Lernens, das auf künstlichen neuronalen Netzen (KNN) basiert. Im Gegensatz zu traditionellen neuronalen Netzen, die normalerweise nur wenige Schichten haben, verfügen tiefe neuronale Netze (Deep Neural Networks) über viele Schichten, die es ihnen ermöglichen, komplexe Zusammenhänge zwischen den Eingabedaten zu erkennen und zu lernen.

Die allgemeine Funktionsweise von Deep Learning ist wie folgt:

1. Eingabe: Ein tiefer neuronaler Netzwerk-Algorithmus nimmt eine große Menge an Eingabedaten auf.
2. Vorwärtspropagation: Die Eingabedaten werden durch die verschiedenen Schichten des Netzwerks geleitet. In jeder Schicht werden die Daten transformiert und durch Gewichtungen und Aktivierungsfunktionen weitergegeben.
3. Verlustfunktion: Eine Verlustfunktion wird verwendet, um den Unterschied zwischen den vorhergesagten Ergebnissen und den tatsächlichen Ergebnissen zu messen.
4. Rückwärtspropagation: Durch den Einsatz von Gradientenabstiegsverfahren werden die Gewichte der Neuronen im Netzwerk in einer Weise angepasst, die die Genauigkeit des Modells erhöht. Die Änderungen werden von der Ausgabeschicht rückwärts durch das Netzwerk propagiert, um die Gewichte in jeder Schicht zu aktualisieren.
5. Optimierung: Der Prozess der Vorwärts- und Rückwärtspropagation wird iterativ durchgeführt, wobei die Gewichte des Netzwerks schrittweise verbessert werden, bis eine akzeptable Genauigkeit erreicht ist.
6. Vorhersage: Sobald das Modell trainiert wurde, wird es verwendet, um Vorhersagen für neue Eingabedaten zu treffen.

NEURONALE NETZE

Neuronale Netze sind eine Art von künstlicher Intelligenz, die auf der Nachbildung der Arbeitsweise des menschlichen Gehirns basiert. Sie bestehen aus einer Reihe von miteinander verbundenen Neuronen, die in Schichten organisiert sind.

Ein typisches neuronales Netzwerk besteht aus einer Eingabeschicht, einer oder mehreren versteckten Schichten und einer Ausgabeschicht. Jede Schicht besteht aus einer bestimmten Anzahl von Neuronen, die jeweils eine bestimmte Anzahl von Eingängen haben.

Jedes Neuron in einem neuronalen Netzwerk ist mit anderen Neuronen in den benachbarten Schichten verbunden. Diese Verbindungen werden als Gewichte bezeichnet und dienen dazu, die Eingaben zu gewichten und zu transformieren, wenn sie durch das Netzwerk fließen.

Die Neuronen in einem neuronalen Netzwerk verwenden Aktivierungsfunktionen, um die gewichtete Summe ihrer Eingänge zu transformieren und ihre Ausgabe zu generieren. Die Ausgabe jedes Neurons wird dann an die nächsten Neuronen in der nächsten Schicht weitergeleitet, bis das Netzwerk eine Ausgabe erzeugt.

Während des Trainings durchlaufen neuronale Netze einen Prozess der Fehlerminimierung, bei dem die Gewichte im Netzwerk angepasst werden, um die Genauigkeit der Vorhersagen zu verbessern. Dies geschieht durch die Berechnung eines Fehlers zwischen der tatsächlichen Ausgabe des Netzwerks und der erwarteten Ausgabe, der dann zurückpropagiert wird, um die Gewichte entsprechend anzupassen.

Je tiefer und komplexer das neuronale Netzwerk ist, desto mehr Schichten und Neuronen hat es und desto komplexere Funktionen kann es erlernen. Neuronale Netze werden in vielen Anwendungen eingesetzt, wie bspw. Bilderkennung, Spracherkennung und Robotik.

AKTIVIERUNGSFUNKTION

Aktivierungsfunktionen werden in neuronalen Netzen verwendet, um die Ausgabe eines Neurons zu berechnen, indem sie die Summe der gewichteten Eingaben des Neurons transformieren. Die Aktivierungsfunktion bestimmt, ob das Neuron aktiviert wird oder nicht, indem sie einen Schwellenwert auf die transformierte Eingabe anwendet.

Es gibt verschiedene Arten von Aktivierungsfunktionen, die in neuronalen Netzen verwendet werden. Jede Aktivierungsfunktion hat ihre eigenen Vorteile und Einschränkungen und eignet sich für verschiedene Anwendungen.

Die Sigmoid-Funktion wird häufig in frühen neuronalen Netzen verwendet. Sie hat den Vorteil, dass sie stetig differenzierbar ist und eine Ausgabe im Bereich von 0 bis 1 erzeugt, was sie besonders geeignet macht, um Wahrscheinlichkeiten zu modellieren.

ReLU (Rectified Linear Unit) ist eine Aktivierungsfunktion, die in tieferen neuronalen Netzen verwendet wird, da sie eine bessere Leistung als die Sigmoid-Funktion bietet. ReLU ist einfach zu berechnen und kann schnell große Mengen von Daten verarbeiten.

Die Tanh-Funktion ist eine symmetrische Variante der Sigmoid-Funktion. Die Tanh-Funktion hat den Vorteil, dass sie stärker an den Nullpunkt zentriert ist und daher besser geeignet ist, um Daten mit negativen Werten zu modellieren.

Die Softmax-Funktion wird häufig verwendet, um Mehrklassenklassifizierungsprobleme zu lösen. Die Softmax-Funktion wandelt eine Eingabe in eine Wahrscheinlichkeitsverteilung um, die angibt, wie wahrscheinlich es ist, dass die Eingabe zu jeder der möglichen Klassen gehört.

