4. Adaline

4.1 Charakterisierung und Lernzyklus von Adaline

Überblick:

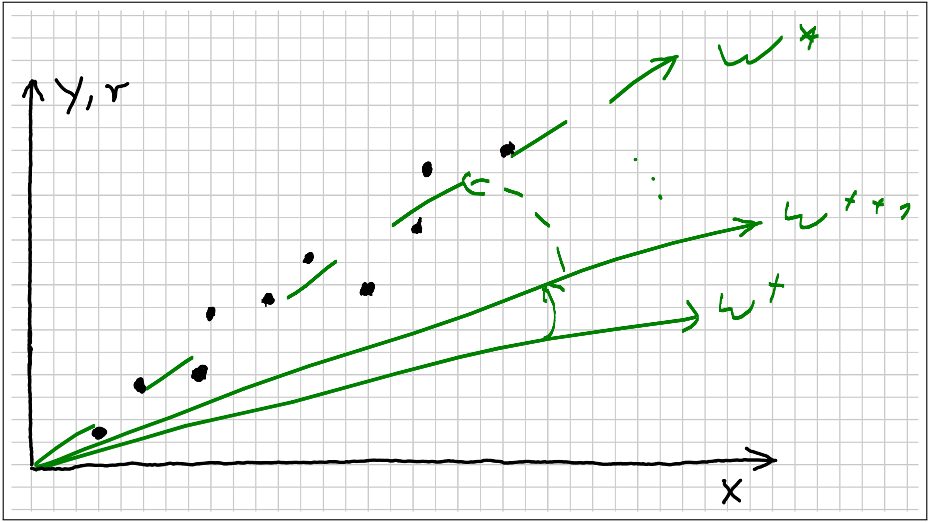
- Charakterisierung von Adaline
- Interpretation des Gewichtsvektors
- Graphische Illustration Lernzyklus

Charakterisierung von Adaline

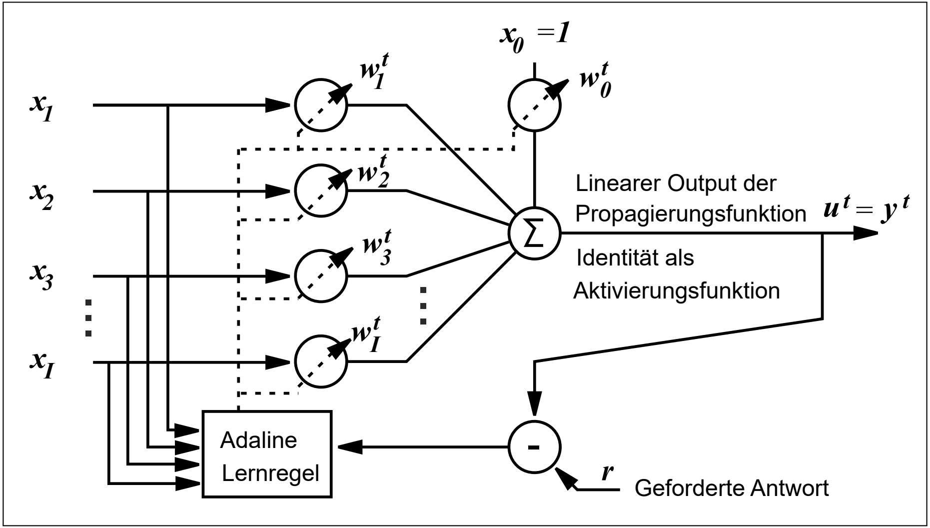
Adaptive **lin**ear **e**lement: Widrow, Hoff (1960)

- Ziel des Lernens ist eine lineare Funktion, mit Minimierung der Abweichung zwischen gelernter Funktion und den Trainingsdaten.
- Input, Gewichtung, Propagierungsfunktion wie Perzeptron, Aktivierungsfunktion ist die Identität.
- Für das Lernen wird der lineare Output der Propagierungsfunktion verwendet.

