Machine Learning 1: Lineare Regression

Ergänzungsfach Informatik, 2021/2022, pro@kswe.ch

16. November 2021

Lernziele:

- Sie ordnen die Begriffe Künstliche Intelligenz, Machine Learning, Neuronale Netze und Deep Learning ein.
- Sie erklären was man unter Machine Learning versteht.
- Sie unterscheiden zwischen Supervised und Unsupervised Learning und geben jeweils Beispiele für Anwendungen aus diesem Bereich.
- Sie verwenden die mathematische Notation aus dem Machine Learning-Bereich.
- Sie erklären die Machine Learning-Fachbegriffe.
- Sie erklären die Idee der linearen Regression.
- Sie erklären an einem Beispiel die Methode der kleinsten Quadrate.
- Sie berechnen das Ergebnis für den Spezialfall der linearen Regression (einfache lineare Ausgleichsgeraden) mit einem Python-Programm.
- Sie erklären an einem Beispiel was ein überbestimmtes, lineares Gleichungssystem ist und wie man es lösen kann.
- Sie notieren ein lineares Gleichungssystem in Matrixform.
- Sie stellen die Ausgleichsfunktion (Minimierungsfunktion) für die mehrdimensionale lineare Regression auf.
- Sie lösen die mehrdimensionale lineare Regression mit einem Python-Programm.

1 Lineare Gleichungssysteme

Bei linearen Gleichungssystemen sind n Punkte gegeben und man stellt n lineare Gleichungen auf. Man sucht eine lineare Funktion, sodass alle Punkte durch die Funktion abgebildet werden. Je nachdem welche Punkte gegeben sind, existiert eine eindeutige Lösung, eine mehrdeutige Lösung oder keine Lösung. Besitzen die gegebenen Punkte die Dimensionalität 1 oder 2, dann spricht man auch davon, dass man eine Gerade (1-D Fall) oder eine Ebene (2-D Fall) durch die Punkte "konstruieren" möchte (geometrische Sichtweise).

Beispiel: Es sind die Punkte $P_1 = (1/2)$ und $P_2 = (2, 5/3)$ gegeben. Man möchte eine Geradengleichung in der Form $f(x) = y = m \cdot x + b$ finden, sodass die Gerade durch die beiden Punkte verläuft. Abbildung 1 zeigt den Graphen.

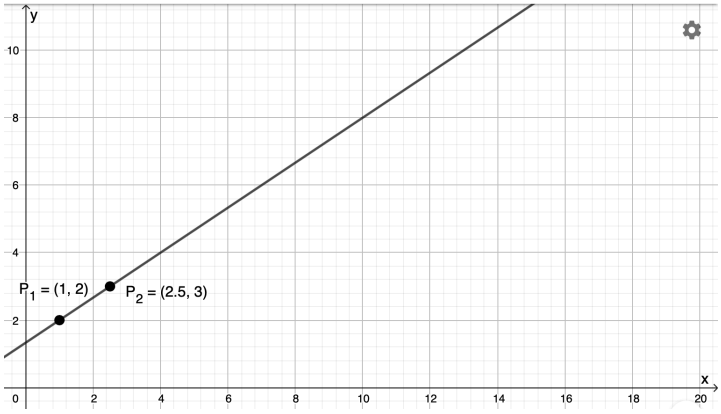


Abbildung 1: Gerade durch die beiden Punkte P_1 und P_2 (GeoGebra-Screenshot).

Dazu muss man die beiden Unbekannten m und b bestimmen. Die Unbekannten kann man bestimmen, indem man ein Gleichungssystem mit zwei Gleichungen aufstellt und löst ($P = (x, y)$):

$$\left| \begin{array}{l} 2 = \mathrm{m} \cdot 1 + \mathrm{b} \\ 3 = \mathrm{m} \cdot 2{,}5 + \mathrm{b} \end{array} \right|$$

Man erhält dann die Lösung $m = \frac{2}{3}$ und $b = \frac{4}{3}$, das heisst $f(x) = y = \frac{2}{3} \cdot x + \frac{1}{3}$.

Eine eindeutige Lösung ist immer dann möglich, wenn man bei n Punkten[1], n lineare Gleichungen aufstellt mit n Unbekannten. Hat man nun mehr Punkte als Unbekannte, dann ist keine eindeutige Lösung mehr möglich.

Beispiel: Es sind die drei Punkte $P_1 = (1/1)$, $P_2 = (2/2)$ und $P_3 = (2, 5/1)$ gegeben. Abbildung 2 stellt die Punkte grafisch dar. Durch diese drei Punkte lässt sich keine Gerade konstruieren.

2 Begriffe und Notation

Im Machine Learning-Bereich spricht man weniger von Punkten, sondern von Samples ("Datenpunkten")

Erklärung 2.1 (Sample) *Die gegebenen Daten (etwa Messdaten, Bilder, Text etc.) werden* ***Daten-Set*** *genannt und ein "Datensatz" (z.B. die Messwerte von einem Patienten) wird* ***Sample*** *genannt.*

Beispiel: Das Diabetes Daten-Set besteht aus $k = 442$ Samples. Für jedes Sample werden 11 Messwerte erhoben.