Die Wahl der Aktivierungsfunktion hängt von der Art des Modells und der Aufgabe ab, die gelöst werden soll. Die Wahl der falschen Aktivierungsfunktion kann zu schlechter Leistung oder Konvergenzproblemen führen.

TRANSFORMER

Transformer sind eine Art von tiefen neuronalen Netzwerken, die für die Verarbeitung von Sequenzen verwendet werden. Sie wurden erstmals 2017 vorgestellt und haben seitdem in verschiedenen Anwendungen wie maschinellem Übersetzen, Textgenerierung, Sprach-erkennung und vielen anderen Anwendungen große Erfolge erzielt.

Die Funktionsweise von Transformern basiert auf der Idee der Aufmerksamkeit (Attention). Bei der Verarbeitung von Sequenzen besteht das Ziel darin, jedem Element in der Sequenz eine Bedeutung zuzuordnen, die von der Bedeutung der anderen Elemente in der Sequenz abhängt. Beispielsweise hängt die Bedeutung eines Wortes in einem Satz von der Bedeutung der anderen Wörter in dem Satz ab.

Transformers verwenden einen Mechanismus der selbstaufmerksamen Verarbeitung (self-attention), der es jedem Element in der Sequenz ermöglicht, sich auf alle anderen Elemente in der Sequenz zu beziehen und ihre Bedeutung zu erfassen. Dies wird erreicht, indem für jedes Element in der Sequenz eine sogenannte Aufmerksamkeitsmaske erstellt wird, die angibt, wie viel Aufmerksamkeit jedem anderen Element in der Sequenz zukommen sollte.

Der Transformer besteht aus mehreren Schichten, die aus mehreren Modulen bestehen. Das Kernmodul des Transformers ist das Multi-Head-Attention-Modul, das für die Selbst-Aufmerksamkeit verwendet wird. In diesem Modul werden die Eingabedaten in mehrere Vektoren aufgeteilt, die dann parallel berechnet werden. Jeder dieser Vektoren repräsentiert einen Aspekt der Aufmerksamkeit der Eingabe.

Neben dem Multi-Head-Attention-Modul gibt es noch andere Module, wie zum Beispiel das Feedforward-Netzwerk, das zur Transformation der Daten in einen anderen Raum verwendet wird. Der gesamte Transformer wird trainiert, um die Zielvariable vorherzusagen, indem er eine Verlustfunktion minimiert.

Insgesamt haben Transformer aufgrund ihrer Fähigkeit, die Beziehungen zwischen Elementen in einer Sequenz zu modellieren, zu einer signifikanten Verbesserung der Leistung in verschiedenen Anwendungen geführt.

NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) ist ein Bereich der künstlichen Intelligenz (KI), der sich mit der Verarbeitung und Analyse von natürlicher Sprache beschäftigt. Es geht darum, Maschinen beizubringen, menschliche Sprache zu verstehen und in verschiedenen Anwendungen zu verwenden.

NLP umfasst mehrere Schritte, um natürliche Sprache in eine maschinenlesbare Form zu bringen, damit sie von Computern verarbeitet werden kann. Dazu gehören folgende Schritte:

1. Tokenisierung: Die Textdaten werden in Wörter, Sätze und Absätze unterteilt.
2. Morphologische Analyse: Die Wörter werden in ihre Grundformen aufgeteilt und grammatikalische Eigenschaften wie Wortart, Numerus und Kasus werden bestimmt.
3. Syntaxanalyse: Hierbei wird die Struktur des Textes untersucht, um die Bedeutung der Wörter im Kontext zu verstehen.
4. Semantische Analyse: Die Bedeutung der Wörter und Sätze im Kontext wird analysiert.
5. Diskursanalyse: Der Text wird in den größeren Zusammenhang eingeordnet und das Verständnis der Bedeutung von Textabschnitten wird verbessert.

Ein Beispiel fpr NLP ist die automatische Übersetzung von Texten in verschiedene Sprachen. Eine weitere Anwendung ist die Erkennung von Stimmungen in Texten, wie sie bei der Sentimentanalyse eingesetzt wird. Chatbots, virtuelle Assistenten und automatische Spracherkennung sind ebenfalls Beispiele für die Anwendung von NLP-Technologien.

S E N T I M E N T A N A L Y S E

Sentimentanalysen sind Verfahren, die eingesetzt werden, um die Stimmung oder das Gefühl in einer Textpassage, einer E-Mail, einem Tweet, einer Bewertung oder einer anderen Art von Schreibarbeit automatisch zu identifizieren und zu bewerten. Das Ziel der Sentimentanalyse ist es, den emotionalen Gehalt eines Textes zu extrahieren und die vorherrschende Stimmung in positiv, negativ oder neutral zu klassifizieren.

Die Funktionweise von Sentimentanalysen basiert auf dem maschinellen Lernen und der Verarbeitung natürlicher Sprache (NLP). Dabei wird ein Text in kleinere Einheiten wie Wörter, Sätze oder Absätze zerlegt und dann jedes Element mit einem vordefinierten Wortschatz, Regeln oder maschinellen Lernalgorithmen bewertet. Der Wortschatz besteht aus einer Sammlung von Wörtern und Phrasen, die mit einem bestimmten Gefühl oder einer bestimmten Bedeutung assoziiert sind. Jedes Wort oder jede Phrase wird mit einem bestimmten Gewicht versehen, das angibt, wie stark es die Stimmung beeinflusst.

Wenn ein Text eingegeben wird, durchläuft er den Algorithmus, der jedes Wort oder jede Phrase in dem Text mit dem Wortschatz abgleicht, um festzustellen, welche Wörter positiv, negativ oder neutral sind. Das Ergebnis ist eine Bewertung, die angibt, ob der Text insgesamt positiv, negativ oder neutral ist. Sentimentanalysen können auch in der Lage sein, die Intensität der Stimmung zu bewerten, indem sie die Anzahl der positiven oder negativen Wörter im Verhältnis zu den neutralen Wörtern berücksichtigen.

