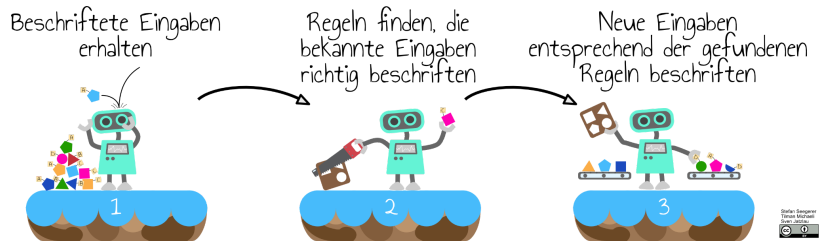


Überwachtes Lernen

Überwachtes Lernen kann verwendet werden, um eine Beschriftung aus einer Reihe von vorgegebenen Beschriftungen wie **beißt** und **beißt nicht** vorherzusagen, es kann aber auch dazu verwendet werden, einen Zahlenwert vorherzusagen, etwa die voraussichtlich anfallenden Schadenssummen bei einer Versicherung oder die Entwicklung von Aktienkursen und Hauspreisen.

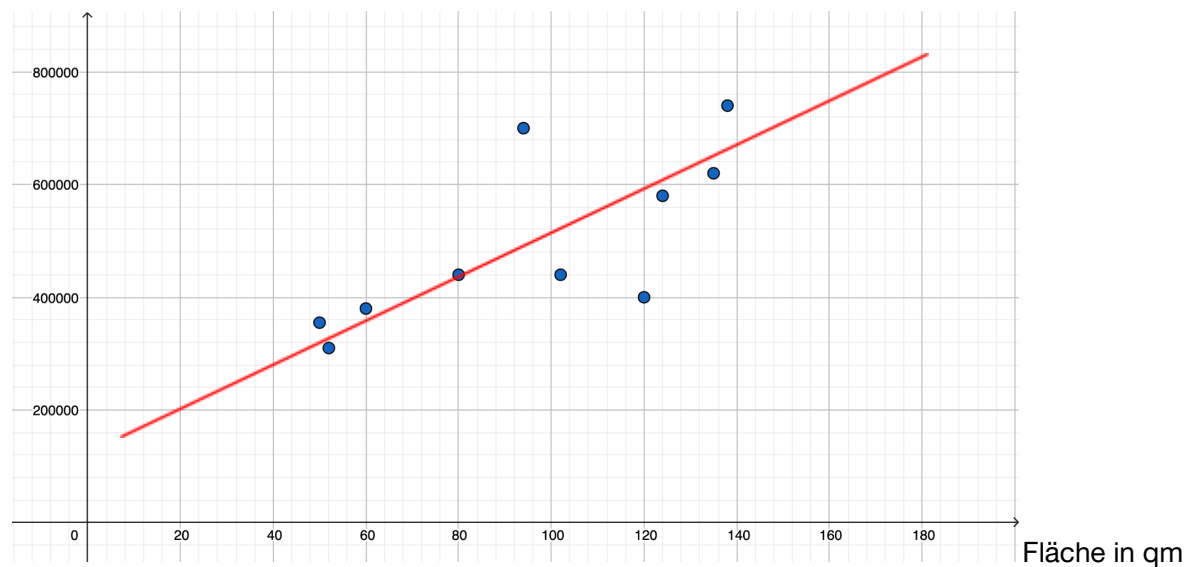


In diesem Arbeitsmaterial setzt du mit linearer Regression ein überwachtes Lernverfahren in Snap! um, um Hauspreise abhängig von ihrer Fläche in Quadratmetern vorherzusagen.

1. Aufgabe 1: Lineare Regression

a) In folgender Abbildung sind die Marktpreise und zugehörige Fläche in Quadratmetern für verschiedene Häuser angezeichnet. Zeichne eine Gerade ein, die sich den gegebenen Punkten bestmöglich annähert.

Marktpreis



b) Triff nun mithilfe deiner Gerade eine Vorhersage wie hoch der Preis für ein Haus mit 180m² Fläche ist : **abhängig von der gewählten Gerade, aber um die 800.000 €**

Überwachtes Lernen wird neben Klassifikationsproblemen auch bei **Regressionsproblemen** eingesetzt, in denen Eingaben keine Klasse (wie **beißt** oder **beißt nicht**) sondern ein numerischer Wert (wie etwa ein Hauspreis) zugeordnet werden soll. Der Computer versucht dazu einen Zusammenhang zwischen Eingabedaten (Fläche) und deren Beschriftung (Hauspreis) zu finden. Das Finden einer Gerade bezeichnet man als **lineare Regression**.

