DHBW Lörrach

Business Intelligence

Vert. Prof. Dr. Aikaterini Nakou

DHBW Lörrach

Statistische Tests

- Testen von Hypothesen (Annahmen, Behauptungen)
- Statistischer Test: Verfahren, mit dessen Hilfe sich bestimmte Hypothesen auf ihre Richtigkeit hin überprüfen lassen.
- Statistische Testverfahren basieren auf Stichprobentheorie
- **Ziel:** Richtigkeit von Aussagen über die Verteilung einer Zufallsvariablen überprüfen.

Statistische Tests: Hypothesen

… für Unterschiede

Nullhypothese (H_0):

- besagt, dass es keinen Mittelwertunterschied zwischen den Gruppen gibt

$$\mu_1 = \mu_2$$

Alternativhypothese (H_1):

- besagt, dass es einen Mittelwertunterschied zwischen den Gruppen gibt

$$\mu_1 \neq \mu_2$$

… für Zusammenhänge

Nullhypothese (H_0):

- besagt, dass es keinen Zusammenhang zwischen den Merkmalen gibt

$$\rho = 0$$

Alternativhypothese (H_1):

- besagt, dass es einen Zusammenhang zwischen den Merkmalen gibt

$$\rho \neq 0$$

DHBW Lörrach

Signifikanztest

- Um entscheiden zu können, ob das Stichprobenergebnis mit der Nullhypothese übereinstimmt, muss eine Grenze für die beschriebene Wahrscheinlichkeit festgelegt werden.

- Dieser Wahrscheinlichkeitsgrenze wird fest vorgeben. Die **Wahrscheinlichkeitsgrenze** wird als **Irrtumswahrscheinlichkeit** α bezeichnet.

- Normalerweise Irrtumswahrscheinlichkeit $\alpha=0.05$

DHBW Lörrach

Schritte- Statistische Tests

1. Formulierung von Null- und Alternativhypothese
2. Bestimmung von Skalenniveau und Verteilungstyp der Merkmalswerte
3. Wahl der entsprechenden Signifikanztests
4. Festlegung der Irrtumswahrscheinlichkeit α, z.B. 0,05
5. Berechnung der Signifikanztests
6. Vergleichen mit der Irrtumswahrscheinlichkeit
7. Ablehnung oder Behalten der Null-Hypothese

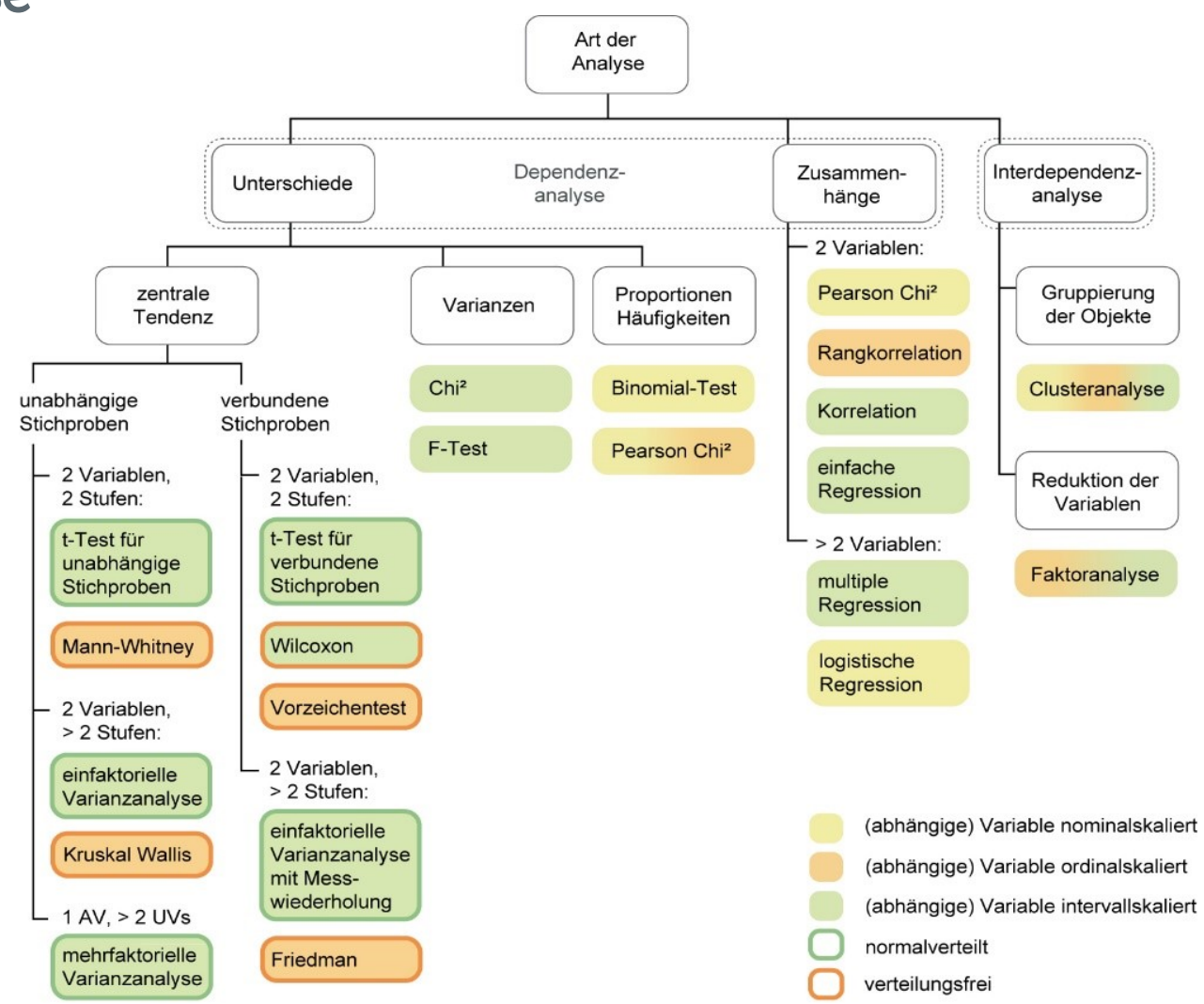
DHBW Lörrach

Signifikanztests

- Bei **parametrischen Signifikanztests** werden angewendet, wenn die die Daten metrisch skaliert sind, da sonst Operationen wie die Berechnung des arithmetischen Mittels oder der Standardabweichung nicht zulässig sind.

- **Nichtparametrische Tests** werden angewendet, wenn die Daten lediglich nominal oder ordinal skaliert sind, oder zwar metrische Daten vorliegen, aber sonstige Voraussetzungen der entsprechenden parametrischen Verfahren nicht erfüllt werden.

Statistische Analyse



Korrelation

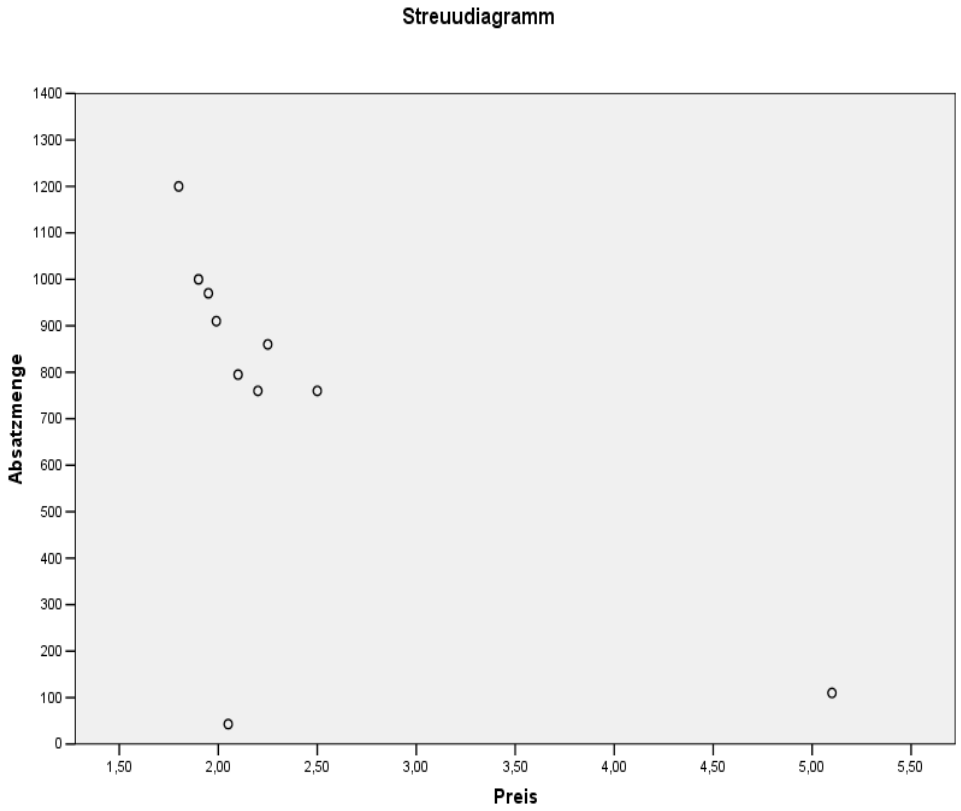
- die Stärke des Zusammenhangs wird durch den Korrelationskoeffizienten r gemessen
- r liegt stets in den Grenzen von -1 bis +1

