Multivariate Datenanalyse

MSc Psychologie WiSe 2021/22

Prof. Dr. Dirk Ostwald

(6) Optimierung

Multivariate Datenanalyse

Datum	Einheit	Thema
15.10.2021	Einführung	(0) Einführung
15.10.2021	Grundlagen	(1) Vektoren
22.10.2021	Grundlagen	(2) Matrizen I
29.10.2021	Grundlagen	(3) Matrizen II
05.11.2021	Grundlagen	(4) Multivariate Normalverteilung
12.11.2021	Latente Variablenmodelle	(5) Hauptkomponentenanalyse
19.11.2021	Latente Variablenmodell	(6) Faktorenanalyse
26.11.2021	Prädiktive Modellierung	(7) Optimierung
03.12.2021	Prädiktive Modellierung	(8) Lineare Diskriminanzanalyse und Logistische Regression
10.12.2021	Prädiktive Modellierung	(9) Support Vektor Maschinen
17.12.2021	Prädiktive Modellierung	(10) Neuronale Netze
	Weihnachtspause	
07.01.2022	Frequentistische Inferenz	(11) T-Tests
14.01.2022	Frequentistische Inferenz	(12) Einfaktorielle Varianzanalyse
21.01.2022	Frequentistische Inferenz	(13) Kanonische Korrelationsanalyse I
28.01.2022	Frequentistische Inferenz	(14) Kanonische Korrelationsanalyse II
22.02.2022	Klausur	12 - 13 Uhr, G26-H1
Jul 2022	Klausurwiederholungstermin	

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Prädiktive Modellierung

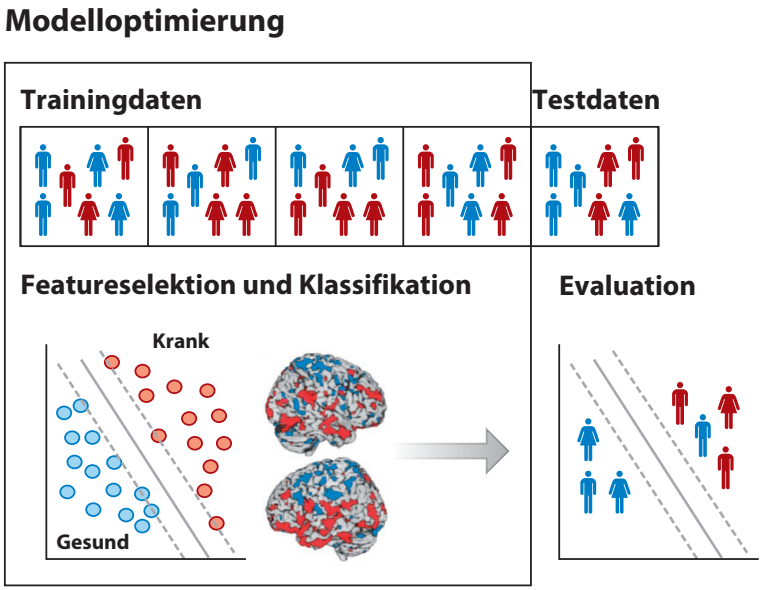
In der Psychologie möchte man gerne Dinge prädizieren, zum Beispiel

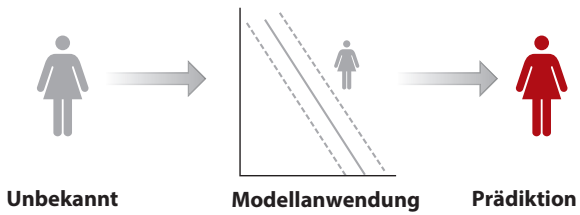
- Risiko psychiatrischer Erkrankung basierend auf Fragebogendaten
- Prognose psychiatrischer Erkrankung basierend auf Hirnbildgebungsdaten
- Psychotherapieerfolg basierend auf klinisch-psychologischen Tests
- Subjektive Wahrnehmung (Bewusstsein) basierend auf funktionellen Hirnbildgebungsdaten
- ...

Prädiktive Modellierung ist ein datenanalytisches Paradigma, das die prädiktive Rhethorik bedient.