[1] Wenn die Punkte "ungeschickt" gewählt sind, dann kann es auch sein, dass es keine Lösung gibt. Die Punkte müssen linear unabhängig sein, wovon wir hier immer ausgehen.

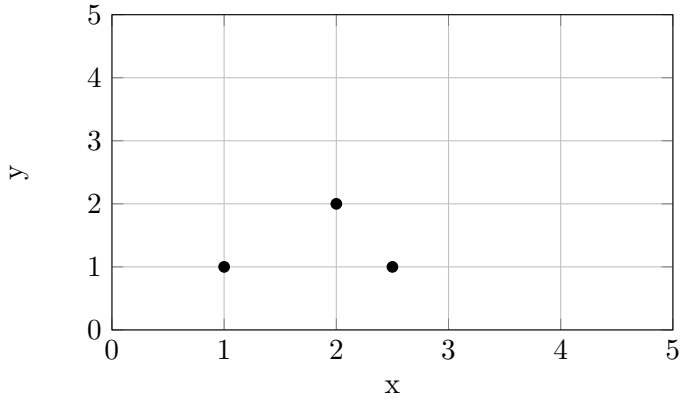


Abbildung 2: Die drei Punkte P_1, P_2 und P_3 im Koordinatensystem.

Jedes Sample besteht aus zwei Teilen: den erhobenen Merkmalen (z.B. Alter, Blutdruck, Pixel, E-Mails, etc.) und dem beobachteten Ergebnis (z.B. Krankheitsfortschreiten, Bild besitzt eine Katze oder nicht, E-Mail ist SPAM oder nicht). Die Merkmale werden **Features** genannt und das beobachtete Ergebnis bezeichnet man als **Label** (oder auch Target).

Beispiel: Ein Sample aus dem Diabetes Daten-Set besteht aus 10 Features und einem Label.

- Features
  - Alter
  - Geschlecht
  - BMI
  - Blutdruck (Durchschnitt)
  - Sechs Blutserummesswerte (z.B. Cholesterin, Blutzuckerspiegel, ...)
- Label (Target)
  - Quantitative Messung über das Krankheitsfortschreiten ("eine Zahl") im Vergleich zu einer "Baseline" (Messung nach einem 1 Jahr).

Erklärung 2.2 (Daten-Set Notation) *Ein Daten-Set wird als Menge von Samples beschrieben und mit D abgekürzt. Pro Sample gibt es einen Feature-Vektor. Der Vektor wird mit $\vec{x}$ abgekürzt und beschreibt die erhobenen Daten. Das Label ist eine Zahl und wird mit y abgekürzt. Bei mehreren Samples werden die Variablen nummeriert. Zusammengefasst: $D = \{(\vec{x_1}, y_1), (\vec{x_2}, y_2), \ldots, (\vec{x_n}, y_n)\}$.*

Beispiel: Auszug aus dem Diabetes Daten-Set: $D = \{(\vec{x_1}, y_1), (\vec{x_2}, y_2), \ldots, (\vec{x_{442}}, y_{442})\}$.

$$\vec{x_1} = \begin{bmatrix} 59 \\ 2 \\ 32.1 \\ 101 \\ 157 \\ 93.2 \\ 38 \\ 4 \\ 4.8598 \\ 87 \end{bmatrix} \quad \vec{x_2} = \begin{bmatrix} 48 \\ 1 \\ 21.6 \\ 87 \\ 183 \\ 103.2 \\ 70 \\ 3 \\ 3.8918 \\ 69 \end{bmatrix} \quad \cdots \quad \vec{x_{442}} = \begin{bmatrix} 36 \\ 1 \\ 19.6 \\ 71 \\ 250 \\ 133.2 \\ 97 \\ 3 \\ 4.5951 \\ 92 \end{bmatrix}$$

und $y_1 = 151$, $y_2 = 75$, $\ldots$ $y_{442} = 57$.

Jeder Sample besteht aus mehreren Features. Die Anzahl der Features bestimmt die Dimension des Vektors $\vec{x}$.

Erklärung 2.3 (Dimension) *Die Anzahl der Elemente in einem Vektor bestimmt die Dimension des Vektors. Repräsentiert ein Vektor ein Sample aus einem Daten-Set, dann ist die Anzahl der Features gleich der Dimension des Vektors. Wir notieren dies wie folgt:* $\vec{x} \in \mathbb{R}^n$*, falls der Vektor n reelle Zahlen beinhaltet oder* $dim(\vec{x}) = n$*, falls der Vektor neben Zahlen auch andere Werte beinhaltet (z.B. Text).*

Beispiel: Jeder Vektor aus dem Diabetes Daten-Set besitzt die Dimension 10. Es gilt somit $\vec{x} \in \mathbb{R}^{10}$.