- für die Stärke des Zusammenhangs ist allein der Betrag des Korrelationskoeffizienten maßgebend

- das Vorzeichen gibt an, ob der Zusammenhang gleichläufig (+) oder gegenläufig (–) ist

Korrelationskoeffizient	Einstufung
$\lvert r \rvert \leq 0.25$	schwache Korrelation
$0.25 < \lvert r \rvert \leq 0.66$	mittlere Korrelation
$0.66 < \lvert r \rvert < 1$	starke Korrelation
$\lvert r \rvert = 1$	perfekte Korrelation

Korrelationskoeffizient nach Pearson

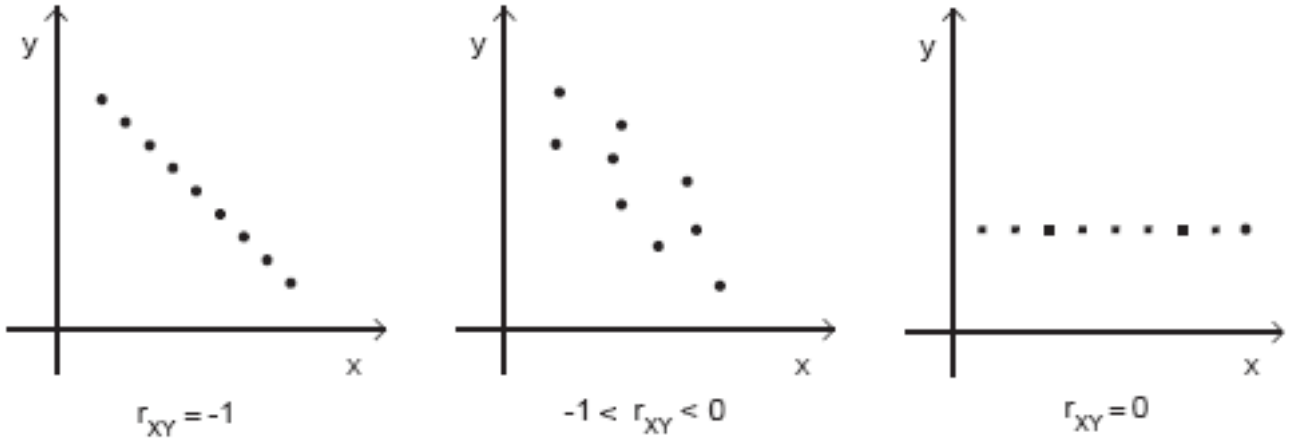


Vermutung:

Zwischen den Variablen Preis und Verkaufsmenge besteht ein linearer und gegenläufiger Zusammenhang;

je höher der Verkaufspreis umso geringer die Absatzmenge.

Korrelationskoeffizient nach Pearson



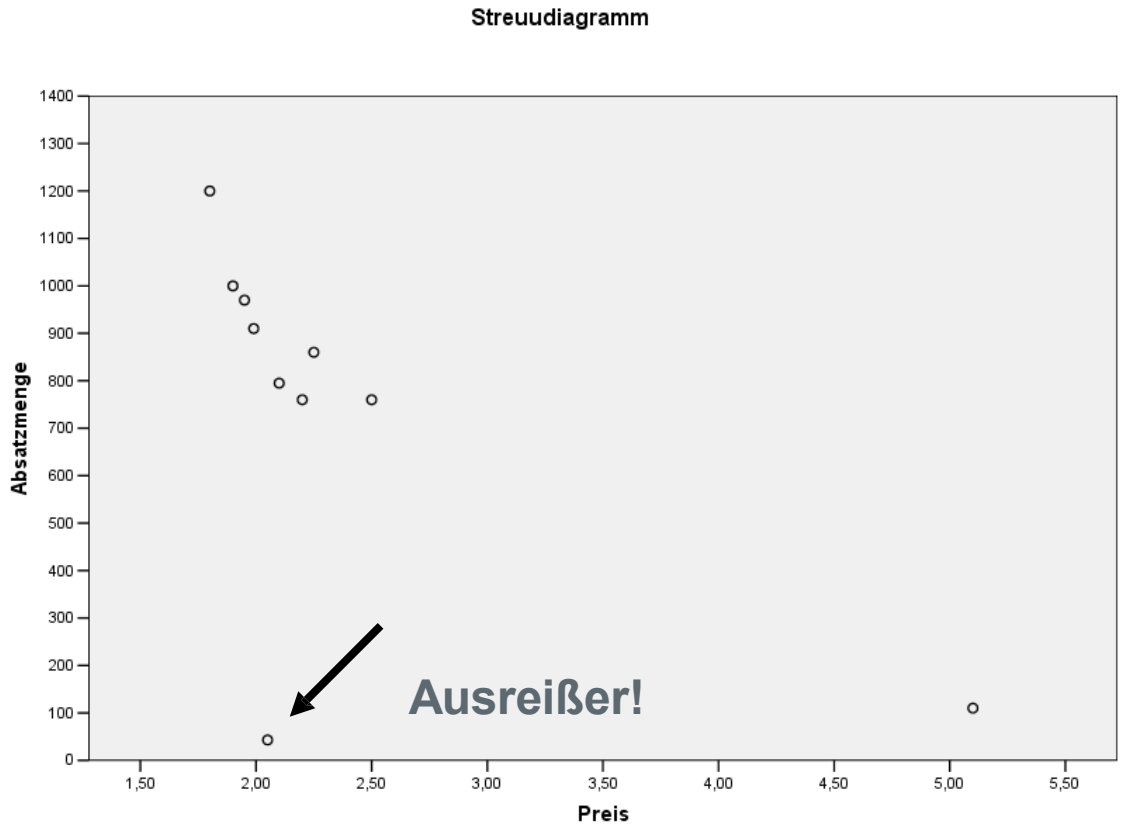
Korrelationskoeffizient nach Pearson

r misst den linearen Zusammenhang zweier Variablen

- **Konsequenz**: einzelne Ausreißer, d.h. einzelne extreme Datenpunkte, können einen starken, unerwünschten Effekt auf den numerischen Wert von r haben; hohe Korrelationen können als gering erscheinen und umgekehrt.

- **Lösung**: Ermittlung von Rangkorrelationskoeffizienten, die von Ausreißern wesentlich weniger beeinflusst werden, da ihre Ermittlung auf den Rängen der Beobachtungen basiert.

