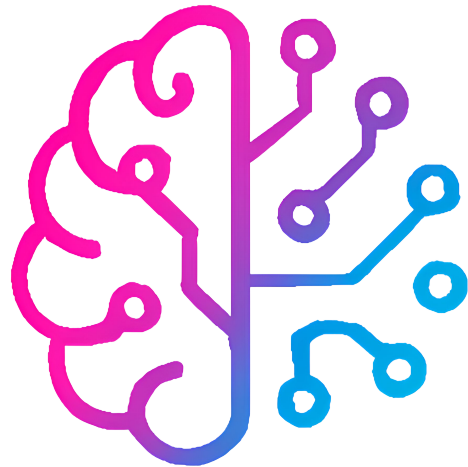


Machine Learning Club

Erstes Treffen



Wer sind wir?

Marius

- 20 Jahre Alt
- Studiert Mathe im 4ten Semester
- Programmiert Machine Learning Projekte in seiner Freizeit

Lars

- 21 Jahre Alt
- Studiert Mathe im 4ten Semester
- Arbeitet mit Machine Learning

Über den Club

Unsere Vision

- Gemeinschaft
- Atmosphäre der Neugier
- Zusammen kreieren
- Alle sind willkommen
- Spaß haben

Konzept

- Wettbewerbe einmal im Monat
- Datensätze werden über die Webseite: <https://machine-learning.club/> veröffentlicht
- Lösungen werden auf der Webseite eingereicht
- Gewinner kriegen Ehrenpreise und stellen ihre Lösungen vor
- Seminarvorträge, Hackathons, Lesegruppen etc...

Klassifikation



Was machen wir in Machine Learning?

- Machine Learning bedeutet: **Computer lernen aus Daten**, ohne explizit programmiert zu sein
- Ziel: **Muster und Zusammenhänge erkennen**, um Vorhersagen treffen zu können
- Es gibt zwei große Arten von Problemen:
 - **Regression** → Vorhersage von Zahlen (z. B. Temperatur, Umsatz)
 - **Klassifikation** → Einteilung in **feste Kategorien** (z. B. Spam / Kein Spam)



Was ist Klassifikation?

- Klassifikation bedeutet, dass ein Modell entscheidet, **zu welcher Klasse ein Objekt gehört**
- Die **Klassen** sind **vorher bekannt**, z. B.:
 - E-Mail → Spam oder Nicht-Spam
 - Bild → Katze, Hund oder Vogel
- Jede Klasse ist eine mögliche Antwort – das Modell wählt **eine davon** aus



Beispiel – Klassifikation im Alltag

 Problem	 Klassen
E-Mail-Filter	Spam / Nicht-Spam
Kreditvergabe	Ja / Nein
Handschriftliche Ziffern	0, 1, 2, ... , 9
Krankheitserkennung	Krank / Gesund

- In allen Fällen gibt es eine **feste Menge an Möglichkeiten**
- Das Modell soll aus Beispielen lernen, wie es neue Fälle einordnet



Wie funktioniert Klassifikation?

- **Daten sammeln** – viele Beispiele mit bekannter Antwort
- **Merkmale wählen** – welche Eigenschaften sind wichtig?
- **Modell trainieren** – Algorithmus findet Regeln in den Daten
- **Vorhersagen treffen** – neue Objekte werden automatisch eingeordnet

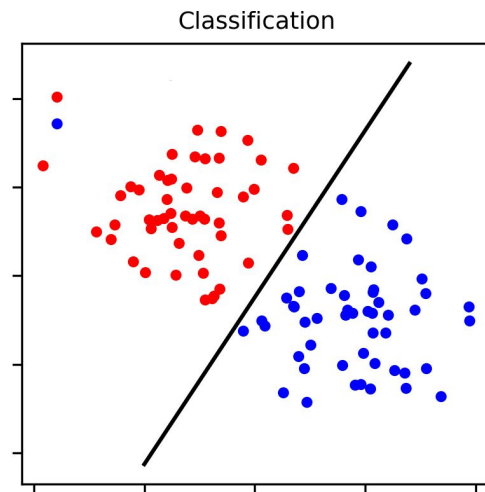
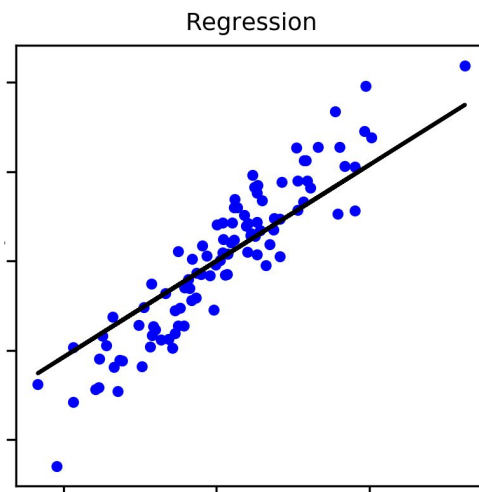


Das Modell sieht viele Beispiele – daraus lernt es, was **typisch für jede Klasse** ist



Ziel: Trennung der Klassen

- Das Modell versucht, eine Grenze zu finden, die die Klassen voneinander trennt
- Bei zwei Klassen (z. B. krank / gesund) wäre das eine Trennlinie im Datenraum
- Je nach Methode ist das eine:
 - Gerade
 - Kurve
 - eine komplexe Fläche





Was braucht das Modell?