Die einzelnen Elemente eines Vektors werden ebenfalls mit Namen versehen. Dabei werden die einzelnen Elemente ebenfalls mit x bezeichnet und nummeriert:

$$\vec{x_1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Achten Sie auf die unterschiedliche Schreibweise. Das Sample ist ein Vektor ($\vec{x_1}$), die Elemente des Vektors sind Werte (z.B. reelle Zahlen).

3 Lineare Regression

Bei der linearen Regression möchte man das gegebene Daten-Set durch eine lineare Funktion möglichst exakt beschreiben (modellieren). Aus der Machine Learning Sicht möchte man eine lineare Funktion finden, welche auch für zukünftige Samples (ohne Labels) ein möglichst gutes Ergebnis (Label) bestimmt. Man möchte mit der Funktion eine Vorhersage machen für neue, noch nicht bekannte Daten.

3.1 Einfache lineare Ausgleichsgeraden

Die einfache lineare Ausgleichsgeraden ist ein Spezialfall der linearen Regression. Man sucht für ein gegebenes Daten-Set eine lineare Funktion mit der Form $f(x) = y = w_1 \cdot x_1 + w_0$ (eine Gerade).

Erklärung 3.1 (Koeffizienten) *Im Machine Learning-Bereich werden die Koeffizienten mit w bezeichnet (eng. weight). Dies soll ausdrücken, dass man mit dem Koeffizienten bestimmen kann, wie viel Gewicht ein Feature erhalten soll.*

Wir sind nun daran interessiert die Koeffizienten w_1 und w_0 zu bestimmen. Falls es nur zwei Samples gibt, dann können wir eine Gerade konstruieren, die durch beide Punkte verläuft. Nun sind in der Regel aber weitaus mehr Samples gegeben (z.B. besteht das Diabetes Daten-Set aus $k = 442$ Samples). Selbst bei nur drei Samples (siehe Abbildung 2) ist es nicht mehr möglich eine Gerade mit dem üblichen Gleichungssystem zu ermitteln. **Es gibt mehr Samples als für die lineare Funktion notwendig**. Die lineare Regression versucht diese Problem nun zu lösen.

Erklärung 3.2 (Lineare Regression) *Finde eine lineare Funktion, welche die Samples aus dem Daten-Set möglichst exakt repräsentiert (modelliert). Optimiere die Funktion bezüglich der Label-Abweichungen aller Samples.*

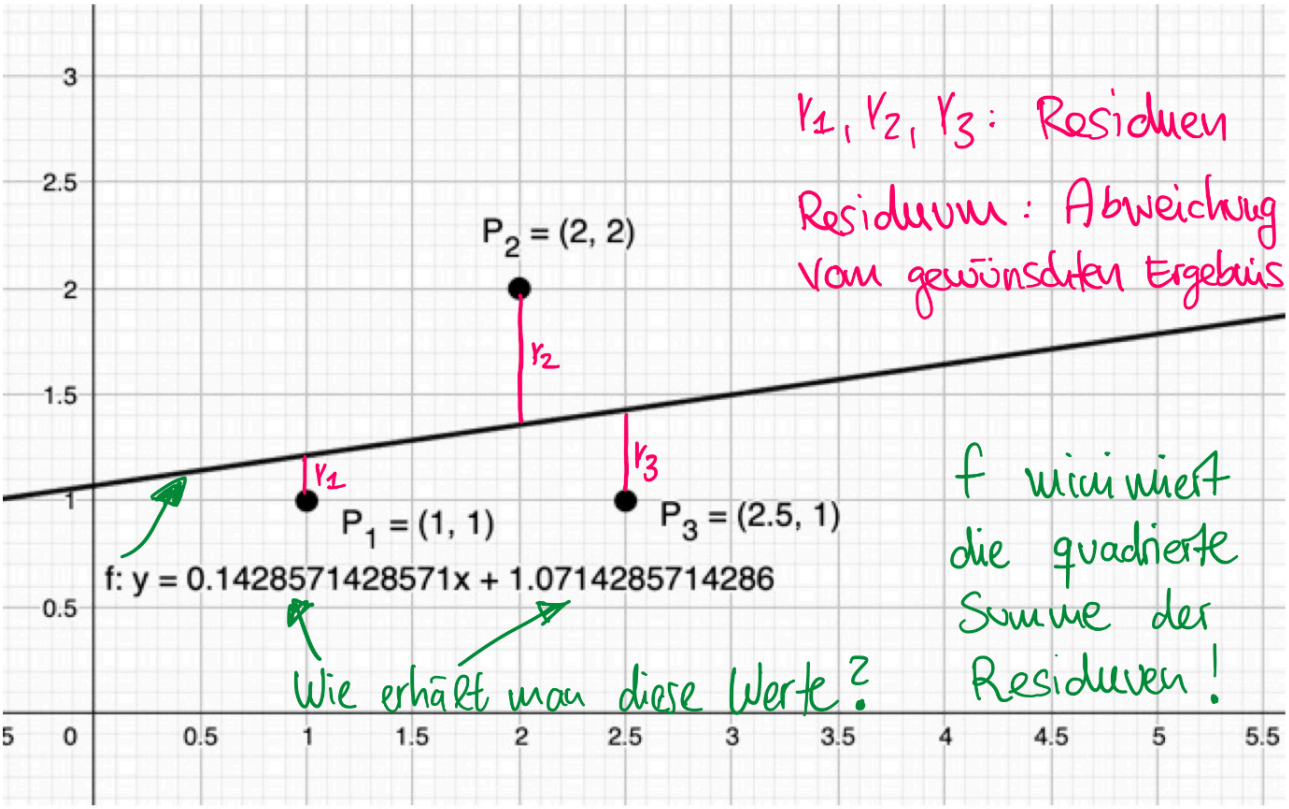
Für den Fall der einfachen linearen Ausgleichsgeraden bedeutet dies, dass man "zwischen die Samples eine Gerade legen möchte". Bei diesem Fall besitzen alle Samples die Dimension 1.