- Die Datengrundlage ist hier oft multivariat, die Vorhersage ist oft univariat.
- Prädiktive Modellierung wird oft mit "Maschinellem Lernen" gleichgesetzt oder verwechselt.
- Prädiktive Modellierung wird oft mit "Künstlicher Intelligenz" gleichgesetzt oder verwechselt.
- Prädiktive Modellierung wird oft mit "Deep Learning" gleichgesetzt oder verwechselt.

Struktur der Prädiktiven Modellierung





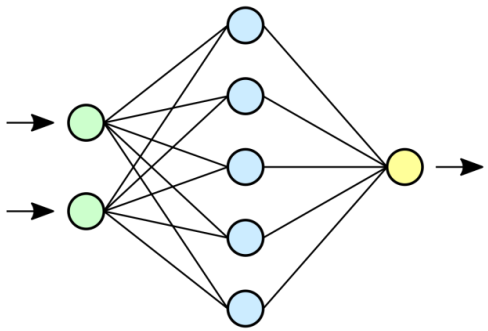
nach Dwyer, Falkai, and Koutsouleris (2018)

Rhethorik der Prädiktiven Modellierung

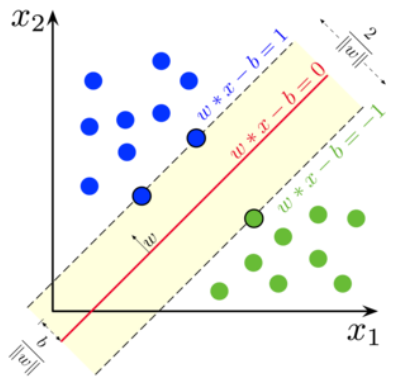
Daten	Trainingsdaten und Testdaten
Statistisches Modell	Modell, Machine Learning Algorithmus
Schätzen von Parametern	Trainieren des Modells, Lernen von Parametern, Supervised Learning