Rangkorrelationskoeffizient nach Spearman



Rangkorrelationskoeffizient nach Spearman

- wird zwischen zwei Variablen berechnet, die mindestens ordinalskaliert sind; für metrisch skalierte Variablen, bei Unsicherheit hinsichtlich der Normalverteilungsanahme misst die Stärke des monotonen Zusammenhangs

- basiert auf Rangzahlen, die den Messwerten zugeordnet sind für beide Variablen wird eine Rangreihe der Werte erstellt, dem höchsten Wert wird der Rangplatz 1 verliehen; bei gleichen Werten werden gemittelte Rangplätze vergeben

- die Differenz d der zugehörigen Rangplatzpaare wird bestimmt

Rangkorrelationskoeffizient nach Spearman

- Wertebereich von rs: -1 bis +1
- gehen mit steigenden x-Werten auch steigende y-Werte einher, so nimmt rs tendenziell einen großen Wert an
- sind die Rangzahlen bei den Merkmalen beider Variablen völlig gleich, so nimmt rs den Wert 1 an (die Rangpaare liegen auf einer Geraden mit positiver Steigung liegen)
- bei entgegengesetzt laufenden Rangzahlen wird rs = -1 (die Rangpaare liegen auf einer Geraden mit negativer Steigung)

Rangkorrelationskoeffizient nach Spearman

Rechenschritte zur Rangkorrelation nach Spearman rs :

Tiefkühlpizza	A	B	C	D	E	F	G	H	I	J
Preis in Euro	5,10	1,80	2,10	2,05	1,99	1,90	2,20	1,95	2,50	2,25
Absatzmenge im Monat	110	1200	100	43	910	1000	760	970	685	860
Rang Preis	1	10	5	6	7	9	4	8	2	3
Rang Absatzmenge	8	1	9	10	4	2	6	3	7	5
d_i	-7	9	-4	-4	3	7	-2	5	-5	-2
d^2_i	49	81	16	16	9	49	4	25	25	4

Rangkorrelationskoeffizient nach Spearman

Es ergibt sich

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d^2}{n(n^2 - 1)} \qquad r_s = 1 - \frac{6 * 278}{10 * (100 - 1)} = -0{,}685$$

Rangkorrelationskoeffizient nach Spearman

Interpretation des Ergebnisses

rs= -0,685

=> starker Zusammenhang

rs< 0 => gegenläufiger monotoner Zusammenhang

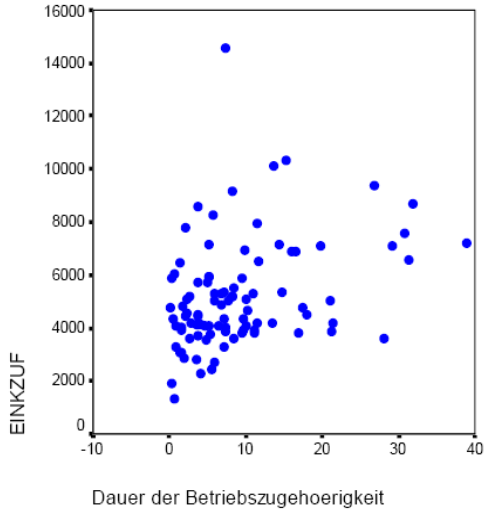
Es zeigt sich ein mittlerer gegenläufiger Zusammenhang zwischen Preis und Absatzmenge: Je höher der Preis einer Tiefkühlpizza, umso niedriger ist die verkaufte Menge an Tiefkühlpizzen.

Kontigenzkoeffizient

- Maß zur Charakterisierung des Zusammenhangs zweier diskreter Variablen
- Da hier nur Häufigkeiten untersucht werden spricht man von Kontingenz statt Korrelation.
- C kann nur positive Werte annehmen mit den theoretischen Grenzen 0 und 1.
- Der Maximale Kontingentkoeffizient ergibt sich aus der Anzahl Spalten und Zeilen einer Kontingenztabelle.

Probleme

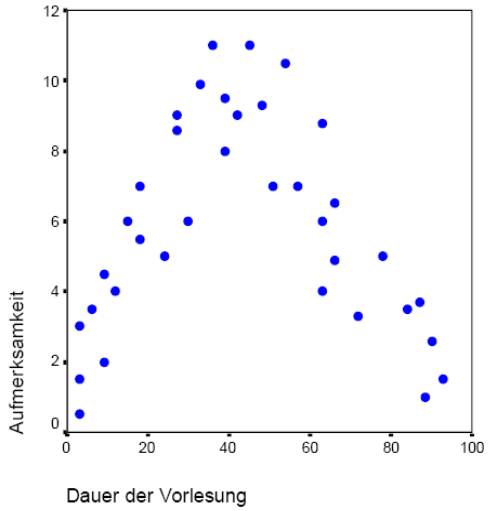
für die Korrelation i.e.S gilt: Einzelne Fälle können einen starken Einfluss auf den Korrelationskoeffizienten ausüben.



Korrelation über alle Fälle: r=0,35.

Korrelation ohne Einkommen über 14.000: r=0,39.

Korrelationen lassen sich für alle Funktionstypen berechnen; allerdings werden nur monotone bzw. lineare Zusammenhänge erfasst.



Korrelation: -0,05, d.h. praktisch gleich null.

Das Beispiel ist fiktiv, Sie sind natürlich ständig aufmerksam! ☺

DHBW Lörrach

Probleme

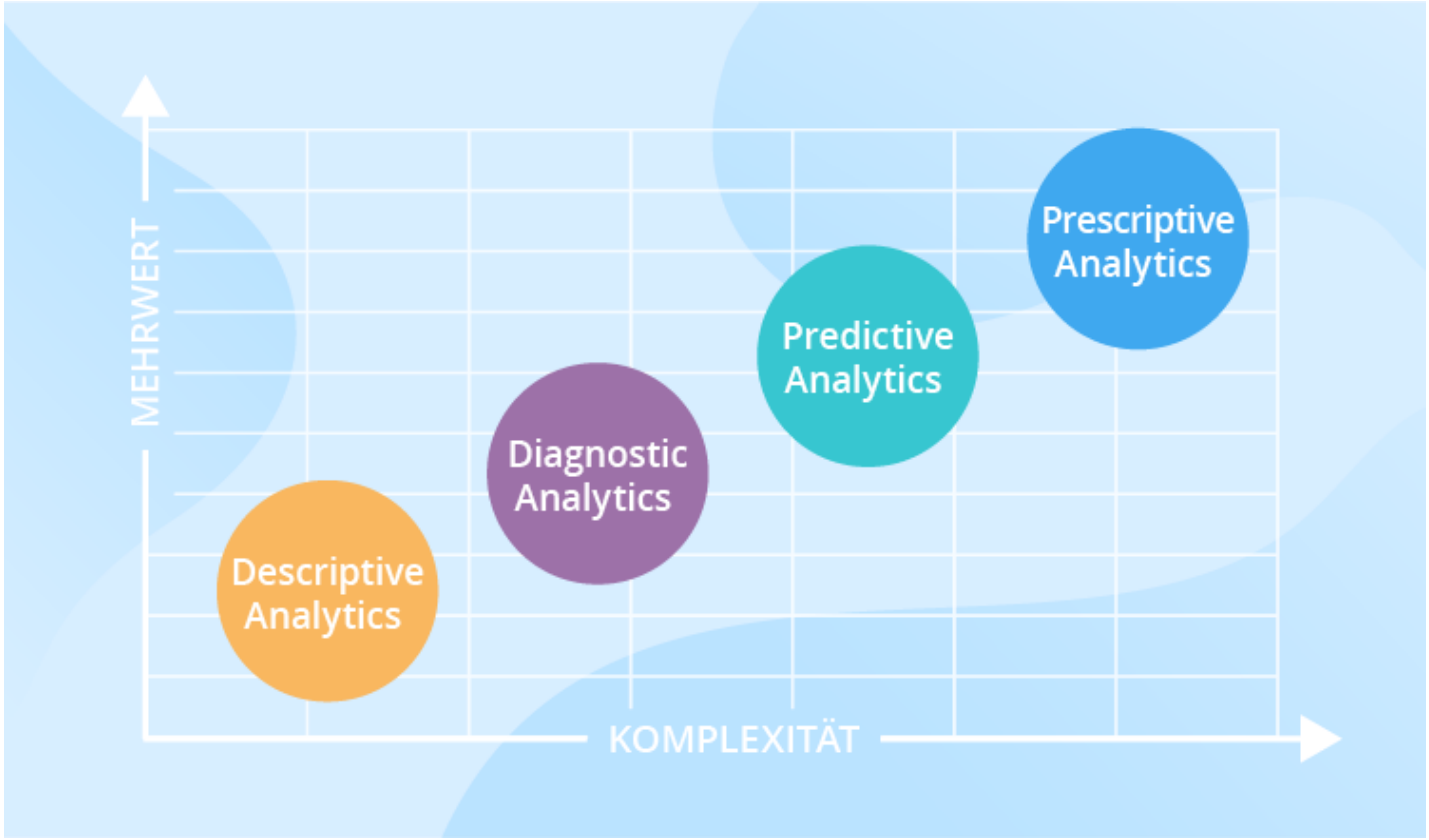
- Kausalzusammenhänge können nicht erfast werden

- Scheinkorrelationen (Korrelation zwischen Merkmalen, die inhaltlich nicht gerechtfertigt ist) können auftreten; Zusammenhänge ergeben sich dann, wenn ein mit beiden beobachtbaren Merkmalen hochkorreliertes drittes Merkmal übersehen wird und unberücksichtigt bleibt.