Sentimentanalysen werden oft in Unternehmen eingesetzt, um Feedback von Kunden zu analysieren, in sozialen Medien, um die öffentliche Meinung zu einem Thema zu messen oder in der Finanzindustrie, um die Stimmung der Anleger auf den Märkten zu bewerten. Sie können auch von Marktforschungsunternehmen oder Regierungsbehörden verwendet werden, um Trends in der öffentlichen Meinung oder das Gefühl der Bürger zu bestimmten Themen zu erkennen.

CLUSTERANALYSE

Eine Clusteranalyse ist ein statistisches Verfahren zur Gruppierung von Objekten (bspw. Datenpunkte, Beobachtungen oder Merkmale) in homogene Untergruppen, die als Cluster bezeichnet werden. Die allgemeine Funktionsweise einer Clusteranalyse besteht aus folgenden Schritten:

1. Auswahl der Daten: Als erstes werden die Daten ausgewählt, die in die Analyse einbezogen werden sollen. Die Daten können aus verschiedenen Quellen stammen, bspw. aus Umfragen, Experimenten oder Sensordaten.
2. Auswahl der Variablen: Anschließend werden die Variablen ausgewählt, die zur Bildung der Cluster herangezogen werden sollen. Die Variablen können quantitativ oder qualitativ sein und können verschiedene Eigenschaften messen, bspw. Größe, Gewicht, Alter, Geschlecht, Einkommen oder Interessen.
3. Wahl der Distanz- oder Ähnlichkeitsmaße: Um die Ähnlichkeit zwischen den Objekten zu bestimmen, werden Distanz- oder Ähnlichkeitsmaße verwendet. Die Wahl der Maße hängt von der Art der Daten und der Fragestellung ab. Einige gängige Maße sind die Euklidische Distanz, die Manhattan-Distanz oder der Kosinus-Ähnlichkeitskoeffizient.
4. Durchführung der Clusteranalyse: Nach der Vorbereitung der Daten, der Auswahl der Variablen und der Maße wird die eigentliche Clusteranalyse durchgeführt. Dabei werden die Objekte so gruppiert, dass innerhalb der Cluster die Ähnlichkeit zwischen den Objekten möglichst hoch und zwischen den Clustern möglichst gering ist. Es gibt verschiedene Verfahren zur Durchführung einer Clusteranalyse, bspw. hierarchische oder partitionierende Verfahren.
5. Interpretation der Ergebnisse: Nach der Durchführung der Clusteranalyse müssen die Ergebnisse interpretiert werden. Dazu können verschiedene statistische und graphische Verfahren verwendet werden, um die Merkmale und Eigenschaften der Cluster zu beschreiben. Dabei ist es wichtig, die Ergebnisse kritisch zu hinterfragen und auf ihre Plausibilität zu überprüfen.
6. Anwendung der Ergebnisse: Die Ergebnisse der Clusteranalyse können zur Identifikation von Gruppen mit ähnlichen Eigenschaften oder Verhaltensweisen verwendet werden. Sie können als Grundlage für die Entwicklung von Marketingstrategien, Kundenprofilen oder zur Segmentierung von Zielgruppen dienen.

K N N – A L G O R I T H M U S

Der KNN-Algorithmus (K-Nearest Neighbors) ist ein einfacher Algorithmus des maschinellen Lernens und wird zur Klassifikation und Regression von Datenpunkten verwendet. Die Funktionsweise des Algorithmus lässt sich in folgende Schritte unterteilen:

1. Datenvorbereitung: Zunächst werden die Trainingsdaten vorbereitet, die zur Erstellung des KNN-Modells verwendet werden sollen. Dabei werden die Datenpunkte in eine n-dimensionale Raumstruktur eingeordnet, wobei jede Dimension ein Merkmal der Datenpunkte repräsentiert.
2. Wahl des K: Als nächstes wird eine passende Anzahl an Nachbarn K gewählt, die zur Bestimmung der Klasse oder des Wertes eines neuen Datenpunktes herangezogen werden sollen. Eine höhere Anzahl an Nachbarn führt zu einer glatteren Entscheidungsgrenze, während eine niedrigere Anzahl zu einer unruhigeren Grenze führt.
3. Bestimmung der K Nachbarn: Nun wird der neue Datenpunkt in der Raumstruktur eingeordnet und die K nächsten Nachbarn bestimmt. Dies erfolgt durch die Berechnung der Abstände zwischen dem neuen Datenpunkt und den bereits vorhandenen Trainingsdaten.
4. Klassifikation/Regression: Wenn die K Nachbarn bestimmt wurden, wird anhand ihrer Labels (bei Klassifikation) oder ihrer Werte (bei Regression) eine Vorhersage für den neuen Datenpunkt getroffen. Dabei wird in der Klassifikation die am häufigsten vorkommende Klasse der K Nachbarn als Vorhersage gewählt, während in der Regression der Durchschnitt der K Werte berechnet wird.
5. Bewertung der Vorhersage: Schließlich wird die Vorhersage des Algorithmus auf ihre Richtigkeit überprüft, indem die Vorhersage mit den tatsächlichen Werten verglichen wird. Hierbei kann eine Vielzahl von Bewertungsmaßen verwendet werden, je nach Anwendungsfall und Datentyp.
6. Anwendung des Algorithmus: Der KNN-Algorithmus kann zur Klassifikation und Regression von neuen Datenpunkten verwendet werden, sobald das KNN-Modell mit den Trainingsdaten erstellt wurde. Der Algorithmus kann in verschiedenen Anwendungsbereichen eingesetzt werden, bspw. in der Bilderkennung, der Spracherkennung oder der Vorhersage von Aktienkursen.