Interpretation des Gewichtsvektors



Graphische Illustration Lernzyklus



4.2 Adaline Proportional Lernregel

Überblick:

- Anpassung der Gewichte
- Änderung des Fehlers
- Graphische Illustration zur Anpassung der Gewichte
- Minimale Störung des bisherigen Lernerfolges

Anpassung der Gewichte

Anpassung der Gewichte eines einzelnen Adaline:

$$\boldsymbol{w}^{t+1} := \boldsymbol{w}^{t} + \boldsymbol{\alpha} \frac{\boldsymbol{d}^{t} \boldsymbol{x}}{\|\boldsymbol{x}\|^{2}}$$

$$\boldsymbol{d}^{t} := \boldsymbol{r} - \underbrace{\boldsymbol{w}^{t,T} \boldsymbol{x}}_{\boldsymbol{u}^{t}}$$

$\boldsymbol{r}$ ist die geforderte reelle Ausgabe zum Eingabevektor $\boldsymbol{x}$.

Die Normierung geschieht durch Division mit der quadrierten Länge des Eingabevektors.

Hinweis: Im Unterkapitel wird das erweiterte Neuronenmodell mit dem Gewichtsvektor $\boldsymbol{w}^{e}$ angenommen, aber zur kürzeren Schreibweise das Symbol $\boldsymbol{e}$ weggelassen.

Änderung des Fehlers

Die Veränderung der Gewichte bewirkt folgende Änderung des Fehlers:

$$\begin{aligned} \Delta d^t &:= (r - w^{t+1,T} x) - (r - w^{t,T} x) \\ &= -(\underbrace{w^{t+1} - w^t}_{\Delta w^t})^T x \end{aligned}$$

Speziell bei Adaline Proportional Lernregel gilt:

$$\Delta w^t := \alpha \frac{d^t x}{\|x\|^2}$$

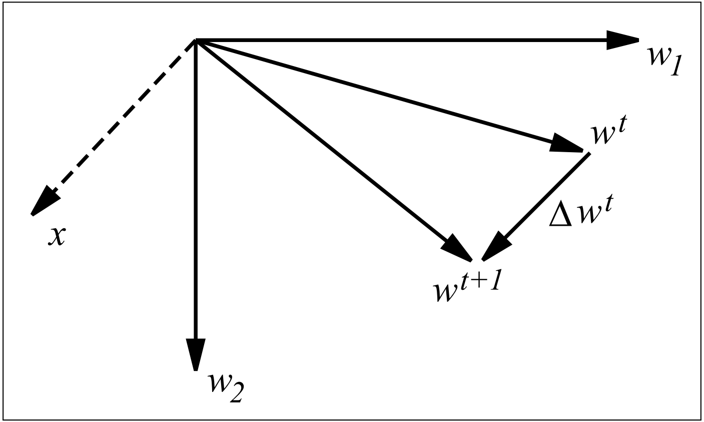
$$\Delta d^t := -\alpha \frac{d^t x^T}{\|x\|^2} x = -\alpha \cdot d^t$$

Änderung des Fehlers

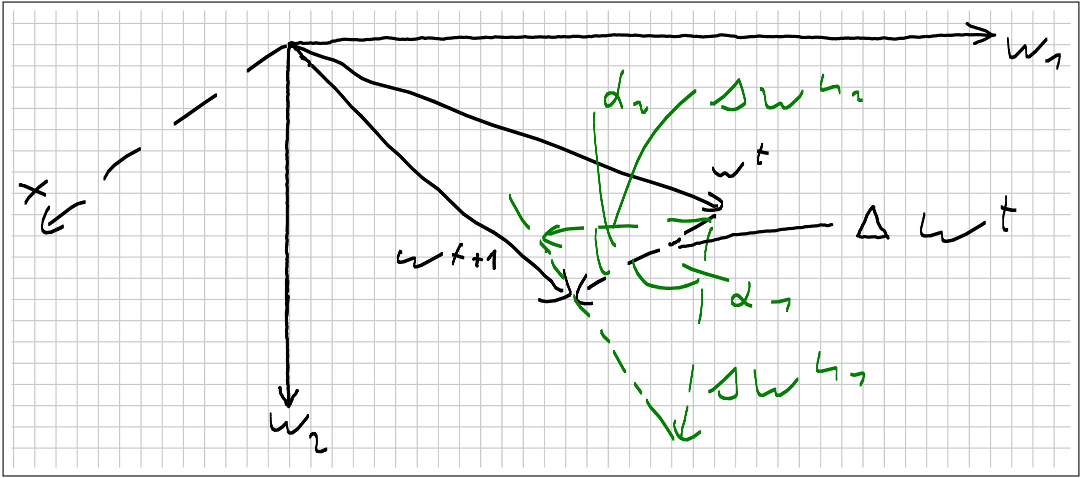
Resumee:

- Lineare Fehlerreduzierung, da proportional zu Fehler, reduziert Anteil $\boldsymbol{\alpha}$ von $\boldsymbol{d}^t$.
- Im Allgemeinen $\mathbf{0} < \boldsymbol{\alpha} < \mathbf{1}$.
  Falls speziell $\boldsymbol{\alpha} = \mathbf{1}$, dann Fehlerkorrektur.
- $\boldsymbol{\Delta w}^t$ parallel zu $\boldsymbol{x}$

Graphische Illustration zur Anpassung der Gewichte



Graphische Illustration zur Anpassung der Gewichte



Graphische Illustration zur Anpassung der Gewichte

$$\Delta w^t \circ x = \|\Delta w^t\| \cdot \|x\| \cdot \underbrace{\cos 0}_{1}$$

$$\Delta w^{h_1} \circ x = \underline{\|\Delta w^{h_1}\|} \cdot \|x\| \cdot \underline{\cos \alpha_1}$$

$$\Delta w^{h_2} \circ x = \underline{\|\Delta w^{h_2}\|} \cdot \|x\| \cdot \underline{\cos \alpha_2}$$

$$\|\Delta w^{h_1}\| \cdot \cos \alpha_1 = \|\Delta w^t\| = \|\Delta w^{h_2}\| \cdot \cos \alpha_2$$

Minimale Störung des bisherigen Lernerfolges

Es gibt verschiedene Varianten der Gewichtsänderung, welche zu gleicher Reduzierung des Fehlers $\boldsymbol{d}^t$ um $\boldsymbol{\Delta d}^t$ führen: z.B.

$$|\boldsymbol{\Delta d}^t| = |\boldsymbol{\Delta w}^{t,T}\boldsymbol{x}| = |\boldsymbol{\Delta w}^{h_1,T}\boldsymbol{x}| = |\boldsymbol{\Delta w}^{h_2,T}\boldsymbol{x}|$$

Wahl von $\boldsymbol{\Delta w}^t$, der parallel zu $\boldsymbol{x}$ ist, bewirkt minimale Gewichtsänderung, da kürzester Verschiebungsvektor.

$\Rightarrow$ Prinzip der minimalen Störung des bisherigen Lernerfolges.

4.3 Adaline Gradientenabstieg Lernregel

Überblick:

- Ensemble von Input-/Output-Tupeln
- Quadratische Fehlerfunktion
- Gradientenabstieg an Fehlerfunktion
- Direkte Ermittlung von Minimum der Fehlerfunktion
- Gradientenabstieg Lernregel

Ensemble von Input-/Output-Tupeln

Elementarfehler bei einem Input-/Output-Tupel:

$$\begin{aligned}(\boldsymbol{d}^t)^2 &= (\boldsymbol{r} - \boldsymbol{w}^{t,T}\boldsymbol{x})^2 \\ &= \boldsymbol{r}^2 - 2\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w}^t + \ \boldsymbol{w}^{t,T}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w}^t\end{aligned}$$

Eigentlich ist aber das Ensemble von Input-/Output-Tupeln, d.h. das Ensemble von zugehörigen Elementarfehlern, zu beachten.

Hinweis: Im Unterkapitel wird das erweiterte Neuronenmodell mit dem Gewichtsvektor $\boldsymbol{w}^e$ angenommen, aber zur kürzeren Schreibweise das Symbol $\boldsymbol{e}$ weggelassen.

