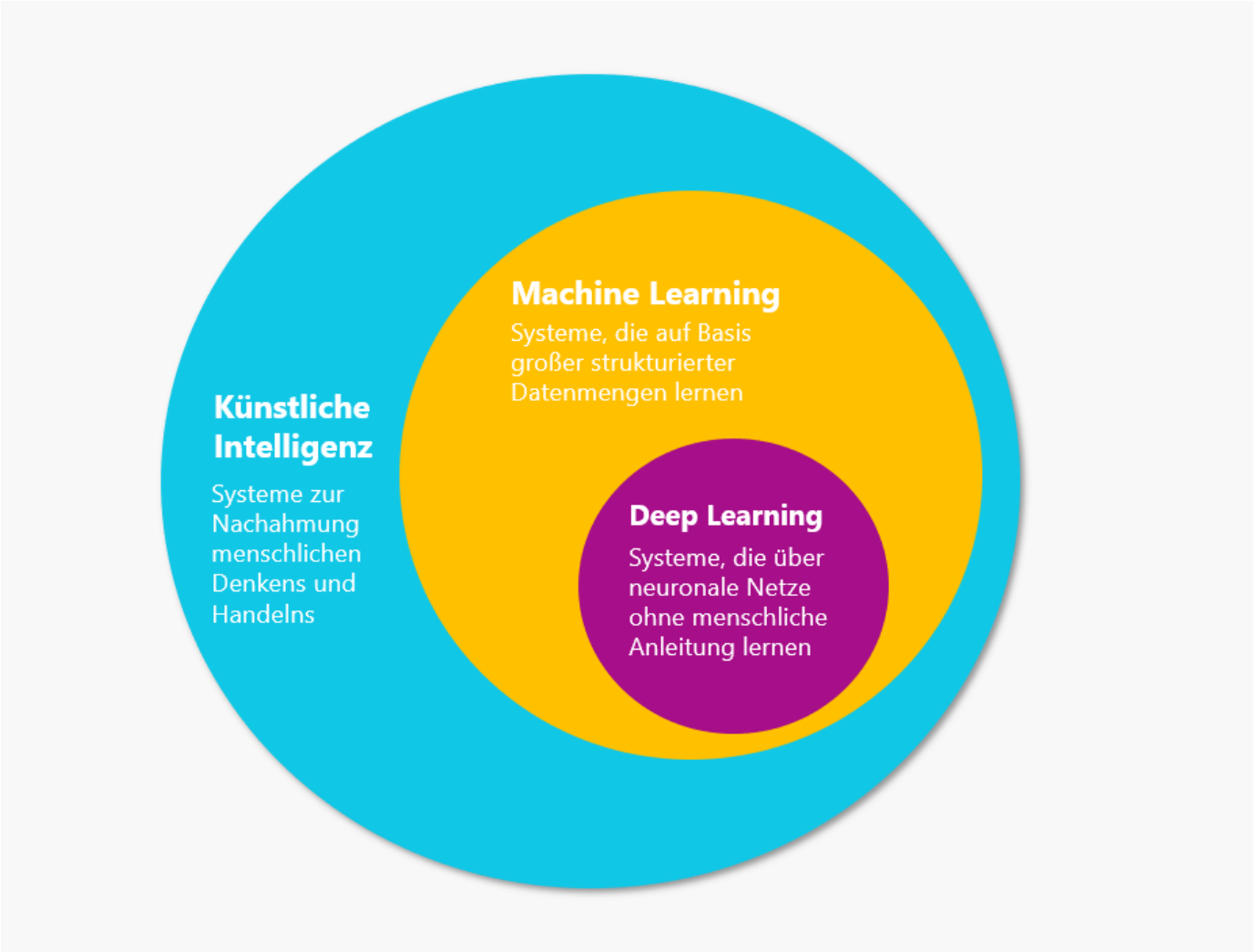
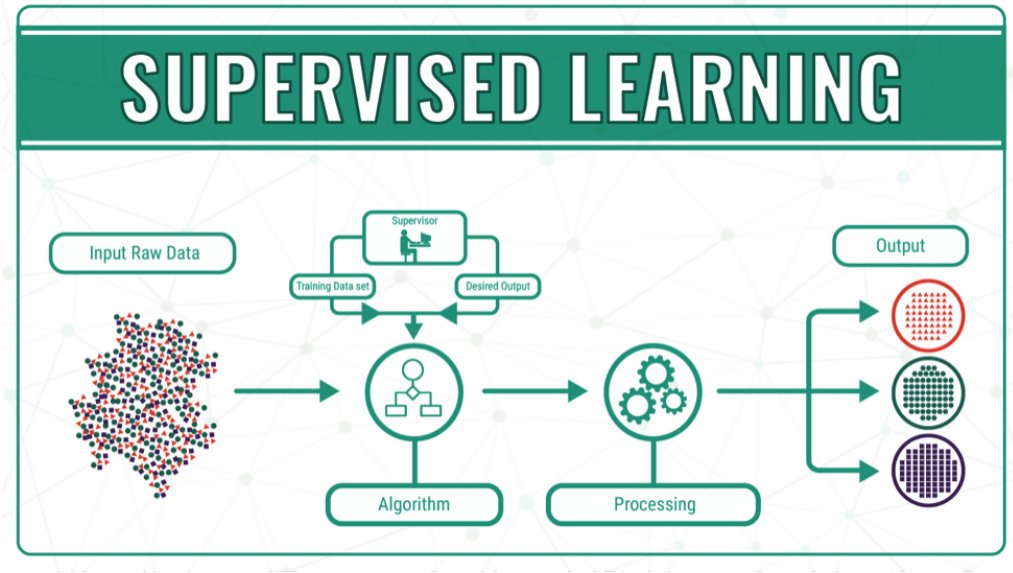
Datenanalyse auf Basis von KI-Methoden