3.2 Abstraktes Beispiel

Das folgende Mini-Daten-Set besteht aus keinen echten Daten. Es ist gerade so gross gewählt, dass man das Prinzip übersichtlich darstellen kann.

Gegeben: $D = \{(1/1), (2/2), (2, 5/1)\}$

Gesucht: $f(x) = y = w_1 \cdot x_1 + w_0$



Wie berechnet man die Residuuen, wenn $f(x) = y = 0.14 \cdot x + 1.07$?

$$
\begin{aligned}
r_1 &= f(1) - 1 = (0.14 \cdot 1 + 1.07) - 1 = 0.21 \\
r_2 &= f(2) - 2 = (0.14 \cdot 2 + 1.07) - 2 = -0.65 \\
r_3 &= f(2,5) - 1 = (0.14 \cdot 2,5 + 1.07) - 1 = 0.42
\end{aligned}
$$

Ein positiver Wert für das Residuum bedeutet, dass die Funktion "zu hoch" liegt. Ein negativer Wert zeigt an, dass die Funktion "zu tief" liegt.

3.3 Methode der kleinsten Quadrate

Wir erhalten die Parameter für eine lineare Funktion, in dem wir die Summe der quadrierten Residuen minimieren. Dieses Verfahren ist unter dem Namen **Methode der kleinsten Quadrate** bekannt. Im 1-D Fall sucht man eine Gerade, welche die Residuen minimiert. Für das Beispiel von oben, suchen wir die besten Koeffizienten w_1 und w_0, um die Summe der quadrierten Residuen zu minimieren. Formal ausgedrückt:

$$
\arg\min_{w_0, w_1} \sum_{i=1}^{3} r_i^2 = \arg\min_{w_0, w_1} (r_1^2 + r_2^2 + r_3^2)
$$

Wir müssen in dieser Gleichung nun die Residuen berechnen und das Minimum suchen. Dies kann man analytisch durchführen ("ableiten" und "auf 0 setzen"). Die Residuen berechnen sich wie folgt:

$$\arg\min_{w_0,w_1}(r_1^2+r_2^2+r_3^2) = \arg\min_{w_0,w_1}((w_1\cdot x_{P_1}+w_0-y_{P_1})^2+(w_1\cdot x_{P_2}+w_0-y_{P_2})^2+(w_1\cdot x_{P_3}+w_0-y_{P_3})^2)$$

Wir haben die Residuen durch die x-Werte und y-Werte der gegebenen Punkte ausgedrückt. Mit x_{P_1} ist der x-Wert des Punktes P_1 gemeint. Wenn wir dies nun lösen, dann erhalten wir folgenden Formel um w_1 und w_0 zu berechnen:

$$\begin{aligned} w_1 &= \frac{\sum_{i=1}^{3}(x_{P_i}-\bar{x})\cdot(y_{P_i}-\bar{y})}{\sum_{i=1}^{3}(x_{P_i}-\bar{x})^2} \\ w_0 &= \bar{y}-w_1\cdot\bar{x} \\ \bar{x} &= \frac{\sum_{i=1}^{3}x_{P_i}}{3} \\ \bar{y} &= \frac{\sum_{i=1}^{3}y_{P_i}}{3} \end{aligned}$$

Mit $\bar{x}$ und $\bar{y}$ bezeichnen wir den Mittelwert aller x bzw. y-Werte. Wir können nun mit dieser Formel die optimalen Werte im Sinne der linearen Regression für das Beispiel mit den drei Punkten einfach ausrechnen.