K – MEANS – ALGORITHMUS

Der K-Means Algorithmus ist ein Algorithmus des maschinellen Lernens, der zur Gruppierung von Datenpunkten in K Cluster verwendet wird. Die Funktionsweise des Algorithmus lässt sich in folgende Schritte unterteilen:

1. Wahl des K: Zunächst wird eine passende Anzahl an Clustern K gewählt, in die die Datenpunkte gruppiert werden sollen.
2. Initialisierung: Anschließend werden K zufällige Zentren (Centroids) in der Datenmenge platziert, die als Startpunkte für die Clusterbildung dienen.
3. Zuordnung der Datenpunkte: Jeder Datenpunkt wird dem am nächsten gelegenen Zentrum zugewiesen und somit einem Cluster zugeordnet. Dabei wird in der Regel die euklidische Distanz zur Bestimmung der Nähe zwischen Datenpunkt und Zentrum verwendet.
4. Aktualisierung der Zentren: Nach der Zuordnung der Datenpunkte werden die Zentren jedes Clusters neu berechnet, indem der Durchschnitt aller Datenpunkte in diesem Cluster gebildet wird. Dadurch werden die Zentren in die Mitte ihrer zugewiesenen Datenpunkte verschoben.
5. Wiederholung: Die Schritte 3 und 4 werden wiederholt, bis sich die Zuordnung der Datenpunkte zu den Clustern nicht mehr ändert oder eine vorgegebene Anzahl an Iterationen erreicht ist.
6. Bewertung der Cluster: Schließlich werden die gebildeten Cluster auf ihre Qualität hin untersucht. Hierbei können verschiedene Bewertungsmaße verwendet werden, bspw. die Varianz innerhalb der Cluster oder der Abstand zwischen den Zentren.
7. Anwendung des Algorithmus: Der K-Means Algorithmus kann zur Gruppierung von Datenpunkten in verschiedenen Anwendungsbereichen eingesetzt werden, bspw. in der Kunden-Segmentierung oder der Analyse von Gesundheitsdaten.

Der K-Means Algorithmus ist ein einfacher und schnell implementierbarer Algorithmus zur Gruppierung von Datenpunkten, der allerdings von der Wahl des Anfangszustandes abhängig ist und zu lokalen Optima führen kann. Aus diesem Grund wird der Algorithmus oft mehrmals mit unterschiedlichen Anfangszuständen ausgeführt, um ein besseres Ergebnis zu erzielen.

RANDOM FOREST–ALGORITHMUS

Der Random Forest Algorithmus ist ein Supervised Learning Algorithmus, der für die Klassifikation und Regression von Daten verwendet wird. Der Algorithmus kombiniert mehrere Entscheidungsbäume zu einem "Wald" und verwendet dann den Durchschnitt der Vorhersagen der einzelnen Bäume als Endergebnis.

Die Funktionsweise des Random Forest Algorithmus lässt sich wie folgt beschreiben:

1. Datensatz teilen: Der Datensatz wird in Trainings- und Testdaten aufgeteilt.
2. Entscheidungsbäume erstellen: Der Random Forest Algorithmus erstellt mehrere Entscheidungsbäume aus dem Trainingsdatensatz. Jeder Baum wird auf einer zufälligen Teilstichprobe des Trainingsdatensatzes erstellt, die Bootstrap-Methode genannt wird. Die Bootstrap-Methode wählt zufällig eine bestimmte Anzahl von Beobachtungen aus dem Trainingsdatensatz aus, um den Baum zu erstellen.
3. Entscheidungsbäume kombinieren: Die Vorhersage jedes Entscheidungsbaumes wird als Endergebnis kombiniert. Wenn der Algorithmus für eine Klassifikation verwendet wird, wird das Endergebnis als Mehrheitsvotum der Vorhersage jedes Entscheidungsbaumes gewählt. Wenn der Algorithmus für eine Regression verwendet wird, wird das Endergebnis als Durchschnitt der Vorhersage jedes Entscheidungsbaumes gewählt.
4. Testen des Modells: Das erstellte Modell wird auf dem Testdatensatz getestet, um die Genauigkeit des Modells zu bewerten.

Die Vorteile des Random Forest Algorithmus sind:

- Random Forest ist ein leistungsstarker Algorithmus, der in der Lage ist, komplexe Zusammenhänge in den Daten zu erfassen.
- Der Algorithmus ist resistent gegenüber Overfitting, da er mehrere Bäume erstellt und kombiniert.
- Der Algorithmus ist auch in der Lage, mit fehlenden Werten und Ausreißern umzugehen.
- Random Forest kann sowohl für Klassifikation als auch Regression eingesetzt werden.

Die Nachteile des Random Forest Algorithmus sind:

- Der Algorithmus kann sehr zeitaufwändig sein, insbesondere wenn der Datensatz sehr groß ist.
- Der Algorithmus kann schwierig zu interpretieren sein, da es schwierig sein kann, die Vorhersagen jedes Entscheidungsbaumes zu verstehen.
- Random Forest kann möglicherweise nicht die bestmögliche Leistung erzielen, wenn die Datenstruktur sehr einfach ist, da es zu Overfitting kommen kann.