KI vs Machine Learning vs Deep Learning



Machine Learning

- Überwachtes Lernen (supervised learning)
- Werden mit Hilfe von positiven/negativen Beispielen trainiert



Regression: „durchgehende“ Ausgabe

- Für jeden Input liefert das Modell einen durchgehenden Wert

Classification: bestimmte Ausgabe

- Für jeden Input liefert das Modell einen von speziellen Werten

Von Merkmal zu Variablen

Merkmal: Isolierte Eigenschaft eines größeren Ganzen, z.B. Intelligenz, Farbe, Einkommen

Ausprägung: Zustand des Merkmals, z.B. IQ =115, Farbe = Rot, Einkommen = hoch

Eine **Variable** wird definiert, indem den Ausprägungen des Merkmals **Zahlen** zugeordnet werden.

Diese Zahlen heißen **Realisationen** oder **Werte.**

Variablen

Eine **diskrete** Variable besitzt zumeist endlich viele und feste Werte, die man über Ganzzahlen beschreiben kann:

- **Dichtome** Variablen haben genau zwei diskrete Werte
- **Polytome** Variablen haben mehr als zwei diskrete Werte

Eine **stetige (kontinuierliche)** Variable kann (unendlich viele) beliebige Werte annehmen, die man über reelle Zahlen beschreibt

R Grundlagen –Pakete-

Pakete sind das Herzstück von R: Sie enthalten Funktionen, die andere Entwickler für uns vorbereitet haben

```
> # Ein Paket installieren
> install.packages("dplyr")
```

Pakete, die auf CRAN verfügbar sind, können einfach installiert werden

Installierte Pakete müssen, bevor ihre Funktionen genutzt werden können, erst geladen werden

```
> # Paket laden
> library(dplyr)
```