DHBW Lörrach

Datenanalyse

Methoden der Datenanalyse



DHBW Lörrach

Datenanalyse

Prädiktive Analyse

Sie ist eine fortgeschrittene Analysemethode, die aktuelle und historische Daten verwendet, um Vorhersagen über zukünftige oder sonst unbekannte Ereignisse zu treffen.

Ziel: Ein Modell / Algorithmus zu erstellen, das eine Zielvariable so genau wie möglich bestimmt oder vorhersagt.

Form der Zielvariablen:

- Regressionsprobleme
- Klassifikationsprobleme

Datenanalyse

Prädiktive Analyse

Um Vorhersagen treffen zu können, werden die Ergebnisse bereits durchgeführter deskriptiver oder diagnostischer Analysen verwendet.

Beispiele:

- Vorhersage der zukünftigen Umsätzen
- Bestimmung des Personalbedarfs
- Risikoanalyse
- etc…

Quelle: IONOS: Deep learning vs. machine learning – what are the differences?

DHBW Lörrach

Künstliche Intelligenz

KI ist ein Mix aus vielen **verschiedenen Technologien**. Sie **befähigt Maschinen** dazu, mit menschenähnlicher Intelligenz zu **verstehen, zu handeln und zu lernen**.

Machine Learning ist ein Teilbereich von KI, der es Computern ermöglicht, **aus Erfahrungen zu lernen**, ohne explizit programmiert zu werden.

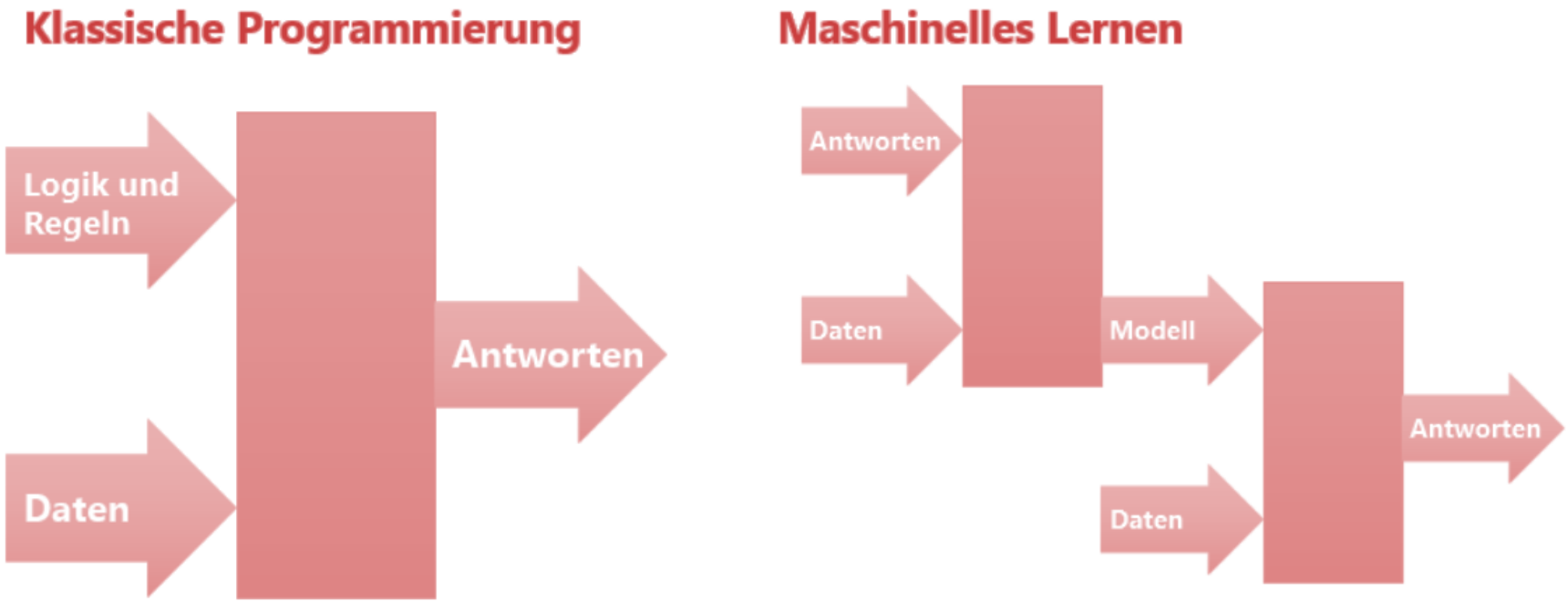
Deep Learning ist eine fortgeschrittene Form des Machine Learning, die auf **künstlichen neuronalen Netzen** basiert.

DHBW Lörrach

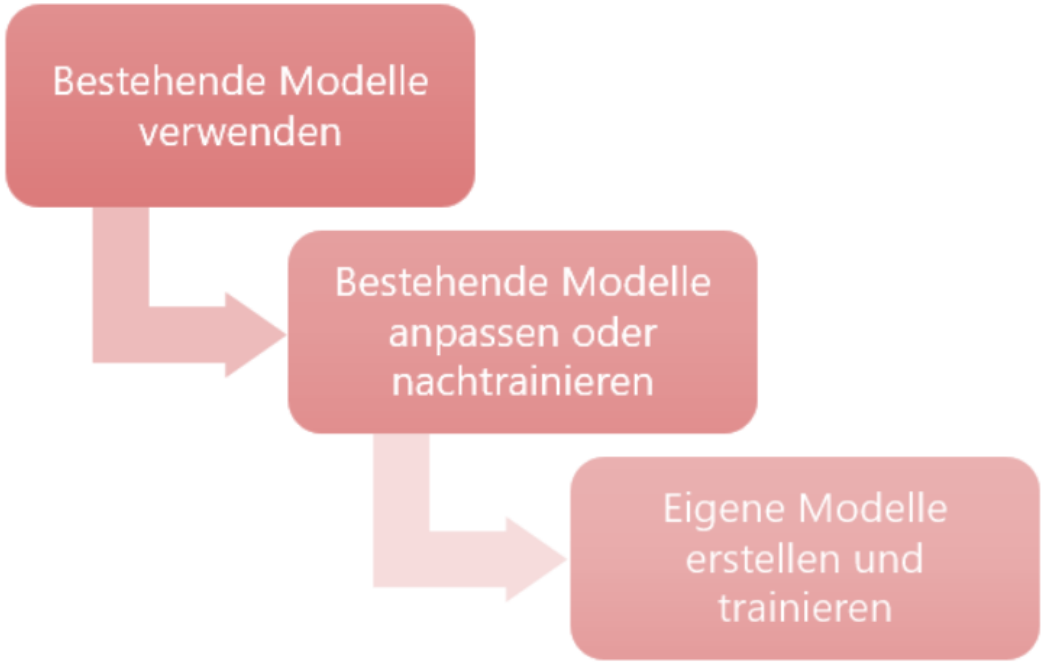
Machine Learning

- Maschinelles Lernen ist der elementare Bereich künstlicher Intelligenz
- mit Hilfe von modellbasierten, flexiblen Algorithmen können eigenständig Lösungen für neue und unbekannte Probleme gefunden werden
- Erstellung von Systemen, die aus Daten selbstständig lernen können

Machine Learning



Machine Learning



DHBW Lörrach

Machine Learning - Herausforderungen

Datenqualität und -verfügbarkeit: Hochwertige Daten sind entscheidend für effektives maschinelles Lernen. Die Beschaffung, Reinigung und Pflege von Daten kann zeitaufwändig sein, und in einigen Fällen sind möglicherweise nicht genügend Daten verfügbar, um ein Modell effektiv zu trainieren.

Repräsentativität der Daten: Die Daten, auf denen ein Modell trainiert wird, müssen eine repräsentative Stichprobe des zugrunde liegenden Problems darstellen. Wenn die Daten nicht repräsentativ sind, kann das Modell falsche Annahmen treffen und schlechte Vorhersagen treffen.