AGGLOMERATIVES CLUSTERING–ALGORITHMUS

Das Agglomerative Clustering ist ein Bottom-up-Clustering-Algorithmus, bei dem jedes Datenobjekt zunächst als eigenes Cluster betrachtet wird und dann schrittweise zu größeren Clustern zusammengefasst wird. Der Algorithmus beginnt mit der Erstellung von N Clustern, wobei jedes Cluster nur ein Datenobjekt enthält. Der Algorithmus arbeitet iterativ und wiederholt den folgenden Schritt, bis alle Datenobjekte zu einem Cluster zusammengefasst sind:

1. Berechnung der Ähnlichkeit: Die Ähnlichkeit zwischen jedem Paar von Clustern wird berechnet. Es gibt verschiedene Möglichkeiten, die Ähnlichkeit zu berechnen, bspw. die euklidische Distanz, die Manhattan-Distanz oder die Korrelation.
2. Zusammenfassung der ähnlichsten Cluster: Die ähnlichsten Cluster werden zu einem größeren Cluster zusammengefasst. Die Ähnlichkeit zwischen den Clustern wird durch einen sogenannten Linkage-Algorithmus berechnet. Es gibt verschiedene Linkage-Methoden, bspw. die Single-Linkage-Methode, die Complete-Linkage-Methode und die Average-Linkage-Methode.
3. Wiederholung des Schritts 1 und 2: Der Prozess wird wiederholt, bis alle Datenobjekte in einem einzigen Cluster zusammengefasst sind.

Das Ergebnis des Agglomerativen Clusterings ist ein Dendrogramm, das die Hierarchie der Zusammenfassung der Cluster zeigt. Ein Dendrogramm ist eine Baumstruktur, die die Abfolge der Zusammenfassung der Cluster zeigt. Die Blätter des Baumes entsprechen den Datenobjekten, während die Knoten des Baumes die Zusammenfassung der Cluster darstellen.

Die Vorteile des Agglomerativen Clusterings sind:

- Der Algorithmus ist einfach zu implementieren und erfordert keine Voraussetzungen über die Struktur oder Verteilung der Daten.
- Es ist ein flexibler Algorithmus, der verschiedene Ähnlichkeitsmaße und Linkage-Methoden verwenden kann.
- Es ist in der Lage, sowohl sphärische als auch nicht-sphärische Clusterstrukturen zu erkennen.

Die Nachteile des Agglomerativen Clusterings sind:

- Es kann zeitaufwändig sein, insbesondere bei großen Datensätzen.
- Es kann schwierig sein, die optimale Anzahl von Clustern zu bestimmen.
- Es kann zu Problemen mit Ausreißern kommen, da diese dazu führen können, dass sich die Cluster falsch zusammenschließen.

BIRCH – ALGORITHMUS

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) ist ein Clustering-Algorithmus, der für große Datenmengen optimiert wurde. Der Algorithmus versucht, die Anzahl der Datenobjekte zu reduzieren und gleichzeitig eine kompakte Datenstruktur zu erhalten, um schnelles Clustering zu ermöglichen. BIRCH verwendet eine Hierarchie von Clustern, wobei jeder Knoten in der Hierarchie ein Cluster darstellt.

Die Funktionsweise von BIRCH lässt sich in folgende Schritte gliedern:

1. Aufbau des CF-Baums: Zunächst wird ein CF-Baum (Clustering Feature Tree) aufgebaut, der eine kompakte Zusammenfassung der Daten darstellt. Der CF-Baum besteht aus einer Hierarchie von Knoten, wobei jeder Knoten einen Cluster repräsentiert. Jeder Knoten im CF-Baum enthält eine Liste von Clustering Features (CF), die als Zusammenfassung der Datenobjekte im Cluster dienen. Die CFs können beispielsweise der Mittelwert und die Varianz der Merkmale sein.
2. Clustering im CF-Baum: Die Datenobjekte werden in den CF-Baum eingefügt, indem sie in das am nächsten gelegene Blatt des Baumes eingefügt werden. Wenn ein Knoten mehr als einen Schwellenwert an CFs enthält, wird er in zwei oder mehrere Knoten aufgeteilt. Dies wird als Clustering im CF-Baum bezeichnet.
3. Clustering auf dem CF-Baum: Nachdem der CF-Baum aufgebaut wurde, wird er rekursiv durchlaufen, um Cluster zu bilden. Zunächst wird der unterste Knoten des Baumes ausgewählt, der die Datenobjekte enthält. Wenn die Anzahl der Datenobjekte unter einem Schwellenwert liegt, wird der Knoten als Cluster betrachtet. Andernfalls werden die CFs des Knotens verwendet, um die Ähnlichkeit zwischen den Clustern zu berechnen. Die ähnlichsten Cluster werden zu einem größeren Cluster zusammengefasst, bis alle Datenobjekte in Clustern enthalten sind.
4. Zusammenfassung der Daten: Nachdem die Cluster gebildet wurden, kann eine Zusammenfassung der Daten in jedem Cluster berechnet werden, indem die CFs des entsprechenden Knotens im CF-Baum verwendet werden. Dadurch wird eine kompakte Repräsentation der Daten erzeugt, die zur schnellen Verarbeitung von neuen Daten verwendet werden kann.

Die Vorteile von BIRCH sind:

- Der Algorithmus verwendet eine kompakte Datenstruktur und ist für große Datenmengen geeignet.
- Er ist für sphärische als auch nicht-sphärische Clusterstrukturen geeignet.

Die Nachteile von BIRCH sind:

- Der optimale Schwellenwert für die Clustertrennung ist schwer zu bestimmen.
- Es ist anfällig für Ausreißer, die die Clusterbildung beeinflussen können.

DBSCAN – ALGORITHMUS

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ist ein Clustering-Algorithmus, der es ermöglicht, Datenpunkte in verschiedene Cluster zu gruppieren. Im Gegensatz zu K-Means, einem häufig verwendeten Clustering-Algorithmus, ist DBSCAN in der Lage, Cluster unterschiedlicher Größe, Form und Dichte zu erkennen.