3.4 Überbestimmte Gleichungssysteme

Im Beispiel aus Unterabschnitt 3.2 besteht das Daten-Set aus mehr Samples als für eine Gerade der Form $f(x)=y=w_1\cdot x_1+w_0$ notwendig. Das Daten-Set besitzt drei Samples, für eine Gerade werden jedoch nur zwei Samples benötigt. Man findet somit *keine* Lösung, sodass eine Gerade durch alle drei Punkte verläuft. Es gibt somit kein w_0 und kein w_1, sodass folgende drei Gleichungen gleichzeitig erfüllt sind:

$$\left|\begin{array}{l} 1 = w_1\cdot 1 + w_0 \\ 2 = w_1\cdot 2 + w_0 \\ 1 = w_1\cdot 2{,}5 + w_0 \end{array}\right|$$

Erklärung 3.3 (Überbestimmte Gleichungssysteme) *Sollen mehrere Gleichungen mit einer oder mehreren Unbekannten gleichzeitig erfüllt sein, dann spricht man von einem Gleichungssystem. Kommen die Unbekannten in den Gleichungen ausschliesslich in der ersten Potenz vor, dann spricht man von einem linearen Gleichungssystem. Gibt es mehr Gleichungen als Unbekannte, dann ist es ein überbestimmtes Gleichungssystem.*

Die lineare Regression versucht in der Regel ein überbestimmtes Gleichungssystem zu lösen. Es ist eine Optimierungsmethode. Das Standardverfahren zur Lösung ist die Methode der kleinsten Quadrate. Dabei werden die Unbekannten so bestimmt, dass die Summe der quadrierten Abstände (Residuen) möglichst gering ist.

Beispiel: Wir betrachten das Alter und den Krankheitsfortschritt aus dem Diabetes Daten-Set. Es gilt somit $D=\{(59/151),(48/75),(72/141),\ldots,(36/57)\}$ mit $|D|=k=442$. Es gibt somit 442 lineare Gleichungen der Form

$$\left|\begin{array}{ll} 151 & = w_1\cdot 59 + w_0 \\ 75 & = w_1\cdot 48 + w_0 \\ 141 & = w_1\cdot 72 + w_0 \\ & \vdots \\ 57 & = w_1\cdot 36 + w_0 \end{array}\right| \tag{1}$$

zu erfüllen. Dies ist ein überbestimmtes, lineares Gleichungssystem. Mithilfe der linearen Regression kann man w_1 und w_0 für eine einfache lineare Ausgleichsgeraden bestimmten.

3.5 Matrixform

Wir können mithilfe der **Matrix-Vektor-Multiplikation** ein lineares Gleichungssystem kompakt in **Matrixform** darstellen. Dies erleichtert die Schreibweise[2] und die Verarbeitung in einem Computerprogramm.

Dazu erstellen wir einen Vektor für die Unbekannten $\vec{w}$, einen Vektor für die y-Werte (linke Seite der Gleichungen) $\vec{y}$ und eine Koeffizientenmatrix X. Voraussetzung ist, dass alle Gleichungen geordnet sind. Alle identischen Unbekannten stehen untereinander, alle Koeffizienten stehen untereinander und alle y-Werte sind links vom Gleichheitszeichen ("linke Seite"). Dann können wir das lineare Gleichungssystem wie folgt notieren:

$$\vec{y} = X \cdot \vec{w}$$

Oft wird $\vec{y}$ auch auf der "rechten Seite" notiert, das heisst $X \cdot \vec{w} = \vec{y}$. Das lineare Gleichungssystem bleibt identisch. Wir erklären nun wie man die Vektoren und die Matrix bestimmt und geben danach ein konkretes Beispiel der Matrixform.

3.5.1 Vektoren

Man erhält den **Vektor für die Unbekannten**, in dem man alle Unbekannten untereinander notiert. Der Vektor wird im Machine Learning-Bereich meist mit $\vec{w}$ bezeichnet, da die Gewichte (eng. weights) unbekannt sind. Vektor-Form:

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \\ w_0 \end{bmatrix} \in \mathbb{R}^{n+1}$$

Mit n bezeichnen wir die **Dimension** der Samples im Daten-Set.

Beispiel: Der Vektor für die Unbekannten aus Gleichung 1 lautet:

$$\vec{w} = \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} \in \mathbb{R}^2$$

Der Vektor für die linke Seite (y-Werte) erhält man, in dem man alle y-Werte untereinander notiert. Der Vektor wird meist mit $\vec{y}$ bezeichnet, da darin alle y-Werte notiert werden. Im Machine Learning-Bereich sind dies die Labels. Vektor-Form:

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_k \end{bmatrix} \in \mathbb{R}^k$$

Mit k bezeichnen wir die **Anzahl** der Samples im Daten-Set.

Beispiel: Der Vektor für die linke Seite aus Gleichung 1 lautet:

$$\vec{y} = \begin{bmatrix} 151 \\ 75 \\ 141 \\ \vdots \\ 57 \end{bmatrix} \in \mathbb{R}^{442}$$

[2] Insbesondere bei mehrdimensionalen Samples.