Typen von Lernalgorithmen (ML)

Überwachtes Lernen

- Werden mit sowohl Eingabe- als auch Ausgabedaten trainiert
- **Ziel**: Zuordnung oder Beziehung zwischen den Eingaben und den entsprechenden Ausgaben zu lernen
- **Anwendungen**: Vorhersage, Klassifizierung

Unüberwachtes Lernen

- Werden nur mit Eingabedaten trainiert.
- **Ziel**: Das Modell sucht nach Mustern und Strukturen in den Daten, ohne auf eine spezifische Ausgabe hingewiesen zu werden.
- **Anwendungen**: Clusteranalyse, Dimensionsreduktion

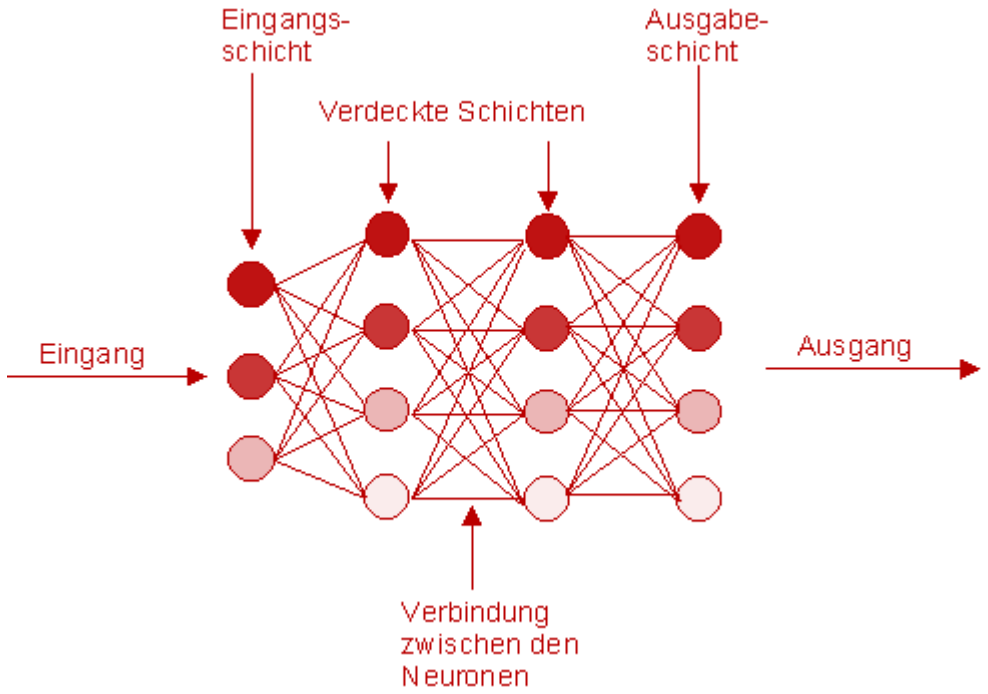
Bestärkendes Lernen

- Lernt durch Belohnung/Bestrafung →Belohnt werden die guten Aktionen, bestraft werden die schlechte Aktionen
- **Ziel**: Durch das Optimieren der Belohnungsfunktion lernt das Modell, die besten Aktionen für bestimmte Situationen zu wählen.
- **Anwendungen**: Prozessoptimierung, Robotik

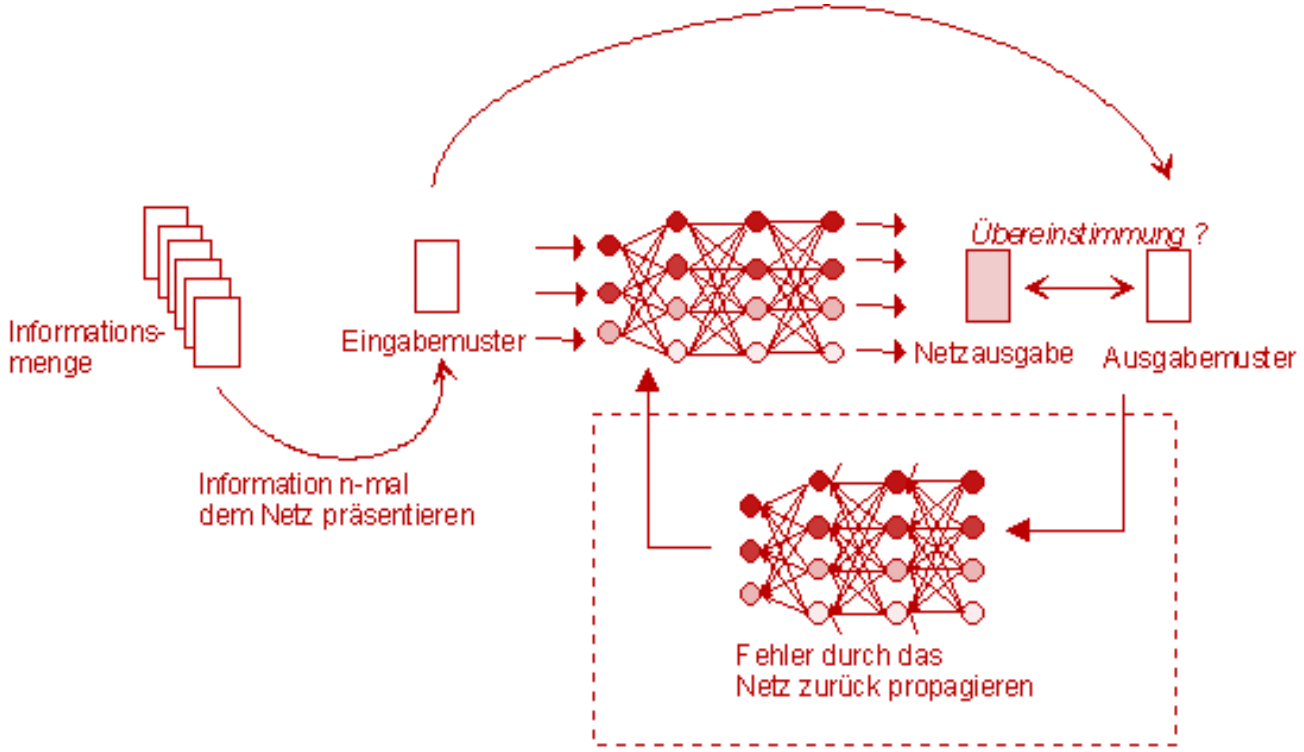
Deep Learning

Deep Learning ist ein Teilbereich des Maschinellen Lernens.