Die Funktionsweise des DBSCAN-Algorithmus lässt sich wie folgt beschreiben:

1. Auswahl des Kernpunktes: Der Algorithmus beginnt damit, einen zufälligen Datenpunkt auszuwählen und überprüft dann, ob der Punkt ein Kernpunkt ist. Ein Kernpunkt ist definiert als ein Datenpunkt, der mindestens K Nachbarn innerhalb eines bestimmten Radius (Epsilon) hat.
2. Erweiterung des Clusters: Wenn ein Kernpunkt gefunden wird, wird ein neues Cluster erstellt und alle erreichbaren Punkte, die innerhalb des Radius Epsilon des Kernpunkts liegen, werden dem Cluster hinzugefügt. Der Prozess wird dann iterativ fortgesetzt, bis alle erreichbaren Punkte hinzugefügt wurden und kein neuer Kernpunkt gefunden wird.
3. Definition von Randpunkten: Wenn ein Punkt innerhalb des Radius Epsilon eines Kernpunkts liegt, aber selbst kein Kernpunkt ist, wird er als Randpunkt bezeichnet. Randpunkte werden dann dem nächstgelegenen Cluster zugeordnet.
4. Identifikation von Ausreißern: Datenpunkte, die keine Kern- oder Randpunkte sind, werden als Ausreißer bezeichnet und in keinem Cluster eingeordnet.

Der DBSCAN-Algorithmus ermöglicht es, komplexe Clusterstrukturen zu erkennen und eignet sich daher besonders für Datensätze, die eine hohe Dichtevariation aufweisen. Ein Nachteil des DBSCAN-Algorithmus ist jedoch, dass er relativ empfindlich gegenüber den Parametern Epsilon und K ist und dass er in Datensätzen mit unterschiedlichen Dichten möglicherweise nicht so gut funktioniert.

SPECTRAL CLUSTERING–ALGORITHMUS

Das Spectral Clustering ist ein Clustering-Algorithmus, der auf der Spektraltheorie der Graphen basiert. Im Gegensatz zu anderen Clustering-Verfahren basiert dieser Algorithmus auf der Eigenschaft, dass die Datenpunkte als Knoten in einem Graphen dargestellt werden können, und die Struktur des Graphen die Struktur der Daten widerspiegelt.

Die Funktionsweise des Spectral Clustering Algorithmus lässt sich wie folgt beschreiben:

1. Erstellung des Graphen: Zunächst wird ein Graph erstellt, indem jeder Datenpunkt als Knoten im Graphen dargestellt wird und die Kanten zwischen den Knoten anhand einer Ähnlichkeitsmatrix definiert werden. Dabei wird die Ähnlichkeit zwischen den Datenpunkten anhand eines Ähnlichkeitsmaßes wie bspw. der euklidischen Distanz oder des Kosinus-Ähnlichkeitsmaßes berechnet.
2. Konstruktion der Laplace-Matrix: Aus der Adjazenzmatrix des Graphen wird die Laplace-Matrix berechnet. Diese Matrix enthält Informationen über die lokale Struktur des Graphen und spiegelt die Ähnlichkeiten der Datenpunkte untereinander wider.
3. Eigenwertzerlegung der Laplace-Matrix: Durch die Eigenwertzerlegung der Laplace-Matrix werden die Eigenwerte und Eigenvektoren berechnet. Die Eigenvektoren werden dann als Basisfunktionen für das Clustering verwendet.
4. Reduktion der Dimensionalität: Die Dimensionalität des Datenraums wird durch die Verwendung der Eigenvektoren reduziert, die die größten Eigenwerte aufweisen. Die Anzahl der ausgewählten Eigenvektoren entspricht der Anzahl der Cluster, die gesucht werden.
5. Anwendung von K-Means: Schließlich werden die reduzierten Datenpunkte mithilfe des K-Means Algorithmus in K-Cluster gruppiert.

Spectral Clustering kann dabei helfen, Datencluster in nicht-linearen Strukturen zu erkennen und eignet sich besonders gut für Datensätze mit vielen Dimensionen. Es gibt jedoch einige Herausforderungen, die bei der Anwendung des Spectral Clustering Algorithmus berücksichtigt werden müssen, wie bspw. die Wahl der Parameter, die Auswahl des passenden Ähnlichkeitsmaßes und die Handhabung von Datensätzen mit vielen Rauschpunkten.

MEAN SHIFT–ALGORITHMUS

Der Mean Shift Algorithmus ist ein Clustering-Algorithmus, der die Dichte der Datenpunkte im Raum verwendet, um Cluster zu bilden. Die Grundidee des Algorithmus ist es, einen Fensterbereich um jeden Datenpunkt zu betrachten und den Mittelpunkt dieses Fensters zu verschieben, um die höchste Dichte von Datenpunkten innerhalb dieses Fensters zu finden. Der Algorithmus findet dann iterativ neue Mittelpunkte, bis ein stabiler Konvergenzpunkt erreicht wird.

Die Funktionsweise des Mean Shift Algorithmus lässt sich wie folgt beschreiben:

1. Initialisierung: Zunächst wird jeder Datenpunkt als möglicher Mittelpunkt betrachtet und ein Fenster um jeden Punkt herum definiert. Die Größe des Fensters wird durch einen Bandbreitenparameter bestimmt.
2. Berechnung des Schwerpunkts: In jedem Schritt wird der Schwerpunkt des Fensters um jeden Punkt berechnet. Dies wird erreicht, indem der gewichtete Durchschnitt aller Punkte im Fensterbereich berechnet wird, wobei jeder Punkt als Gewicht seine Dichte innerhalb des Fensters hat.
3. Verschiebung des Fensters: Das Fenster wird nun um den neuen Schwerpunkt verschoben, und der Schwerpunkt wird erneut berechnet. Dieser Prozess wird so lange fortgesetzt, bis der Schwerpunkt innerhalb des Fensters konvergiert.
4. Zusammenfassen der Konvergenzpunkte: Konvergente Mittelpunkte werden zusammengefasst und als Cluster bezeichnet.