Ensemble von Input-/Output-Tupeln

Übergang auf Erwartungswert $\boldsymbol{E}$ des Fehlers für alle Elemente aus $\boldsymbol{\Omega_T}$.

$$\underbrace{\boldsymbol{E((d^t)^2)}}_{=:\,\mathsf{MSE}(\boldsymbol{w^t})} = \boldsymbol{E((r)^2) - 2 \cdot \underbrace{E(rx^T)}_{=:\,V^T} \cdot w^t + w^{t,T} \cdot \underbrace{E(xx^T)}_{=:\,A} \cdot w^t}$$

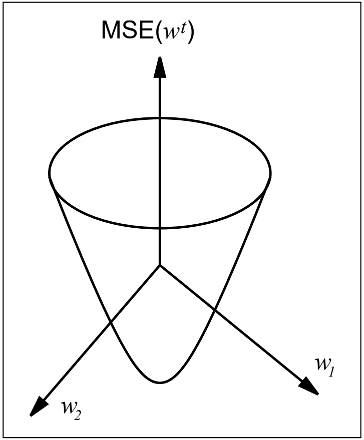
Mit $\boldsymbol{V}$ ein $(\boldsymbol{I} + 1)$-dimensionaler Spaltenvektor,

$\boldsymbol{A}$ eine quadratische $(\boldsymbol{I} + 1) \times (\boldsymbol{I} + 1)$ Matrix,

MSE die Mean Squared Error Funktion (im Gewichtsraum definiert).

Quadratische Fehlerfunktion

Die Fehlerfunktion MSE($\boldsymbol{w}^t$) ist quadratisch in den Gewichten, d.h. konvex mit globalem Minimum. Die Position eines Punktes entspricht der Wahl der Gewichte mit zugehörigem MSE-Wert.



Gradientenabstieg an Fehlerfunktion

- Ableiten der Fehlerfunktion (Fehlerfläche) liefert Gradientenvektor.
- Abstieg an Fehlerfläche in der negierten Gradientenrichtung, d.h. maximales Gefälle.
- Entsprechende Anpassung der Gewichte führt zum Minimum der Fehlerfunktion.

Direkte Ermittlung von Minimum der Fehlerfunktion

Falls komplettes Ensemble von Input-/Output-Tupeln vorliegt, und die Fehlerfunktion quadratisch ist, dann kann das Minimum der Fehlerfunktion, d.h. der optimale Gewichtsvektor, direkt ermittelt werden.

$$\boldsymbol{\nabla}\mathsf{MSE}(\boldsymbol{w}^t) := \begin{pmatrix} \frac{\partial \mathsf{MSE}(\boldsymbol{w}^t)}{\partial \boldsymbol{w}_0} \\ \vdots \\ \frac{\partial \mathsf{MSE}(\boldsymbol{w}^t)}{\partial \boldsymbol{w}_I} \end{pmatrix} = -\mathbf{2} \cdot \boldsymbol{V} + \mathbf{2} \cdot \boldsymbol{A} \cdot \boldsymbol{w}^t \stackrel{!}{=} \vec{\mathbf{0}}$$

$$\Longrightarrow \boldsymbol{w}^* := \boldsymbol{A}^{-1} \cdot \boldsymbol{V}$$

Gradientenabstieg Lernregel

Falls aber wie beim Online-Lernen die Input-/Output-Tupel nur sequentiell auftreten, dann wird für das aktuelle Tupel der momentane Gradientenvektor berechnet und damit der Gewichtsvektor angepasst.

$$\begin{aligned} \boldsymbol{w}^{t+1} &:= \boldsymbol{w}^{t} - \boldsymbol{\mu} \left(\frac{\boldsymbol{\partial (d^t)^2}}{\boldsymbol{\partial w^t}} \right) \\ &= \boldsymbol{w}^{t} + \boldsymbol{2 \mu d^t x} \end{aligned}$$

Hinweis: Beim Gradientenabstieg hat man eine garantierte aber langsame Konvergenz zum (nächsten lokalen) Minimum.