3.5.2 Koeffizientenmatrix

Die Koeffizientenmatrix ist eine Matrix, die aus den Faktoren der Unbekannten besteht. Besitzt eine Unbekannte keinen expliziten Faktor, dann ist der Faktor 1 zu verwenden. Die Koeffizienten einer Gleichung entspricht einer Zeile der Matrix. Pro Spalte stehen somit die Koeffizienten bezüglich derselben Unbekannten untereinander. Die Koeffizienten werden von links nach rechts und von oben nach unten notiert.
Im Machine Learning-Bereich besteht die Koeffizientenmatrix gerade aus den **Feature-Vektoren**. Ein Feature-Vektor muss als **Zeile** einer Matrix notiert werden. Form:

$$X = \begin{bmatrix} \overrightarrow{x_1}^T & 1 \\ \overrightarrow{x_2}^T & 1 \\ \overrightarrow{x_3}^T & 1 \\ \vdots & \vdots \\ \overrightarrow{x_k}^T & 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n & 1 \\ x_1 & x_2 & \dots & x_n & 1 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ x_1 & x_2 & \dots & x_n & 1 \end{bmatrix} \in \mathbb{R}^{k \times n+1}$$

Mit k bezeichnen wir die **Anzahl** der Samples im Daten-Set. Mit n bezeichnen wir die **Dimension** der Samples im Daten-Set. Mit $\overrightarrow{x_1}^T$ einen Zeilenvektor (Transponieren von $\overrightarrow{x_1}$). Die letzte Spalte der Matrix sind die Faktoren für w_0.

Beispiel: Die Koeffizientenmatrix X für die rechte Seite aus Gleichung 1 lautet:

$$X = \begin{bmatrix} 59 & 1 \\ 48 & 1 \\ 72 & 1 \\ \vdots & \vdots \\ 36 & 1 \end{bmatrix} \in \mathbb{R}^{442 \times 2}$$

3.5.3 Matrixform der Gleichung 1

Die Matrixform des linearen Gleichungssystems aus Gleichung 1 lautet somit:

$$\overrightarrow{y} = X \cdot \overrightarrow{w} \Leftrightarrow \begin{bmatrix} 151 \\ 75 \\ 141 \\ \vdots \\ 57 \end{bmatrix} = \begin{bmatrix} 59 & 1 \\ 48 & 1 \\ 72 & 1 \\ \vdots & \vdots \\ 36 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_0 \end{bmatrix}$$

3.5.4 Kurzaufgabe

1. Notieren Sie für $D = \{(1/1), (2/2), (2,5/1)\}$ und $f(x) = y = w_1 \cdot x_1 + w_0$ das lineare Gleichungssystem in Matrixform.

Lösungsvorschlag:

$$\overrightarrow{y} = X \cdot \overrightarrow{w} \Leftrightarrow \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2,5 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_0 \end{bmatrix}$$

3.6 Mehrdimensionale lineare Ausgleichsfunktion

Bei der einfachen linearen Ausgleichsgeraden besitzen die Samples aus dem Daten-Set die Dimension 1. Nun möchten wir Samples mit einer höheren Dimension betrachten.

Beispiel: Das Diabetes Daten-Set $D = \{(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), \ldots, (\overrightarrow{x_{442}}, y_{442})\}$ besitzt Samples der Dimension 10.

$$\overrightarrow{x_1} = \begin{bmatrix} 59 \\ 2 \\ 32.1 \\ 101 \\ 157 \\ 93.2 \\ 38 \\ 4 \\ 4.8598 \\ 87 \end{bmatrix} \quad \overrightarrow{x_2} = \begin{bmatrix} 48 \\ 1 \\ 21.6 \\ 87 \\ 183 \\ 103.2 \\ 70 \\ 3 \\ 3.8918 \\ 69 \end{bmatrix} \quad \cdots \quad \overrightarrow{x_{442}} = \begin{bmatrix} 36 \\ 1 \\ 19.6 \\ 71 \\ 250 \\ 133.2 \\ 97 \\ 3 \\ 4.5951 \\ 92 \end{bmatrix}$$

und $y_1 = 151$, $y_2 = 75$, $\ldots y_{442} = 57$.

Ziel der linearen Regression ist es wieder eine lineare Funktion zu ermitteln, welche die Samples möglichst aussagekräftig beschreibt. Da die Samples nun eine höhere Dimension besitzen, spricht man nicht mehr von einer Ausgleichsgeraden, sondern von einer Ausgleichsfunktion. Geometrisch interpretiert sucht man zum Beispiel für zweidimensionale Samples eine Ebene. Bei höheren Dimensionen handelt es sich um eine Hyperebene, was grafisch nicht mehr darstellbar ist.