Der Mean Shift Algorithmus ist in der Lage, Cluster unterschiedlicher Größe und Form zu erkennen, da die Größe des Fensters dynamisch angepasst wird, um die höchste Dichte von Datenpunkten innerhalb des Fensters zu finden. Der Algorithmus ist jedoch empfindlich gegenüber der Wahl der Bandbreitenparameter, und es kann schwierig sein, geeignete Parameter für bestimmte Datensätze zu finden. Außerdem kann der Algorithmus bei sehr großen Datensätzen ineffizient sein, da jeder Punkt als potenzieller Mittelpunkt betrachtet wird.

GAUSSIAN MIXTURE–ALGORITHMUS

Der Gaussian Mixture Algorithmus ist ein probabilistisches Clustering-Verfahren, das auf der Annahme basiert, dass die Datenpunkte aus einer Mischung von mehreren normalverteilten Clustern stammen. Der Algorithmus verwendet die Maximum-Likelihood-Methode, um die Parameter der Verteilungen zu schätzen und die Datenpunkte den entsprechenden Clustern zuzuordnen.

Die Funktionsweise des Gaussian Mixture Algorithmus lässt sich wie folgt beschreiben:

1. Initialisierung: Zunächst werden die Anzahl der Cluster und die Parameter der normalverteilten Verteilungen initialisiert. Dazu gehören die Mittelwerte, Varianzen und Gewichte der Verteilungen.
2. Berechnung der Wahrscheinlichkeiten: Für jeden Datenpunkt wird die Wahrscheinlichkeit berechnet, dass er aus jedem der Cluster stammt. Dies wird erreicht, indem die Dichte jeder normalverteilten Verteilung an der Position des Datenpunktes berechnet wird und mit dem Gewicht der Verteilung multipliziert wird.
3. Zuteilung der Datenpunkte: Jeder Datenpunkt wird dem Cluster zugeordnet, für den er die höchste Wahrscheinlichkeit hat.
4. Schätzung der Parameter: Die Parameter der normalverteilten Verteilungen werden durch Maximierung der Likelihood-Funktion geschätzt. Dies wird erreicht, indem die Schätzungen der Mittelwerte, Varianzen und Gewichte der Verteilungen iterativ aktualisiert werden, um die beste Übereinstimmung zwischen den Datenpunkten und den Verteilungen zu erreichen.
5. Konvergenz: Der Algorithmus konvergiert, wenn die Veränderung der Parameter innerhalb einer bestimmten Schranke liegt oder wenn eine maximale Anzahl von Iterationen erreicht ist.

Der Gaussian Mixture Algorithmus kann verschiedene Arten von Clusterstrukturen erkennen und eignet sich besonders für Datensätze mit hoher Dimensionalität. Er hat jedoch auch einige Einschränkungen, wie bspw. die Wahl der Anzahl der Cluster, die Empfindlichkeit gegenüber Ausreißern und die Möglichkeit, in lokalen Optima stecken zu bleiben.

MAXIMUM–LIKELIHOOD–METHODE

Die Maximum-Likelihood-Methode ist ein statistisches Verfahren zur Schätzung der Parameter einer Wahrscheinlichkeitsverteilung, das häufig in der statistischen Inferenz und verwendet wird. Die Methode basiert auf der Annahme, dass die beobachteten Daten aus einer bestimmten Wahrscheinlichkeitsverteilung stammen und zielt darauf ab, die Werte der Parameter zu finden, die am wahrscheinlichsten die beobachteten Daten erklären.

Die Funktionsweise der Maximum-Likelihood-Methode lässt sich wie folgt beschreiben:

1. Annahme einer Wahrscheinlichkeitsverteilung: Zunächst wird eine Wahrscheinlichkeitsverteilung für die beobachteten Daten angenommen. Diese Verteilung kann eine bekannte parametrische Verteilung wie die Normalverteilung oder eine nichtparametrische Verteilung wie eine Kernel Density Estimation sein.
2. Definition der Likelihood-Funktion: Die Likelihood-Funktion ist die Wahrscheinlichkeit, dass die beobachteten Daten bei einer gegebenen Wahl der Verteilungsparameter beobachtet wurden. Diese Funktion wird als Produkt der Wahrscheinlichkeiten der einzelnen Datenpunkte berechnet, die sich aus der Wahrscheinlichkeitsverteilung ergeben.
3. Schätzung der Parameter: Die Schätzung der Parameter wird durch Maximierung der Likelihood-Funktion erreicht. Das bedeutet, dass die Werte der Parameter gesucht werden, die die größte Wahrscheinlichkeit für die beobachteten Daten liefern. Dies kann durch Ableiten der Likelihood-Funktion nach den Parametern und dem Setzen der Ableitungen auf Null erreicht werden. In einigen Fällen kann es jedoch nicht möglich sein, die Likelihood-Funktion analytisch zu maximieren, und es sind numerische Optimierungsmethoden erforderlich.
4. Konvergenz: Die Schätzungen der Parameter konvergieren, wenn die Änderungen der Schätzungen innerhalb einer bestimmten Schranke liegen oder wenn eine maximale Anzahl von Iterationen erreicht ist.

WITHIN CLUSTER SUM OF SQUARES

Within Cluster Sum of Squares (WCSS) ist eine Methode zur Bewertung der Qualität von Clustering-Ergebnissen. Es misst die Summe der quadratischen Abweichungen (Sum of Squares) aller Punkte innerhalb jedes Clusters von dessen Zentrum.