4.4 Anwendungen von Adaline

Überblick:

- Adaline als Adaptiver Filter
- Adaline zur System-Modellierung
- Adaline zur Prädiktion

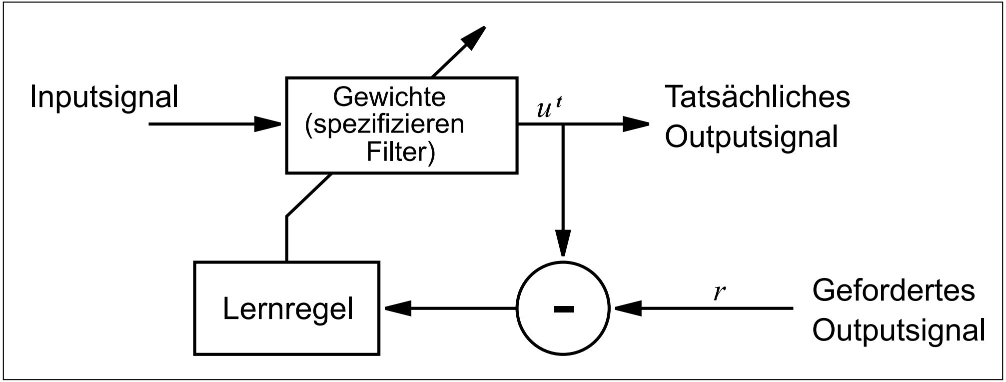
Adaline als Adaptiver Filter

Aufgabe: Ermitteln eines Operators (Filters) für Verarbeitung eines Signals (z.B. eines Bildes).

Lösung:

- Gefordertes Output-Signal mit tatsächlichem Output-Signal vergleichen.
- Gewichtsadaption (Filteradaption) aufgrund berechneter Differenz.

Adaline als Adaptiver Filter



Adaline als Adaptiver Filter

Beispiel für einen gelernten Filter (Operatormuster, Faltungskern), um Grauwertkanten (in horizontaler Richtung) zu extrahieren. Siehe Grundlagen der Bildverarbeitung.

1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	2	2	2	2	2	2	1	1	1
1	1	1	2	2	2	2	2	2	1	1	1
1	1	1	2	2	2	2	2	2	1	1	1
1	1	1	2	2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1

Faltung mit

1	0	-1

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	-1	-1	0	0	0	0	1	1	0	0
0	0	-1	-1	0	0	0	0	1	1	0	0
0	0	-1	-1	0	0	0	0	1	1	0	0
0	0	-1	-1	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Relevant bei sog. Convolutional Neural Networks (Deep Learning).

Adaline zur System-Modellierung

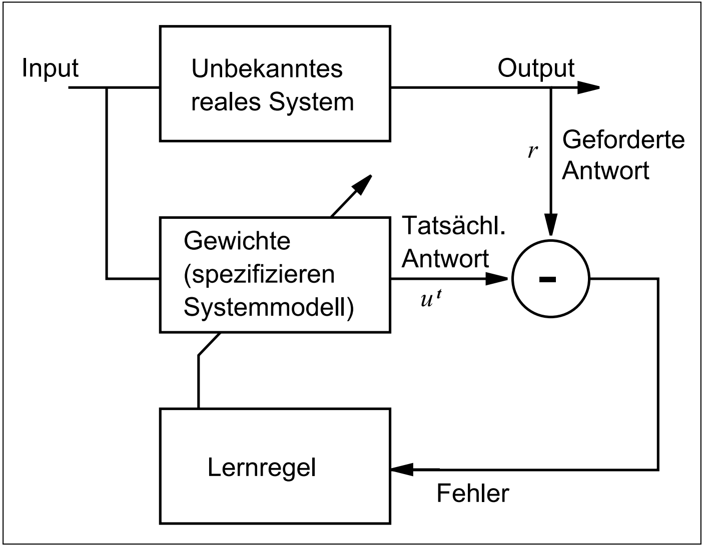
Synonym: System-Modellierung $=$ System-Identifikation

Aufgabe: Aus beobachtbarer Input-Output-Relation soll ein System modelliert werden.