Erklärung 3.4 (Mehrdimensionale lineare Regression) *Die Idee der linearen Regression bleibt auch für höhere Dimensionen identisch. Die Samples des Daten-Sets bilden ein überbestimmtes, lineares Gleichungssystem. Es gibt keine exakte Lösung. Man sucht deshalb eine Hyperebene, welche die Summe der quadrierten Residuen (Abstand zwischen Hyperebene und y-Wert) minimiert.*

Die mehrdimensionale lineare Regression verallgemeinert die einfache lineare Regression. Es wird deshalb eine Ausgleichsfunktion der Form

$$f(x) = y = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + \cdots + w_n \cdot x_n + w_0$$

gesucht. Die Unbekannten sind w_0, w_1, $\ldots$, w_n. Die Matrixform des Gleichungssystems lautet:

$$\overrightarrow{y} = X \cdot \overrightarrow{w}$$

Vektoren und Matrix werden wie in Unterabschnitt 3.5 notiert. $\overrightarrow{y}$ beinhaltet die y-Werte der Samples, X beinhaltet die Feature-Vektoren und $\overrightarrow{w}$ die Unbekannten.

3.7 Abstraktes Beispiel

Das folgende Mini-Daten-Set besteht aus keinen echten Daten. Es ist gerade so gross gewählt, dass man das Prinzip übersichtlich darstellen kann.

Gegeben: $D = \{(\overrightarrow{x_1} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 1 \end{bmatrix}, 4), (\overrightarrow{x_2} = \begin{bmatrix} 20 \\ 2 \end{bmatrix}, 5), (\overrightarrow{x_3} = \begin{bmatrix} 30 \\ 1 \end{bmatrix}, 5), (\overrightarrow{x_4} = \begin{bmatrix} 40 \\ 3 \end{bmatrix}, 7)\}$

Gesucht: $f(x) = y = w_1 \cdot x_1 + w_2 \cdot x_2 + w_0$

Lineares Gleichungssystem:

$$\left| \begin{array}{ll} 4 & = w_1 \cdot 10 + w_2 \cdot 1 + w_0 \\ 5 & = w_1 \cdot 20 + w_2 \cdot 2 + w_0 \\ 5 & = w_1 \cdot 30 + w_2 \cdot 1 + w_0 \\ 7 & = w_1 \cdot 40 + w_2 \cdot 3 + w_0 \end{array} \right| \tag{2}$$

Lineares Gleichungssystem in Matrixform: $\vec{y} = X \cdot \vec{w} \Leftrightarrow$

$$\begin{bmatrix} 4 \\ 5 \\ 5 \\ 7 \end{bmatrix} = \begin{bmatrix} 10 & 1 & 1 \\ 20 & 2 & 1 \\ 30 & 1 & 1 \\ 40 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_0 \end{bmatrix}$$

3.8 Methode der kleinsten Quadrate

Das Verfahren löst das überbestimmte lineare Gleichungssystem, in dem die Unbekannten so gewählt werden, dass die Summe der quadrierten Residuen minimal ist. Dieses Verfahren wird Methode der kleinsten Quadrate genannt.
Residuen:

$$\left| \begin{array}{ll} r_1 & = w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n + w_0 \text{ - } y_1 \\ r_2 & = w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n + w_0 \text{ - } y_2 \\ & \vdots \\ r_k & = w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n + w_0 \text{ - } y_k \end{array} \right| \tag{3}$$

Es werden Werte für die Unbekannten w_0, w_1, …, w_n gesucht, sodass die Summe der quadrierten Residuen minimal ist:

$$\underset{w_0,w_1,w_2,\ldots,w_n}{\arg\min} \sum_{i=1}^{k} r_i^2 = \underset{w_0,w_1,w_2,\ldots,w_n}{\arg\min} \sum_{i=1}^{k} (w_1 \cdot x_{1i} + w_2 \cdot x_{2i} + \cdots + w_n \cdot x_{ni} + w_0 - y_i)^2$$

Die Notation $\underset{w_0,w_1,w_2,\ldots,w_n}{\arg\min}$ bedeutet, dass wir die Summe minimieren möchten und dabei die Unbekannten (auch Argumente genannt) “frei” wählen dürfen. Mit x_{1i} ist das i-te Sample gemeint und davon die 1. Komponente.

Residuen in Matrixform: $\vec{r} = X \cdot \vec{w} - \vec{y}$

Summe der quadrierten Residuen in Matrixform:

$$\underset{\vec{w}}{\arg\min}\ \vec{r}^T \cdot \vec{r} = \underset{\vec{w}}{\arg\min}\ ||\vec{r}||_2^2 = \underset{\vec{w}}{\arg\min}\ ||X \cdot \vec{w} - \vec{y}||_2^2$$

3.9 Abstraktes Beispiel

Wir führen das Beispiel mit dem Konzept aus Unterabschnitt 3.8 (Gleichung 3) fort.

