

Zusammenfassung Dimensionsreduktion & Embeddings

Hauptkomponentenanalyse (PCA)	<ul style="list-style-type: none"> - Ein statistisches Verfahren, um Daten in einfacheren, aber aussagekräftigen Strukturen zu transformieren - Alltagsbeispiel: Von einer Vielzahl Autos das passende auszuwählen - Ziel der PCA: Reduzierung der Komplexität von Datensätzen durch Fokussierung auf die wichtigsten Merkmale
Notwendigkeit von PCA	<ul style="list-style-type: none"> - Bei überwältigenden Daten mit zu vielen Features erschwert die Übersicht - Lösung durch PCA: Reduktion auf wesentliche Merkmale, um Entscheidungen einfacher zu machen
Wie funktioniert die PCA (siehe auch extra Blatt)?	<p>Schritt 1:</p> <ul style="list-style-type: none"> - Darstellung der Merkmale in einem Koordinatensystem <p>Schritt 2:</p> <ul style="list-style-type: none"> - Berechnung des Mittelwerts und Minimierung der Distanz zu einer optimalen Geraden, der ersten Hauptkomponente <p>Schritt 3:</p> <ul style="list-style-type: none"> - Projektion der Datenpunkte auf die Komponentenachse zur Unterscheidung <p>Schritt 4:</p> <ul style="list-style-type: none"> - Iteratives Verfahren - Erzeugung weiterer Komponenten, die orthogonal(senkrecht) zu vorherigen stehen, um alle Dimensionen der Daten zu erfassen
Vorteile und Grenzen der PCA	<p>Vorteile:</p> <ul style="list-style-type: none"> - Ermöglicht Entscheidungsfindung basierend auf den wichtigsten Datenmerkmalen - Vereinfachung von Entscheidungen, Identifizierung der wichtigsten Merkmale, keine Fachkenntnisse erforderlich <p>Grenzen:</p> <ul style="list-style-type: none"> - Nicht immer intuitiv interpretierbare Komponenten - Anwendbarkeit ist begrenzt, wenn interne Streuung gross ist
Multidimensionale Skalierung (MDS) Definition	Multidimensionale Skalierung (MDS) ordnet Datenpunkte so an, dass ihre Abstände möglichst genau den tatsächlichen Unterschieden entsprechen. Dadurch kann man hochdimensionale

	Daten einfach in 2D oder 3D darstellen, ohne die Beziehungen stark zu verzerren.
Multidimensionale Skalierung (MDS)	<ul style="list-style-type: none"> - Ziel von MDS: Vereinfachung komplexer, hochdimensionaler Datensätze in eine verständliche, meist zweidimensionale Darstellung - Hauptgedanke: Bewahrung der relativen Distanzen zwischen den Datenpunkten beim Übergang zu einem niedrigen dimensionierten Raum - Anwendungsbereich: Geeignet für eine Vielzahl von Daten
Wie funktioniert MDS?	<p>Schritt 1: Erstellung der Distanzmatrix</p> <ul style="list-style-type: none"> - Ausgangspunkt für MDS ist eine Distanzmatrix, die die Ähnlichkeiten oder Unähnlichkeiten zwischen den Datenpunkten erfasst - Bspw. Euklidischer Abstand, um die "Distanz" zwischen zwei Punkten im Raum zu berechnen - Diese Distanzen werden für alle möglichen Paare von Datenpunkten berechnet, um eine Distanzmatrix zu erstellen - Wichtigkeit der Distanzmatrix: Sie bildet die Grundlage für die Anordnung der Punkte in einem neuen Raum <p>Schritt 3: Multidimensionale Skalierung</p> <ul style="list-style-type: none"> - Schrittweise Optimierung: Anpassung der Punkte im Zielraum, um die originalen Distanzen so genau wie möglich abzubilden - Iterativer Ansatz: MDS passt die Platzierung der Datenpunkte schrittweise an, um eine optimale Anordnung zu erreichen - Visualisierung: Ergebnis ist eine grafische Darstellung, die die relativen Distanzen im niedrigen dimensionierten Raum zeigt
MDS vs. PCA	<ul style="list-style-type: none"> - Dateneigabe: MDS arbeitet direkt mit Distanzmatrizen, während PCA tabellarische Daten benötigt - Zielsetzung: PCA fokussiert auf die Maximierung der Varianz entlang der Hauptkomponenten. MDS zielt auf die Bewahrung der originalen Distanzen ab - Ansatz: PCA verwendet einen linear-projektiven Ansatz. MDS nutzt einen iterativen Optimierungsprozess. - Herausforderungen bei MDS: Die Distanzen im niedrigen dimensionierten Raum können niemals vollständig die Komplexität des hochdimensionalen Raums erfassen - Bedeutung der Achsen: Bei PCA haben die Achsen statistische Bedeutung. Bei ;DS