Python Grundlagen –Bibliotheken-

Bibliotheken sind das Herzstück von Python: Sie enthalten Funktionen, die andere Entwickler für uns vorbereitet haben

```
> # Ein Bibliothek installieren
> pip install pandas
```

Installierte Bibliotheken müssen, bevor ihre Funktionen genutzt werden können, erst geladen werden

```
> # Bibliothek laden
> import pandas
```

Daten laden und aufbereiten

CSV-Datei (Comma-separated Values, .csv)

- Standard-Format zum Austausch von strukturierten Daten
- Wie eine Tabelle: Zellen sind durch Trennzeichen getrennt, meistens , (Komma) oder ; (Semikolon)

```
lfdn;age;group;outcome
1;18;1;4
2;23;0;4
3;22;1;3
```

Daten in R laden

Legt das Arbeitsverzeichnis auf den Ordner, in dem ihr die Beispieldatensätze abgelegt habt

```
> setwd("C:/statistik")
```

Die Funktion read.csv2() ladet die csv-Datei

```
> df <- read.csv2("statistik.csv", header = TRUE, sep= ";", dec=".")
```

Daten in Python laden

Bei PyCharm wird automatisch als Arbeitsverzeichnis der Ordner, in dem ihr die Beispieldatensätze abgelegt habt, identifiziert

Die Funktion read_csv() von pandas ladet die csv-Datei

```
> import pandas as pd

> df = pd.read_csv("statistik.csv")
```

Datensatz kennenlernen

- In R

> str(df)
> summary(df)
> head(df)
> ncol(df)
> nrow(df)

- In Python

> df.head()
> df.info()

Daten aufbereiten

- Daten, die wir sammeln sind selten direkt für die Analyse bereit
- Wir haben fehlende Daten, brauchen neue Variablen, ggf. haben unterschiedliche Mitarbeiter unterschiedlich codiert, usw.
- Datenaufbereitung ist ein wichtiger und notwendiger Schritt in der Datenanalyse

Daten aufbereiten-Rechnen mit Variablen

- In R

```
> df$Angebot <- df$Price_euros - 100
```

- In Python

```
> df['Angebot'] = df['Price_euros'] -100
```

Daten aufbereiten-Variablen umbenennen

- In R

```
> df= rename(df, maxAngebot = Angebot)
```

- In Python

```
> df.rename(columns = {'Angebot':'maxAngebot'}, inplace = True)
```

Daten aufbereiten-Filtern

- In R

```
> Apple=filter(df, Company == "Apple")
> laptop_unt_1000=filter(df, Price_euros <= 1000)
```

- In Python

```
> is_apple =  df['Company']=="Apple"
> apple = df[is_apple]
> unt_1000 =  df['Price_euros']<= 1000
> laptop_unt_1000 = df[unt_1000]
```