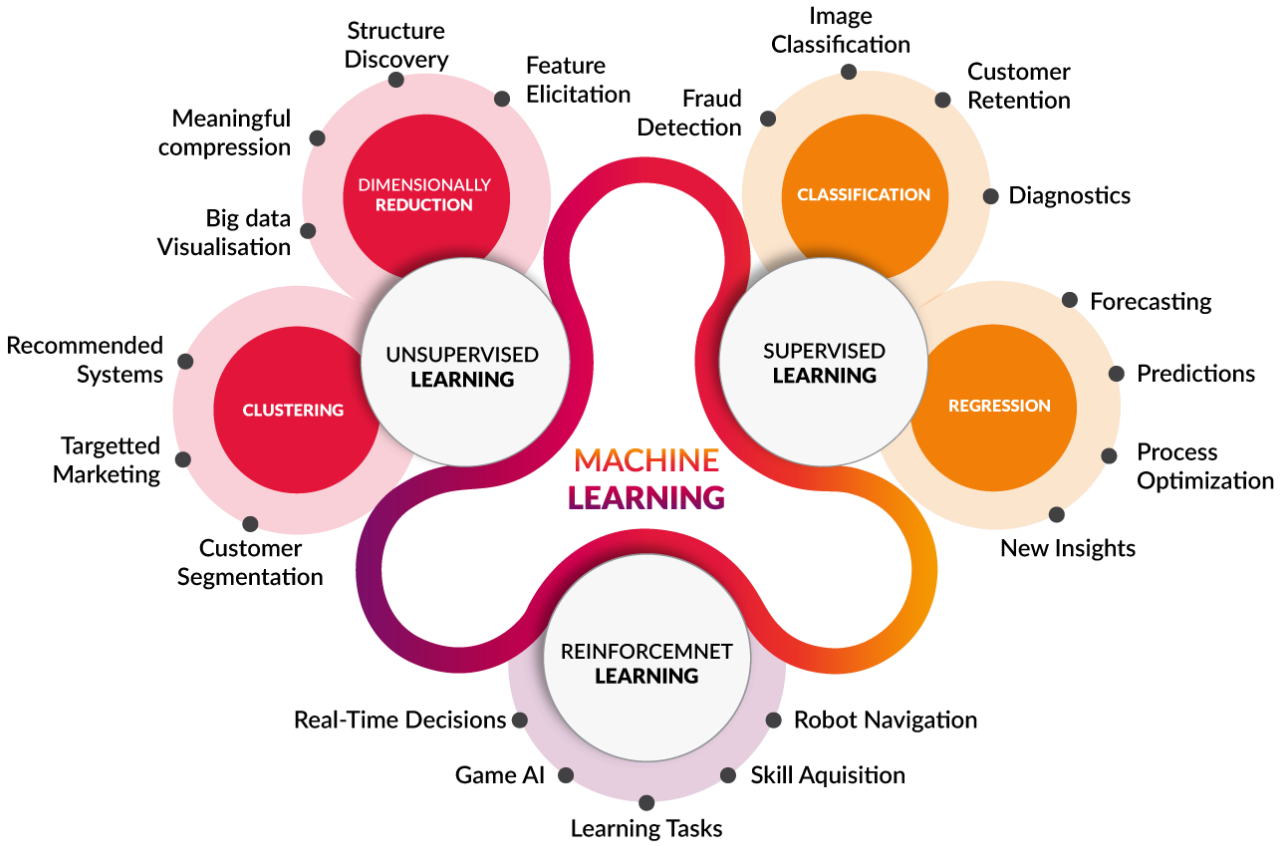
Grundlage sind Neuronale Netze mit mindestens einem Hidden-Layer



Deep Learning

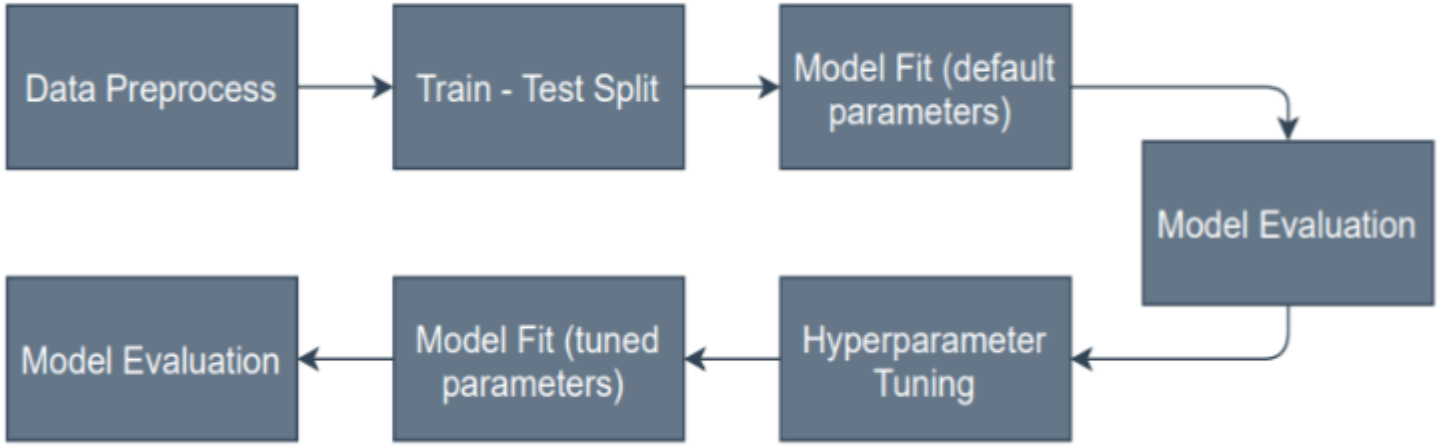


Anwendungsfälle KI



DHBW Lörrach

Modellbildung Machine Learning



DHBW Lörrach

Merkmalsauswahl

Die **Merkmalsauswahl** beinhaltet die Identifizierung der **wichtigsten Merkmale** für das **Modell**.

Dies kann durch:

- statistische Methoden wie Korrelationsanalyse oder

- iteratives Vorgehen wie „forward/backward selection“ erfolgen.

Eine gute Merkmalsauswahl kann die **Modellgenauigkeit verbessern** und **Overfitting verhindern**.

Einfache Lineare Regression

Repräsentation der Punktwolke durch eine Gerade der allgemeinen Form:

$$Y = b0 + b1 * X$$

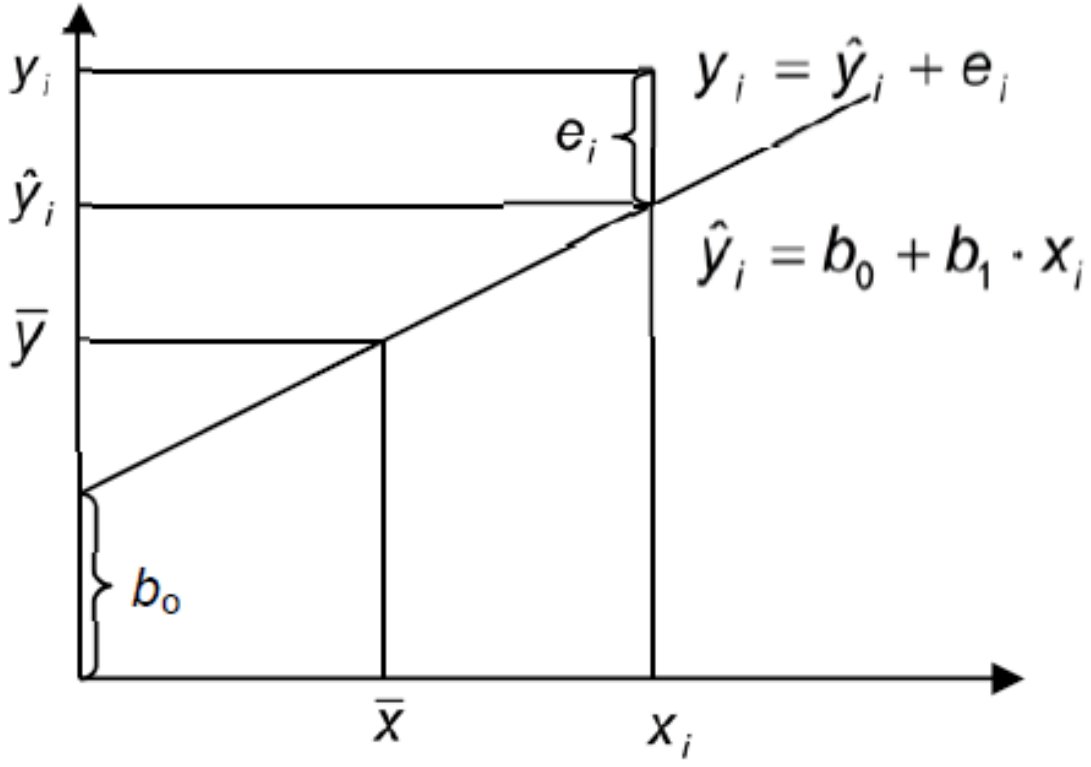
Dabei stehen:

- *y* für die abhängige Variable,
- *x* für die unabhängige Variable,
- *b*0 für den Schnittpunkt der Geraden mit der *y*-Achse des Koordinatensystems
- *b*1 für die Steigung der Geraden, auch Regressionskoeffizient genannt

Regressionsgerade

- Zur Berechnung der Geraden werden in ein Koordinatensystem die Wertepaare übertragen und eine Punktwolke zeigen.
- Legt man nun rein graphisch irgendeine Gerade hinein, so sind stets Abweichungen der Einzelwerte *yi* von der Geraden festzustellen.
- Diese Abweichungen werden als Residuen *ei* bezeichnet.