Lösung:

- Gemeinsamer Input für Adaline und unbekanntem System.
- Tatsächlicher Output des unbekannten Systems sollte der Geforderte Output des Systemmodells sein.

Adaline zur System-Modellierung



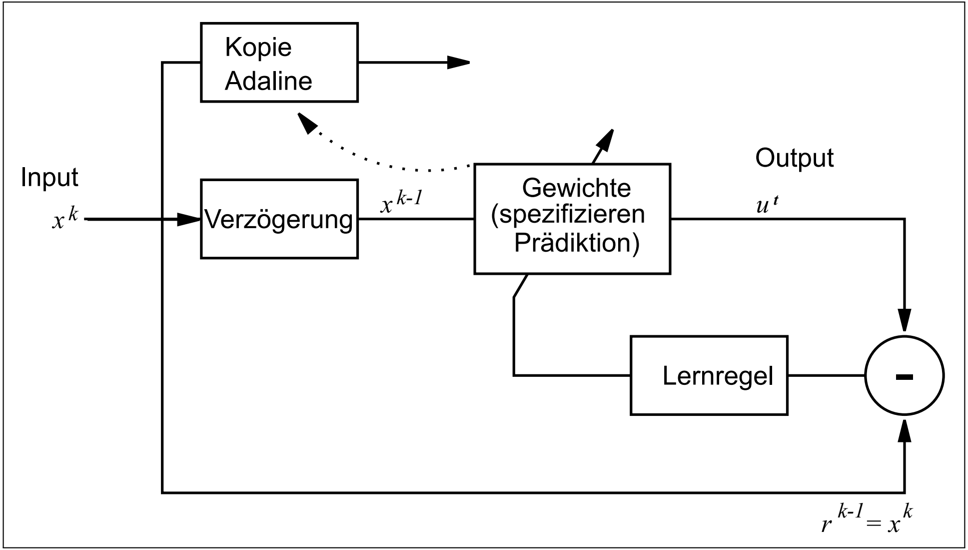
Adaline zur Prädiktion

Aufgabe: Schätzung späterer Signale aus früheren Signalen.

Lösung:

- Früheres Signal als Input, späteres Signal als geforderter Output, Gewichtsadaption.
- Kopie der optimalen Gewichte als fertiges Adaline zur Prädiktion.

Adaline zur Prädiktion



4.5 Lineare Regression durch Batch-Lernen

Überblick:

- Aufgabenstellung der Linearen Regression
- Gleichungssystem, Parameteroptimierung
- Dimension Eingaberaum versus Anzahl Trainingselemente

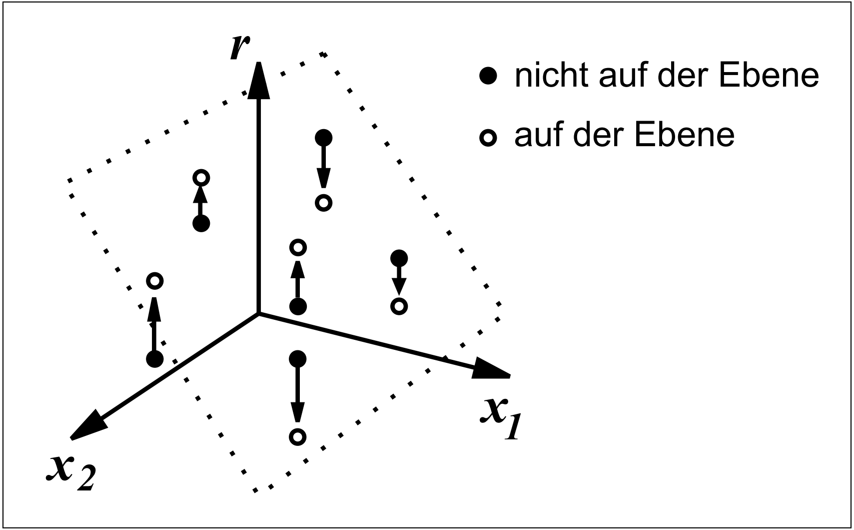
Aufgabenstellung der Linearen Regression

Annahme: Adaline; und alle Trainingselemente vorhanden für Batch-Lernen.