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

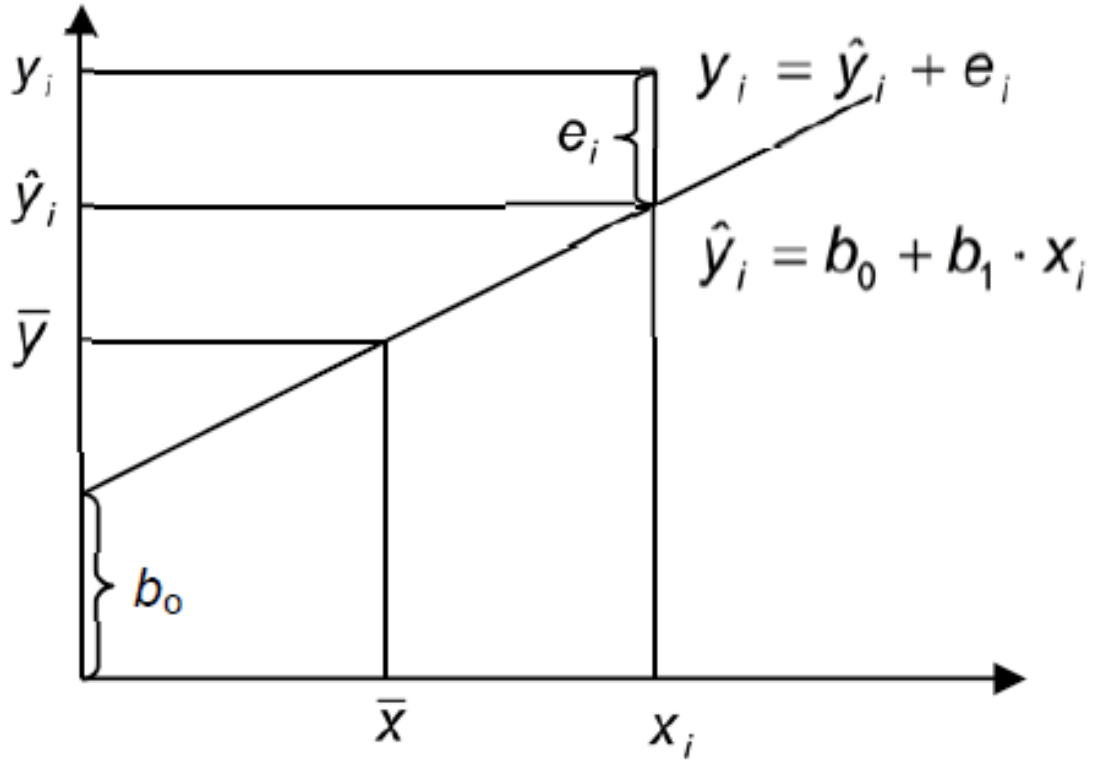
Regressionsgerade

- 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:

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

Methode der kleinsten Quadrate

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

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

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

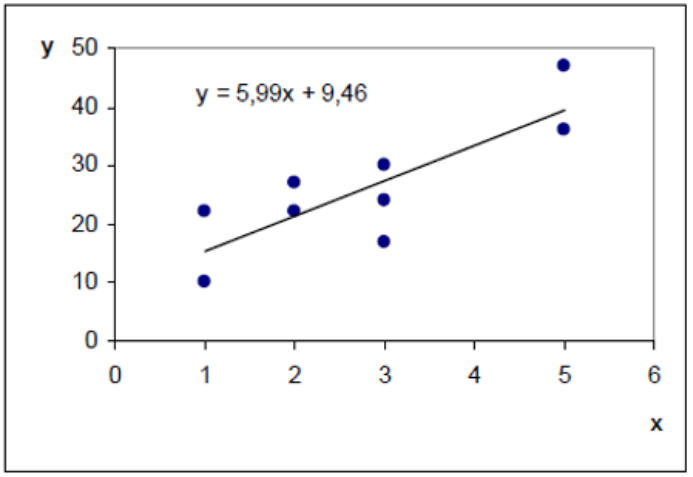
Beispiel

	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**:

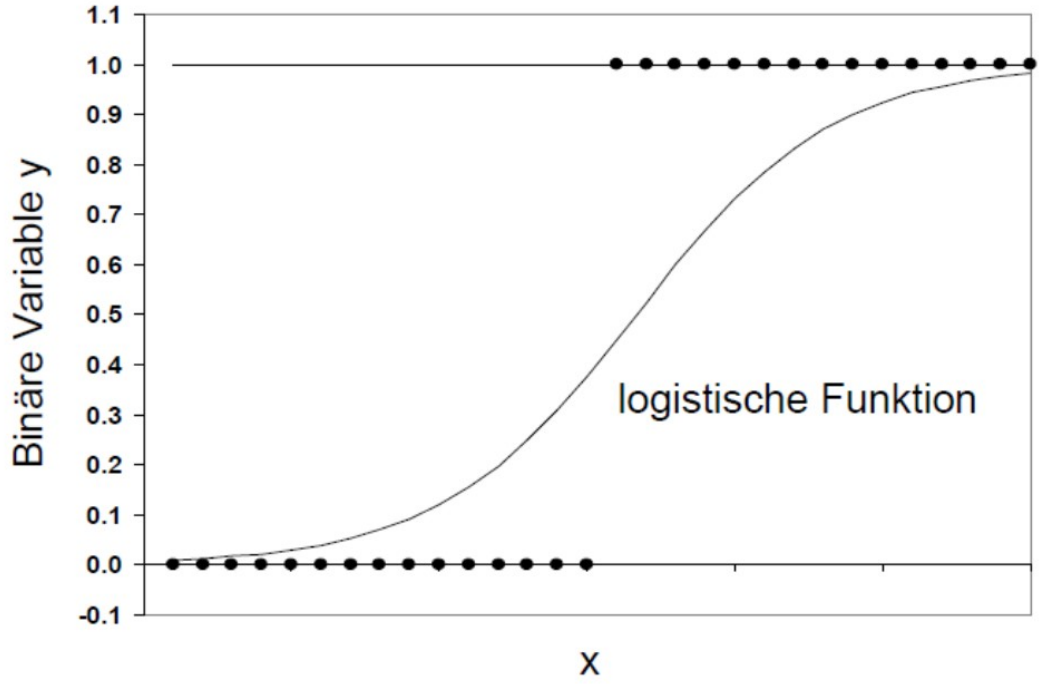
$$r^2 = \frac{SQ_{Reg}}{SQ_{Ges}} = \frac{b_1 \cdot \left[\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})\right]}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = b_1 \cdot \frac{SP_{xy}}{SQ_{xy}} = b_1^2 \cdot \frac{SQ_x}{SQ_y} \qquad ($$

Für das obige Beispiel folgt:

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

Logistische Regression

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



Logistische Regression - Modelgüte

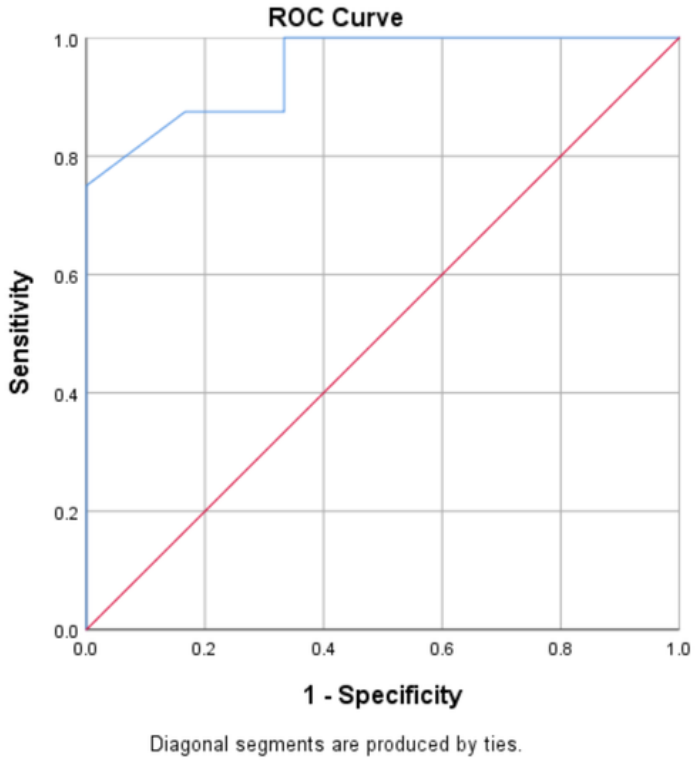
Um zu beurteilen, wie gut ein logistisches Regressionsmodell zu einem Datensatz passt, können wir die folgenden zwei Metriken betrachten:

- **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.



Vielen Dank

Für Ihre

Aufmerksamkeit!