Darstellung- Regressionsgerade



Regressionsgerade

- Damit das Datenmaterial durch die Regressionsgerade möglichst gut repräsentiert wird, muss die Abweichung der Einzelwerte yi von der Geraden minimiert werden.
- Ein Kriterium für die beste Anpassung der Regressionsgerade an die Beobachtungen muss gefunden werden.
- Methode der kleinsten Quadrate vorgestellt werden, die die Quadratsumme der Residuen minimiert.

Methode der kleinsten Quadrate

Die Regressionsgerade ist diejenige Gerade, die die Summe der quadrierten Residuen (Abweichungen, Vorhersagefehler) minimiert.

Es gilt:

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - (b_0 + b_1 \cdot x_i)$$

$$e_i^2 = [y - (b_0 + b_1 \cdot x_i)]^2$$

Gefordert ist:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y - (b_0 + b_1 \cdot x_i)]^2 \rightarrow Min$$

DHBW Lörrach

Methode der kleinsten Quadrate

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_j - \bar{x})^2} = \frac{\sum_{j=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i\right) \Big/ n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}$$

$$b_1 = \frac{Summe\ der\ Abweichungsprodukte_{xy}}{Summe\ der\ Abweichungsquadrate_{xy}} = \frac{SP_{xy}}{SQ_{xy}}$$

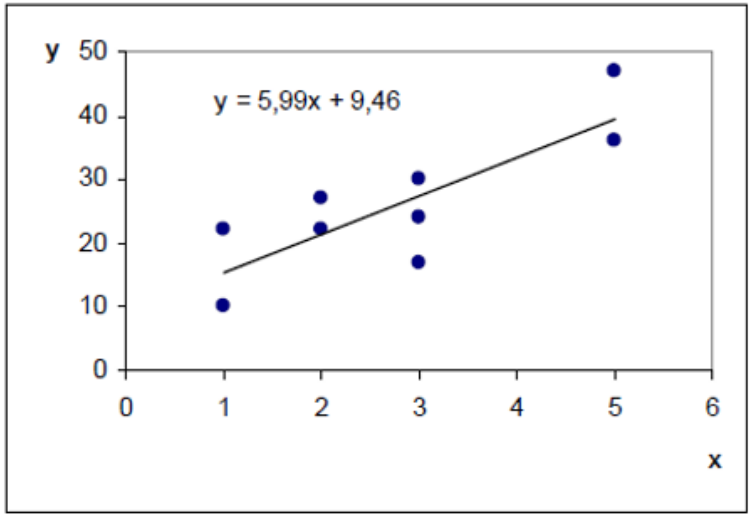
Beispiel

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*
Schnittpunkt	9,4618	4,8596	1,9471	0,0926
X Variable 1	5,9937	1,5630	3,8347	0,0064

Die Regressionsgerade lautet damit:

$$\hat{y} = b_0 + b_1 \cdot x = 9{,}4618 + 5{,}9937 \cdot x$$

In Worten: Ändert sich die Einflussgröße x um eine Einheit, so ändert sich die Zielgröße y um 5,9937 Einheiten. Ist die Einflussgröße = 0, so beträgt der Wert der Zielgröße = 9,4618.



Anpassungsgüte

Den Anteil der durch die Regression erklärten Streuung an der Gesamtstreuung bezeichnet als **Bestimmtheitsmaß r^2**:

$$r^2 = \frac{SQ_{Reg}}{SQ_{Ges}} = \frac{b_1 \cdot \left[\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})\right]}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = b_1 \cdot \frac{SP_{xy}}{SQ_{xy}} = b_1^2 \cdot \frac{SQ_x}{SQ_y} \qquad ($$

Für das obige Beispiel folgt:

$$r^2 = \frac{5{,}9937 \cdot 105{,}2222}{930{,}8889} = 0{,}6775$$

Was sind Zeitreihen?

Eine **Zeitreihe** ist eine **sequenzielle** Abfolge von Datenpunkten, die in **regelmäßigen Zeitintervallen gemessen oder beobachtet** werden.

Zeitreihen finden Anwendung in verschiedenen Bereichen wie Wirtschaft, Klimatologie, Finanzwesen und mehr.

Quelle: Hayes S., 2021: Finding Seasonal Trends in Time-Series

Komponenten einer Zeitreihe

Trendkomponente

- langfristigen Verlauf der Daten
- Auf- oder Abwärtstrend im Laufe der Zeit

Saisonale Komponente

- wiederkehrende Muster oder Schwankungen, die sich innerhalb eines bestimmten Zeitraums wiederholen
- jährliche, monatliche oder wöchentliche Schwankungen

Restkomponente

- zufällige oder unvorhersehbare Variationen in den Daten
- durch unvorhersehbare Ereignisse oder Messfehler

Was ist Zeitreihenanalyse?

Die Zeitreihenanalyse beschäftigt sich mit statistischen Methoden zur **Analyse und Modellierung** einer geordneten Folge von Beobachtungen (**Zeitreihe**).

Diese Modellierung führt zu einem Prozessmodell für das System, das die Daten erzeugt hat.

Somit können anhand dieses Modells **zukünftige Ereignisse vorhergesagt** werden.

Quelle: Özen A., 2021: Seasonality Analysis and Forecast in Time Series

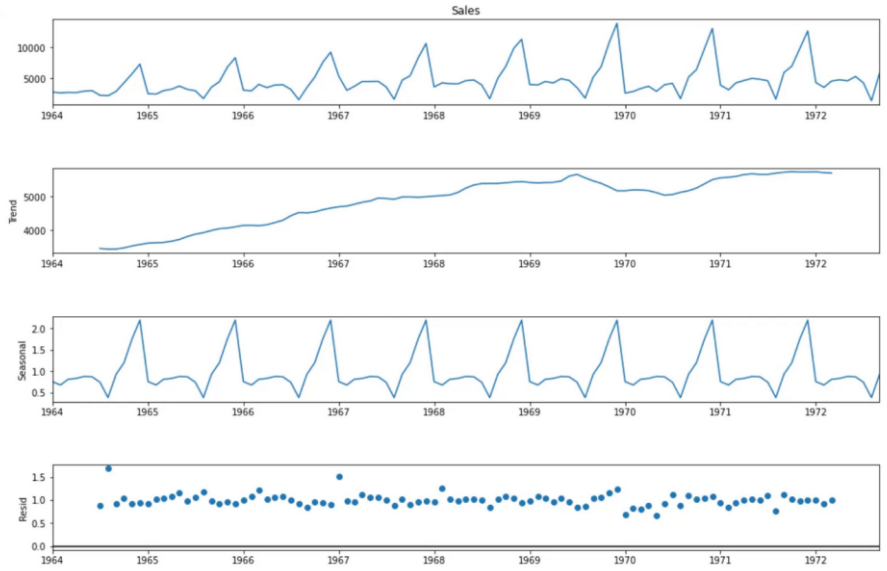
Decomposition - Zerlegung

Bei der Zeitreihenzerlegung können:

- **additive** oder
- **multiplikative**

Bei **additiven** Modellen wird die Zeitreihe als **Summe** von Trend, saisonaler Komponente und Restkomponente dargestellt.

Bei **multiplikativen** Modellen wird die Zeitreihe als **Produkt** dieser Komponenten modelliert.



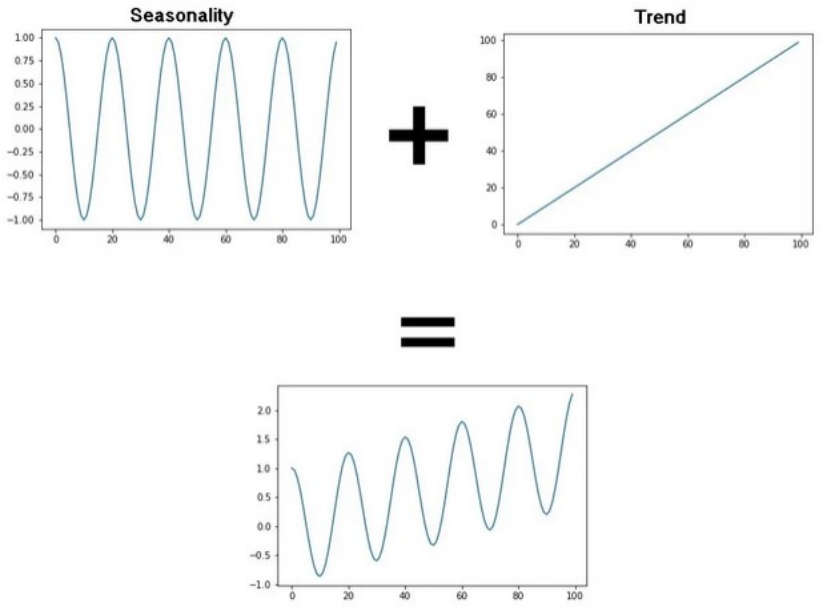
Quelle: Özen A., 2021: Seasonality Analysis and Forecast in Time Series

Additive Komponentenmodelle

Das additive Modell für eine Zeitreihe y(t) kann folgendermaßen dargestellt werden:

y(t) = Trendkomponente(t) + saisonale Komponente(t) + Restkomponente(t)

Additive Modelle eignen sich insbesondere dann gut, wenn die *saisonalen Schwankungen* in der Zeitreihe **nicht proportional zum Trend** sind und eher als additive Effekte auftreten.