Die Berechnung des WSS beginnt damit, dass jeder Punkt im Datensatz einem bestimmten Cluster zugeordnet wird. Anschließend wird für jedes Cluster das Zentrum berechnet, das als der durchschnittliche Wert aller Punkte innerhalb des Clusters definiert ist. Dann wird die Summe der quadratischen Abweichungen jedes Punktes innerhalb des Clusters von seinem Zentrum berechnet und aufsummiert. Dies wird für jedes Cluster im Datensatz durchgeführt und die Ergebnisse werden addiert, um den Gesamtwert des WCSS zu erhalten.

Eine niedrigere WCSS-Zahl bedeutet, dass die Punkte innerhalb jedes Clusters näher beieinander liegen und somit eine höhere Dichte aufweisen. Das bedeutet, dass das Clustering besser ist, da es klarere und besser definierte Cluster gibt. Eine höhere WCSS-Zahl deutet hingegen darauf hin, dass die Punkte innerhalb jedes Clusters weiter auseinander liegen und somit eine geringere Dichte aufweisen, was darauf hinweisen kann, dass das Clustering schlechter ist.

In der Praxis wird der WCSS oft zusammen mit anderen Metriken wie der Silhouette-Analyse verwendet, um eine fundierte Bewertung der Qualität der Clustering-Ergebnisse zu erhalten und das beste Clustering-Modell auszuwählen.

FAKTORENANALYSE

Eine Faktorenanalyse ist ein statistisches Verfahren, das zur Identifizierung von verborgenen, zugrunde liegenden Variablen oder Faktoren in einem Datensatz verwendet wird. Diese Faktoren können dazu beitragen, die Beziehungen zwischen verschiedenen Variablen zu erklären und somit die Daten auf eine überschaubare Anzahl von Dimensionen zu reduzieren.

Die Funktionsweise einer Faktorenanalyse kann in mehrere Schritte unterteilt werden:

1. Vorbereitung der Daten: Zunächst müssen die Daten aufbereitet werden, um eine sinnvolle Analyse zu ermöglichen. Das kann beinhalten, fehlende Werte zu ergänzen oder Ausreißer zu entfernen.
2. Auswahl der Faktorenanalyse-Methodik: Es gibt verschiedene Arten von Faktorenanalysen, darunter die Hauptkomponentenanalyse (PCA), die Exploratorische Faktorenanalyse (EFA) und die Konfirmatorische Faktorenanalyse (CFA). Je nach Ziel und Struktur der Daten kann die Auswahl der passenden Methode variieren.
3. Bestimmung der Anzahl der Faktoren: Der nächste Schritt besteht darin, die Anzahl der Faktoren zu bestimmen, die in den Daten enthalten sein könnten. Dies kann durch eine Analyse der Eigenwerte oder durch Überprüfung der kumulativen Varianzaufklärung erfolgen.
4. Schätzung der Faktoren: In diesem Schritt werden die Faktoren geschätzt und die Daten werden auf diese Faktoren projiziert. Dies kann durch eine lineare Transformation der Daten erreicht werden.
5. Interpretation der Faktoren: Nach der Schätzung der Faktoren werden diese interpretiert, um ihre Bedeutung zu verstehen. Dies kann beinhalten, die Faktoren zu benennen und zu beschreiben, welche Variablen am stärksten mit jedem Faktor korrelieren.
6. Überprüfung der Ergebnisse: Schließlich müssen die Ergebnisse der Faktorenanalyse auf Validität und Reliabilität überprüft werden. Dies kann durch verschiedene statistische Tests und Techniken erfolgen, um sicherzustellen, dass die Faktoren tatsächlich die Daten auf sinnvolle Weise erklären.

Insgesamt ist die Faktorenanalyse ein leistungsfähiges Werkzeug, um komplexe Datensätze zu analysieren und auf eine überschaubare Anzahl von Faktoren zu reduzieren, die dazu beitragen können, die Beziehungen zwischen Variablen zu erklären und zu verstehen.

E I G E N W E R T E

In der Statistik beziehen sich Eigenwerte auf die charakteristischen Werte einer Matrix. Eine Matrix ist eine Anordnung von Zahlen in Form eines Rechtecks oder Quadrats. Eine quadratische Matrix hat dieselbe Anzahl von Zeilen und Spalten, und ihre Eigenwerte sind eine spezielle Art von Skalaren, die der Matrix zugeordnet sind.

Der Eigenwert einer Matrix ist eine Zahl, die angibt, wie die Matrix auf einen Vektor wirkt. Ein Vektor ist ein Objekt, das eine Größe und eine Richtung hat. Wenn man eine Matrix auf einen Vektor anwendet, multipliziert man den Vektor mit der Matrix und erhält einen neuen Vektor. Der Eigenwert gibt an, um welchen Faktor der neue Vektor in Richtung des ursprünglichen Vektors vergrößert oder verkleinert wird.

In der Statistik werden Eigenwerte häufig verwendet, um die Varianz in einem Datensatz zu erklären. Wenn man beispielsweise eine Faktorenanalyse durchführt, kann man die Eigenwerte der Kovarianzmatrix der Variablen berechnen. Diese Eigenwerte geben an, wie viel Varianz jeder Faktor in den Daten erklärt. Faktoren mit höheren Eigenwerten erklären mehr Varianz in den Daten als Faktoren mit niedrigeren Eigenwerten.

Eigenwerte werden auch in anderen statistischen Methoden verwendet, wie z.B. der Hauptkomponentenanalyse (PCA) und der linearen Regression. In der PCA werden die Eigenwerte verwendet, um zu bestimmen, wie viele Hauptkomponenten in den Daten enthalten sind. In der linearen Regression können die Eigenwerte der Kovarianzmatrix der unabhängigen Variablen verwendet werden, um zu bestimmen, welche Variablen die meisten Informationen zur Vorhersage der abhängigen Variablen liefern.