- **Trainingsdaten** mit **bekannten Klassen**
- Jede Zeile besteht aus:
 - **Features** → messbare Eigenschaften (z. B. Alter, Blutdruck, Raucher: ja/nein)
 - **Label** → die Klasse (z. B. Schlaganfall: ja/nein)



Die **Features liefern** die **Informationen**, aus denen das Modell lernt








Training und Test

- Die Daten werden aufgeteilt in:
 - **Trainingsdaten** → zum Lernen
 - **Testdaten** → zum Überprüfen
- Wichtig: Das **Modell darf die Testdaten nicht vorher sehen!**
- Nur so können wir prüfen, ob es wirklich **verallgemeinern** kann









Häufige Klassifikationsverfahren

Modell	Beschreibung
 Logistische Regression	Einfaches Modell, gibt Wahrscheinlichkeiten für Klassen an
 Entscheidungsbaum	Wenn-Dann-Regeln, leicht verständlich
 Random Forest	Viele Bäume, robust und genau
 k-Nächste Nachbarn	Vergleicht mit den ähnlichsten Beispielen
 Neuronale Netze	Sehr flexibel, auch für Bilder, Sprache, etc.



Je nach Problem eignen sich **verschiedene Modelle besser**

Typischer Workflow

1.  **Datenvorbereitung** – aufräumen, normalisieren, codieren
2.  **Modellauswahl** – welches Modell passt?
3.  **Training & Validierung** – lernen und testen
4.  **Hyperparameter-Tuning** – Feineinstellungen optimieren
5.  **Evaluierung** – wie gut ist das Modell wirklich?
6.  **Einsatz** – auf neue Daten anwenden

 Dieser Ablauf **wiederholt** sich oft, um das Modell zu verbessern



Wie misst man, wie gut das Modell ist?




- **Accuracy** – wie viele Vorhersagen waren korrekt?
- **Precision** – wie viele der als positiv erkannten Fälle sind wirklich positiv?
- **Recall** – wie viele der tatsächlich positiven Fälle wurden erkannt?
- **Confusion Matrix** – zeigt alle Treffer und Fehler auf einen Blick



Nicht nur Accuracy zählt – bei medizinischen Daten ist Recall oft wichtiger!



Warum ist Klassifikation wichtig?

- Klassifikation hilft, **klare Entscheidungen** zu treffen
- Sie ist die Grundlage vieler moderner Anwendungen:
 -  Sicherheitsfilter
 -  Medizinische Diagnostik
 -  Wirtschaftliche Vorhersagen
- Machine Learning kann hier helfen, **schneller, objektiver und skalierbar** zu entscheiden

Erste Challenge: Schlaganfälle vorhersagen



Schlaganfälle Übersicht

- Zweithäufigste Todesursache Weltweit
- Risiko Faktoren
 - Beeinflussbar: Hoher Blutdruck, Rauchen, Diabetes, Übergewicht, und ein bewegungsarmer Lebensstil
 - Nicht Beeinflussbar: Alter, Geschlecht, und familien Geschichte
- 80 Prozent aller Schlaganfälle sind verhinderbar!











Unser Datensatz

- Vorhersage basierend auf Attributen wie Alter, BMI, Raucher, Vorerkrankungen etc.
- Ungefähr 5000 Datenpunkte
- Stark unbalanciert (mehr Leute die keinen Schlaganfall erlitten)



Aufbau der Daten

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
	Female 59% Male 41% Other (1) 0%					Private 58% Self-employed 16% Other (1076) 26%	Rural 50% Urban 50%			never smoked 37% Unknown 31% Other (1337) 33%	
48841	Male	31.0	0	0	No	Self-employed	Rural	64.85	23.0	Unknown	0
55244	Male	40.0	0	0	Yes	Self-employed	Rural	65.29	28.3	never smoked	0
70992	Female	8.0	0	0	No	children	Urban	74.42	22.5	Unknown	0
38207	Female	79.0	1	0	Yes	Self-employed	Rural	76.64	19.5	never smoked	0
8541	Female	75.0	0	0	Yes	Govt_job	Rural	94.77	27.2	never smoked	0
2467	Female	79.0	1	0	Yes	Self-employed	Rural	92.43		never smoked	0
19828	Female	56.0	1	0	Yes	Private	Rural	97.37	34.1	smokes	0
621	Male	69.0	0	0	Yes	Private	Rural	101.52	26.8	smokes	0
54975	Male	7.0	0	0	No	Self-employed	Rural	64.06	18.9	Unknown	0
57485	Female	1.48	0	0	No	children	Rural	55.51	18.5	Unknown	0
51746	Female	37.0	0	0	Yes	Govt_job	Rural	67.07	27.4	never smoked	0
30677	Female	3.0	0	0	No	children	Urban	82.91	19.9	Unknown	0
47357	Female	60.0	0	0	Yes	Private	Rural	62.78	36.4	Unknown	0
12204	Female	51.0	0	0	No	Govt_job	Rural	116.14	20.9	never smoked	0
42117	Male	43.0	0	0	Yes	Self-employed	Urban	143.43	45.9	Unknown	1





Evaluation

		POSITIVE	NEGATIVE
ACTUAL VALUES	POSITIVE	TP	FN
	NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$





Regeln

- Max. 4 Leute pro Team
- Wettbewerb endet am **31.05 um Mitternacht**
- Öffentliches und Privates Leaderboard
- Max. 5 Abgaben pro Tag

Einleitung zu Kaggle

<https://www.kaggle.com/competitions/schlaganfall-vorhersage-wettbewerb/leaderboard>



Preise

- Erster Platz: Titel des “Classification Casanovas” mit Pokal
- Bragging Rights ein Leben lang
- Gewinner stellen beim nächsten Treffen ihre Lösungen vor

Kontakt

Machine Learning Club

contact@machine-learning.club

<https://machine-learning.club>