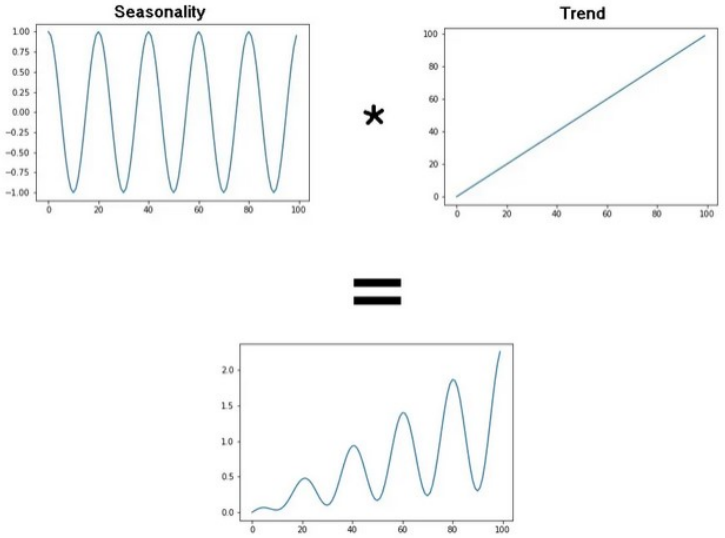
Quelle: Hayes S., 2021: Finding Seasonal Trends in Time-Series

Multiplikative Komponentenmodelle

Das multiplikative Modell für eine Zeitreihe y(t) kann folgendermaßen dargestellt werden:

y(t) = Trendkomponente(t) * saisonale Komponente(t) * Restkomponente(t)

Multiplikative Modelle eignen sich besonders gut, wenn die **saisonalen Effekte relativ** zur **Trendgröße** variieren, beispielsweise wenn die saisonalen Schwankungen mit zunehmendem Trend stärker werden.



Quelle: Hayes S., 2021: Finding Seasonal Trends in Time-Series

DHBW Lörrach

Methoden zur Prognostizierung

Statistische Verfahren

Komponenten-modelle

Überwachtes lernen

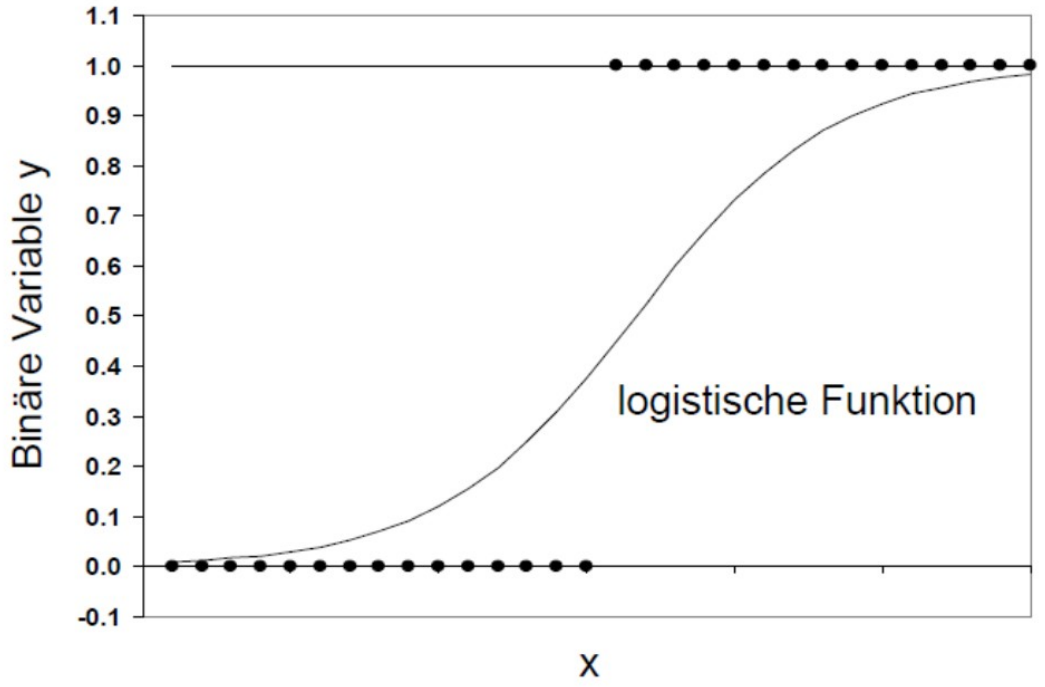
Regression
Generallisierte additive Modelle(GAM)

Deeplearning

RNN LSTM

Logistische Regression

Die (binär) logistische Regressionsanalyse testet, ob ein Zusammenhang zwischen mehreren unabhängigen und einer binären abhängigen Variable besteht.



DHBW Lörrach

Logistische Regression - Modelgüte

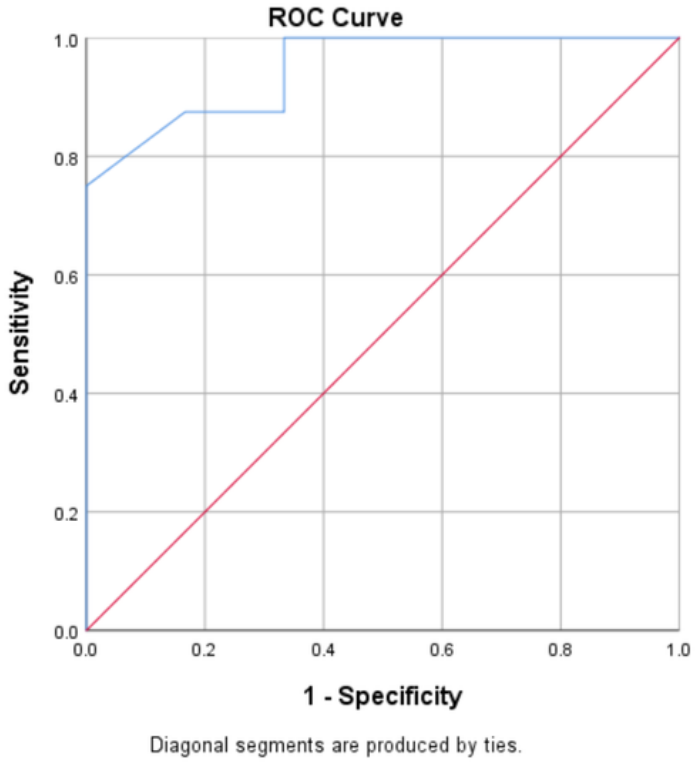
Um zu beurteilen, wie gut ein logistisches Regressionsmodell zu einem Datensatz passt, können wir die folgenden zwei Metriken betrachten:

- **Sensitivität**: Die Wahrscheinlichkeit, dass das Modell ein positives Ergebnis für eine Beobachtung vorhersagt, wenn das Ergebnis tatsächlich positiv ist.
- **Spezifität**: Die Wahrscheinlichkeit, dass das Modell ein negatives Ergebnis für eine Beobachtung vorhersagt, wenn das Ergebnis tatsächlich negativ ist.

Eine einfache Möglichkeit, diese beiden Metriken zu visualisieren, besteht darin, eine **ROC-Kurve** zu erstellen.

Logistische Regression – ROC-Kurve

ROC-Kurve ist ein Diagramm, das die Sensitivität und Spezifität eines logistischen Regressionsmodells anzeigt.



DHBW Lörrach

Vielen Dank!