Gegeben: Trainingselemente $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M$ ($\boldsymbol{I}$-dimensional) mit zugehörigen skalaren Werten $\boldsymbol{r}^1, \ldots, \boldsymbol{r}^M$.

Gesucht: Parameter der $\boldsymbol{I}$-dimensionalen Ebene, so daß $(\boldsymbol{x}^m, \boldsymbol{r}^m)$ zur Ebene minimalen Abstand hat, für alle $\boldsymbol{m} \in \{1, \ldots, \boldsymbol{M}\}$.

Aufgabenstellung der Linearen Regression



Gleichungssystem, Parameteroptimierung

Gesucht sind optimale Werte für die Parameter $\boldsymbol{w_0, w_1, \ldots, w_I}$ in den Formeln $\quad \boldsymbol{r^m = w_0 x_0^m + w_1 x_1^m + w_2 x_2^m + \ldots + w_I x_I^m + \delta^m}$,
mit $\boldsymbol{x_0^m = 1}$, und $\boldsymbol{m \in \{1, \cdots, M\}}$,
sodaß die Fehler $\boldsymbol{\delta^m}$ minimal sind.

Matrixschreibweise:

Sei die Menge von $\boldsymbol{M}$ Trainingslementen $\boldsymbol{(x^{m,T}, r^m)}$ repräsentiert in Matrix $\boldsymbol{H}$ und Vektor $\boldsymbol{R}$, sowie die Gewichte $\boldsymbol{w_i}$ in Vektor $\boldsymbol{w^e}$ und die Fehler $\boldsymbol{\delta^m}$ in Vektor $\boldsymbol{\Delta}$.

Dann gilt folgendes System von Gleichungen: $\quad \boldsymbol{R = H \cdot w^e + \Delta}$

Gleichungssystem, Parameteroptimierung

Definitionen:

$$\boldsymbol{H} := \begin{pmatrix} \mathbf{1} & \boldsymbol{x}_1^1 & \cdots & \boldsymbol{x}_I^1 \\ \vdots & & \vdots & \\ \mathbf{1} & \boldsymbol{x}_1^M & \cdots & \boldsymbol{x}_I^M \end{pmatrix}; \qquad (\boldsymbol{M} \times (\boldsymbol{I} + \mathbf{1})) \text{ Eingabedaten-Matrix}$$

$$\boldsymbol{R} := (\boldsymbol{r}^1, \cdots, \boldsymbol{r}^M)^T; \qquad \boldsymbol{M} - \text{dimensionaler Solldaten-Vektor}$$

$$\boldsymbol{w}^e := (\boldsymbol{w}_0, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_I)^T; \qquad (\boldsymbol{I} + \mathbf{1}) - \text{dimens. Gewichts-Vektor}$$

$$\boldsymbol{\Delta} := (\boldsymbol{\delta}^1, \cdots, \boldsymbol{\delta}^M)^T; \qquad \boldsymbol{M} - \text{dimensionaler Fehler-Vektor}$$

Dimension Eingaberaum, Anzahl Trainingselemente

Fall a: $(\boldsymbol{I} + 1) = \boldsymbol{M} = \text{Rang}(\boldsymbol{H})$

Trainingselement $(\boldsymbol{x}^{m,T}, \boldsymbol{r}^{m})$ liefert Hyperebene im Gewichtsraum.

Betrifft Dualität Eingabe-/Gewichtsraum (siehe auch Unterkapitel 3.5).

Insgesamt erhält man $\boldsymbol{M}$ Hyperebenen der Form:
$(\mathbf{1}, \boldsymbol{x}^{m,T}) \cdot \boldsymbol{w}^{e} = \boldsymbol{r}^{m}$, $\boldsymbol{m} \in \{\mathbf{1}, \ldots, \boldsymbol{M}\}$.