Residuen:

$$\left| \begin{array}{ll} r_1 & = w_1 \cdot 10 + w_2 \cdot 1 + w_0 \text{ - } 4 \\ r_2 & = w_1 \cdot 20 + w_2 \cdot 2 + w_0 \text{ - } 5 \\ r_3 & = w_1 \cdot 30 + w_2 \cdot 1 + w_0 \text{ - } 5 \\ r_4 & = w_1 \cdot 40 + w_2 \cdot 3 + w_0 \text{ - } 7 \end{array} \right|$$

Summe der quadrierten Residuen:

$$\begin{aligned} \underset{w_0,w_1,w_2}{\arg\min} \sum_{i=1}^{4} r_i^2 &= \underset{w_0,w_1,w_2}{\arg\min} \sum_{i=1}^{4} (w_1 \cdot x_{1i} + w_2 \cdot x_{2i} + w_0 - y_i)^2 \\ &= (w_1 \cdot 10 + w_2 \cdot 1 + w_0 - 4)^2 + (w_1 \cdot 20 + w_2 \cdot 2 + w_0 - 5)^2 + (w_1 \cdot 30 + w_2 \cdot 1 + w_0 - 5)^2 \\ &\quad + (w_1 \cdot 40 + w_2 \cdot 3 + w_0 - 7)^2 \end{aligned}$$

Residuen in Matrixform: $\vec{r} = X \cdot \vec{w} - \vec{y} \Leftrightarrow$

$$\vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 10 & 1 & 1 \\ 20 & 2 & 1 \\ 30 & 1 & 1 \\ 40 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_0 \end{bmatrix} - \begin{bmatrix} 4 \\ 5 \\ 5 \\ 7 \end{bmatrix}$$

Summe der quadrierten Residuen:

$$\arg\min_{\vec{w}} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}^T \cdot \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \arg\min_{\vec{w}} \begin{bmatrix} r_1 & r_2 & r_3 & r_4 \end{bmatrix} \cdot \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \arg\min_{\vec{w}} \left\| \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} \right\|_2^2 = \arg\min_{\vec{w}} \left\| \vec{r} \right\|_2^2$$

Dies können wir explizit als Matrix-Vektor-Multiplikation notieren:

$$\arg\min_{\vec{w}} \left\| \vec{r} \right\|_2^2 = \left\| \begin{bmatrix} 10 & 1 & 1 \\ 20 & 2 & 1 \\ 30 & 1 & 1 \\ 40 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_0 \end{bmatrix} - \begin{bmatrix} 4 \\ 5 \\ 5 \\ 7 \end{bmatrix} \right\|_2^2$$

3.10 Lösung für die Methode der kleinsten Quadrate

Wir haben nun das überbestimmte, lineare Gleichungssystem für die mehrdimensionale lineare Regression kennengelernt. Wir haben auch gesehen, dass die Methode der kleinsten Quadrate versucht das Problem zu lösen, in dem die Residuen quadriert und summiert werden. Nun geht es konkret darum, wie man die Unbekannten bestimmen kann. Wie löst man also das folgende Minimierungsproblem:

$$\arg\min_{\vec{w}} ||X \cdot \vec{w} - \vec{y}||_2^2$$

In den meisten Fällen (es kommt auf die Matrix X an), lässt sich das Minimierungsproblem analytisch lösen. Dies bedeutet, man kann mit mathematischen Techniken eine Lösung finden. Da es ein Minimierungsproblem ist, können wir die Ableitung von $||X \cdot \vec{w} - \vec{y}||_2^2$ bestimmen und diese Nullsetzen und das Minimum bestimmen (wir suchen für eine mehrdimensionale Funktion die Extremstelle). Dadurch erhalten wir folgende Gleichunng:

$$X^T \cdot X \cdot \vec{w} = X^T \cdot \vec{y}$$

Wir sind nun daran interessiert den Vektor $\vec{w}$ zu bestimmen. Deshalb müssen wir die Gleichung umformen, sodass $\vec{w}$ auf der "linken Seite steht":

$$X^T \cdot X \cdot \vec{w} = X^T \cdot \vec{y}$$

Wir multiplizieren beide Seiten mit $(X^T \cdot X)^{-1}$und erhalten:

$$(X^T \cdot X)^{-1} \cdot X^T \cdot X \cdot \vec{w} = (X^T \cdot X)^{-1} \cdot X^T \cdot \vec{y}$$

Da $(X^T \cdot X)^{-1} \cdot X^T \cdot X = (X^T \cdot X)^{-1} \cdot (X^T \cdot X) = I$ die Einheitsmatrix ergibt, können wir die Gleichung vereinfachen:

$$I \cdot \vec{w} = (X^T \cdot X)^{-1} \cdot X^T \cdot \vec{y}$$

Die Einheitsmatrix verändert einen Vektor nicht, wir erhalten somit:

$$\vec{w} = (X^T \cdot X)^{-1} \cdot X^T \cdot \vec{y}$$

Diese Rechnung ergibt den gesuchten Vektor $\vec{w}$. Zur Berechnung sind alle Informationen vorhanden: Die Matrix X und der Vektor $\vec{y}$ können durch das Daten-Set bestimmt werden.