Mathematisch wird eine Gerade durch die **Steigung** a und den y -Achsenabschnitt b wie folgt beschrieben:

$$y = a \cdot x + b$$

Aufgabe des Computers ist es also, geeignete Werte für a und b zu finden. Die Parameter a und b stellen für den Computer also das Modell dar, mit deren Hilfe dann auch Vorhersagen für weitere Häuser mit ihrer gegebenen Fläche getroffen werden können.

2. Aufgabe 2: Vorhersage in Snap!

Zunächst benötigen wir einen Block, der für ein bereits existierendes Modell Vorhersagen trifft.

Fertiges Projekt als Lösung für Aufgabe 2 und 3: <https://bit.ly/A1-reg-l>

a) Öffne die Vorlage des Projekts "Hauspreise": <https://bit.ly/A1-reg>

Verwende anschließend den Block **plotte 2D Daten: Häuser**, um die Trainingsdaten auf der Bühne anzeigen zu lassen.

b) Implementiere den Block **sage Wert vorher für Eingabe: mit Modell:**,

indem du via Rechtsklick>Bearbeiten... in den Bearbeitungsmodus wechselst und den Wert berechnest und über den berichte Block zurückgibst.

Zur Erinnerung: $\text{Hauspreis} = a \cdot x + b$ mit Fläche x



Die Vorlage speichert die Parameter a und b in den beiden gleichnamigen Skriptvariablen

c) Rufe den gerade implementierten Block mit den folgenden Eingaben auf und überprüfe, ob du jeweils das angegebene Ergebnis erhältst.

sage Wert vorher für Eingabe: 30 **mit Modell:** Liste 4000 12000 **Ergebnis:** 132000

sage Wert vorher für Eingabe: 70 **mit Modell:** Liste 4000 12000 **Ergebnis:** 292000

3. Aufgabe 3: Lernen

Noch fehlt uns allerdings der wichtigste Schritt eines maschinellen Lernverfahrens: das Lernen des Modells und damit das Finden geeigneter Parameter a und b . Der Computer löst das aber nicht wie wir vorhin mit "draufgucken", sondern nimmt eine Gerade und passt sie schrittweise an.

Implementiere nun den Block **lerne Modell aus mit Lernrate:**, indem du via Rechtsklick>Bearbeiten... in den Bearbeitungsmodus wechselst.

Einige vorbereitende Schritte sind bereits erledigt: Die Variable **Eingabedaten** enthält alle Eingabedaten, die Variable **Erwartete Ausgaben** die zugehörigen Beschriftungen.

Hinweis: Da wir hier immer dieselben Operationen für alle Eingabedaten und erwarteten Ausgaben durchführen, können wir den Blöcken als Eingaben statt Einzelwerten auch Listen übergeben.

a) Rufe zunächst den Block **sage Wert vorher für Eingabe:** mit Modell: **Eingabedaten** und dem aktuellen Modell **Liste a b** auf. Speichere das Ergebnis in der Variable **Vorhersagen**.

b) Diese Vorhersagen sollen nun mit den erwarteten Ausgaben verglichen werden. Bestimme dazu den Fehler, indem du die Differenz aus **Vorhersagen** und **Erwartete Ausgaben** bestimmst.

c) Um diesen Fehler zu reduzieren, muss das Modell und damit die Variablen a und b entsprechend angepasst werden. Passe dazu a und b gemäß folgender Formel an:

setze **a** auf

$$a = \text{Lernrate} \times \frac{\text{Summe von Fehler} \times \text{Eingabedaten}}{\text{Länge von Trainingsdaten}}$$

setze **b** auf

$$b = \text{Lernrate} \times \frac{2 \times \text{Summe von Fehler}}{\text{Länge von Trainingsdaten}}$$

d) Überprüfe die gelernte Gerade visuell, indem Du den folgenden Block ausführst:

plotte Modell: lerne Modell aus Häuser mit Lernrate: 0.0001 im Maßstab des Datensatzes: Häuser

4. Aufgabe 4: Mehr als nur Quadratmeter

In der Realität hängt der Hauspreis aber natürlich nicht nur von der Fläche, sondern auch von vielen weiteren Faktoren ab. Sammle weitere für den Hauspreis relevante Merkmale:

Etwa Lage/Viertel, Baujahr, Energieverbrauch, letzte Sanierung, Stockwerk, Anzahl der Räume, ...