Eindeutige Lösung $\boldsymbol{w}^{e,*}$, als Schnittpunkt der $\boldsymbol{M}$ Hyperebenen im Gewichtsraum.

Gleichungssystem $\boldsymbol{H} \cdot \boldsymbol{w}^{e} = \boldsymbol{R}$, Lösung $\boldsymbol{w}^{e,*} := \boldsymbol{H}^{-1} \cdot \boldsymbol{R}$

Fehlervektor $\boldsymbol{\Delta}$ ist der Nullvektor, d.h. Fehlerfunktion $\boldsymbol{D}(\boldsymbol{w}^{e,*}) = \mathbf{0}$

Dimension Eingaberaum, Anzahl Trainingselemente

Fall b: Rang$(\boldsymbol{H}) = (\boldsymbol{I}+1) < \boldsymbol{M}$

Überbestimmtes Gleichungssystem: $\quad \boldsymbol{R} = \boldsymbol{H} \cdot \boldsymbol{w}^e + \boldsymbol{\Delta}$

Es existiert keine fehlerfreie, aber eine eindeutig bestimmte Lösung $\boldsymbol{w}^{e,*}$.
Diese wird gefunden am globalen Minimum der Fehlerfunktion

$$\boldsymbol{D}(\boldsymbol{w}^e) := \sum_{m=1}^{M} \left(r^m - \sum_{i=0}^{I} w_i x_i^m \right)^2$$

Bestimmung von Vektor $\boldsymbol{w}^{e,*}$ durch Minimierung von $\boldsymbol{D}(\boldsymbol{w}^e)$.
Dabei gilt unter Verwendung von Matrizen:

$$\boldsymbol{D}(\boldsymbol{w}^e) = \|\boldsymbol{\Delta}\|^2 := (\boldsymbol{R} - \boldsymbol{H} \cdot \boldsymbol{w}^e)^T \cdot (\boldsymbol{R} - \boldsymbol{H} \cdot \boldsymbol{w}^e)$$

Dimension Eingaberaum, Anzahl Trainingselemente

Lösung: Partielle Ableitungen der Fehlerfunktion $\boldsymbol{D}(\boldsymbol{w}^e)$ nach den Parametern $\boldsymbol{w}_i$, und diese jeweils Null-Setzen.

$$\nabla_{\boldsymbol{w}^e}[(\boldsymbol{R} - \boldsymbol{H} \cdot \boldsymbol{w}^e)^T \cdot (\boldsymbol{R} - \boldsymbol{H} \cdot \boldsymbol{w}^e)] = \\ -2 \cdot \boldsymbol{H}^T \cdot \boldsymbol{R} + 2 \cdot \boldsymbol{H}^T \cdot \boldsymbol{H} \cdot \boldsymbol{w}^e \stackrel{!}{=} \vec{0}$$

Daraus folgt $\boldsymbol{w}^{e,*} := \underbrace{(\boldsymbol{H}^T \cdot \boldsymbol{H})^{-1} \cdot \boldsymbol{H}^T}_{\text{Pseudo-Inverse von } \boldsymbol{H}} \cdot \boldsymbol{R}$

Bei linearer Regression und quadratischer Fehlerfunktion führt der Gradientenabstieg zur Pseudo-Inversen. Die Lösung $\boldsymbol{w}^{e,*}$ ergibt sich direkt aus obiger Formel.

Dimension Eingaberaum, Anzahl Trainingselemente

Fall c: Rang$(\boldsymbol{H}) = \boldsymbol{M} < (\boldsymbol{I} + 1)$

Unterbestimmtes Gleichungssystem: $\boldsymbol{R} = \boldsymbol{H} \cdot \boldsymbol{w}^e + \boldsymbol{\Delta}$

Es existiert keine eindeutige Lösung, sondern sogar mehrere Lösungen für $\boldsymbol{w}^e$, und alle diese Lösungen sind fehlerfrei. Die Fehlerfunktion hat an diesen Stellen also den Wert Null.

Hinweis zur Bestimmung dieser Lösungen: Singulärwertzerlegung.