	<p>haben die Koordinaten keine inhärente Bedeutung</p> <ul style="list-style-type: none"> - Vorteil von MDS: Kann Distanzmatrizen direkt verarbeiten und ist flexibel in der Anwendung auf verschiedenen Datentypen
t-SNE Definition	<p>t-SNE stellt Datenpunkte so dar, dass ähnliche Punkte nah beieinander und unähnliche weit entfernt sind, wobei es sich auf lokale Nachbarschaften konzentriert.</p> <p>Es eignet sich besonders gut, um Cluster in komplexen, hochdimensionalen Daten sichtbar zu machen.</p>
t-SNE vs. MDS	<ul style="list-style-type: none"> - Bessere Gruppierung: t-SNE neigt dazu, Datenpunkte ähnlicher Typen enger und klarer zu gruppieren als MDS. - Wichtig bei grossen Datensätzen: Der Unterschied in der Gruppierungsleistung zwischen t-SNE und MDS wird mit der Zunahme der Datensatzgrösse signifikanter. - Fokus auf Lokalität: Während MDS die globalen Distanzen zu bewahren sucht, fokussiert t-SNE auf die Bewahrung der lokalen Nachbarschaftsbeziehungen. - Flexibilität bei komplexen Datensätzen: t-SNE zeigt besonders bei Datensätzen mit komplexen, nicht-linearen Strukturen Vorteile. - Erkenntnisgewinnung: Ermöglicht tiefere Einblicke in die Struktur der Daten durch klare visuelle Trennung von Clustern. - Breites Anwendungsspektrum: Von der genetischen Forschung und Proteinanalyse bis hin zu Kundenverhaltensanalysen.
Embeddings	<ul style="list-style-type: none"> - Embeddings sind fortgeschrittene Techniken zur Dimensionalitätsreduktion, die es ermöglichen, hochdimensionale Daten in einen Raum geringerer Dimensionen zu "betten", wobei die wesentlichen Strukturen und Beziehungen erhalten bleiben. - Embeddings bieten eine leistungsstarke Methode zur Vereinfachung und Analyse komplexer Datensätze, indem sie intuitive, visuelle Einsichten in die Datenstruktur ermöglichen.
Embeddings Anwendungsbereiche	<ul style="list-style-type: none"> - Natursprachverarbeitung (NLP): In NLP ermöglichen Word Embeddings eine effektive Kodierung der semantischen Bedeutung von Wörtern in dichten

	<p>Vektorräumen, was für Aufgaben wie Textklassifikation, Sentiment-Analyse und maschinelle Übersetzung genutzt wird.</p> <ul style="list-style-type: none"> - Bildverarbeitung: In der Bildverarbeitung werden Embeddings genutzt, um Bilder in einen Vektorraum zu kodieren, wodurch Algorithmen des maschinellen Lernens effektiver Muster erkennen und Klassifikationen durchführen können. - Empfehlungssysteme: Embeddings finden Anwendung in Empfehlungssystemen, indem sie Nutzer und Produkte in denselben Vektorraum projizieren, was die Entdeckung von Ähnlichkeiten und die Generierung personalisierter Empfehlungen ermöglicht.
Word Embeddings Definition	<ul style="list-style-type: none"> - Techniken, um Wörter in Zahlen umzuwandeln - Jedes Wort wird als Vektor in einem hochdimensionalen Raum repräsentiert
Ziel, Möglichkeiten, Zukünftige Entwicklungen	<p>Ziel:</p> <ul style="list-style-type: none"> - Transformation von Text in eine Form, die von maschinellen Lernmodellen verarbeitet werden kann - Bewahrung semantischer Bedeutungen zwischen Wörtern, sodass ähnliche Wörter nahe beieinander im Vektorraum liegen <p>Möglichkeiten von Word Embeddings:</p> <ul style="list-style-type: none"> - Ermöglichen die Anwendung maschineller Lerntechniken auf Textdaten - Können Feature-Extraktion in komplexen NLP-Aufgaben wie Sentiment-Analyse, Textklassifikation und mehr verwendet werden <p>Zukünftige Entwicklungen:</p> <ul style="list-style-type: none"> - Weiterentwicklung der Modelle zur besseren Erfassung von Feinheiten in der Sprache und Reduktion von Ambiguitäten (Mehrdeutigkeiten)
SBERT	<ul style="list-style-type: none"> - Eine Modifikation des BERT-Modell, die für die Einbettung ganzer Sätze optimiert ist - Trainiert, um Text in fixierte Längen von Vektoren umzuwandeln, während semantische Bedeutungen beibehalten werden - SBERT berücksichtigt den Kontext von Wörtern in Sätzen, kann jedoch unerwartete Ergebnisse liefern, wenn Wörter ausserhalb ihres typischen Kontextes betrachtet werden

	<ul style="list-style-type: none">- Jedes Wort wird durch 384 numerische Werte repräsentiert, die seine Position im Vektorraum definieren
FastText	<ul style="list-style-type: none">- FastText ist eine Erweiterung von Word2Vec, entwickelt von Facebook, die auch Subwortinformationen berücksichtigt- FastText ist auf Worte optimiert- Jedes Wort wird als Vektor mit 300 Merkmalen dargestellt