Typische Modelle der Prädiktiven Modellierung

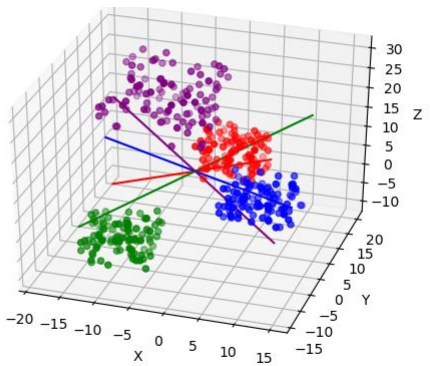
Neuronale Netze | Deep Learning



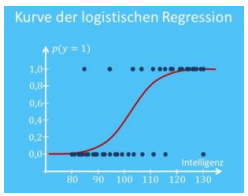
Support Vektor Maschinen



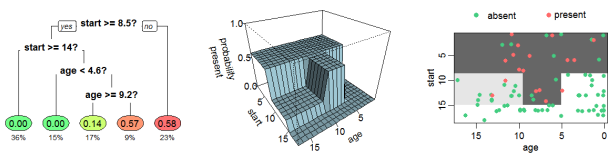
Lineare Diskriminanzanalyse



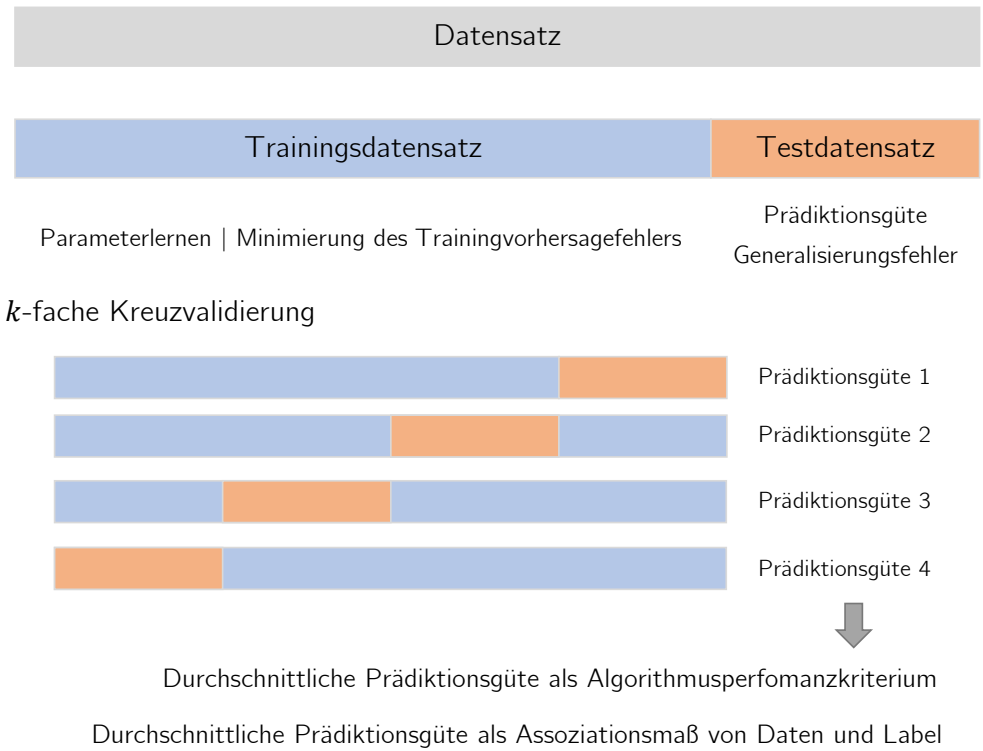
Logistische Regression



Entscheidungsbäume



Modellschätzung- und Modellevaluationsansatz der Prädiktiven Modellierung

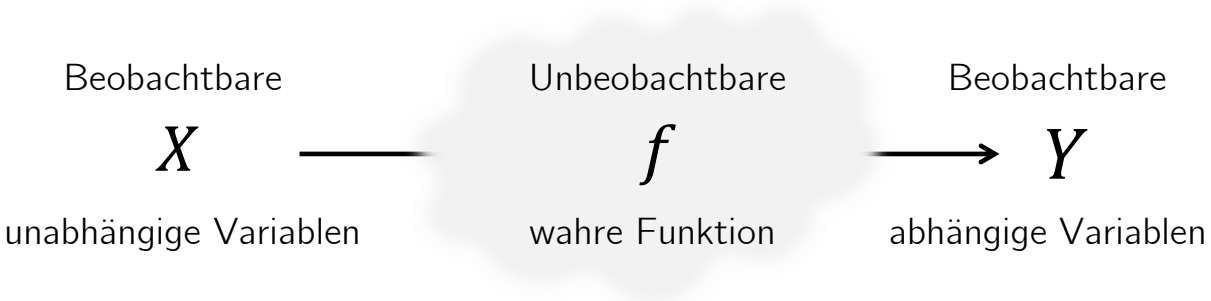


Die theoretische Analyse dieses Ansatzes heißt “Statistische Lerntheorie” (Vapnik 2010)

Explanatorische Modellierung vs. Prädiktive Modellierung

Explanatorische Modellierung ⇔ Wissenschaft

Bestimmung von $\hat{f} \coloneqq \text{argmin} \| f - \hat{f} \|$



Bestimmung von $\tilde{f} \coloneqq \text{argmin} \| Y - \tilde{f}(X) \|$

Prädiktive Modellierung ⇔ Anwendung

- Es gibt keinen Grund anzunehmen, dass immer $\tilde{f} = \hat{f}$ gilt.
- Für ein Beispiel mit $\tilde{f} \neq \hat{f}$. , siehe z.B. Shmueli (2010).

Technik der Prädiktiven Modellierung

Das Lernen von Modellparametern ist das zentrale Problem der Prädiktiven Modellierung

- Parameterlernen impliziert die Minimierung einer Zielfunktion (objective function).
- An der Stelle eines Minimums ist die erste Ableitung einer Funktion gleich Null.
- An der Stelle eines Minimums ist die zweite Ableitung einer Funktion positiv.

Optimierungsverfahren identifizieren Parameterwerte, für die diese Minimumsbedingungen gelten.

Modell	Optimierungsverfahren
Lineare Diskriminanzanalyse	Analytische Optimierung
Logistische Regression	Gradientenverfahren
Support Vektor Maschinen	Optimierung mit Nebenbedingungen
Neuronale Netze	Gradientenverfahren

Anmerkungen

- In der statistischen Modellierung hat die Zielfunktion meist probabilistische Konnotation.
- Typische Zielfunktionen der statistischen Modellierung sind Likelihood Funktionen.
- Prädiktive Modellierung verzichtet zum Teil auf die Repräsentation von Unsicherheit.

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Definition (Funktionenarten)

In der statistischen Anwendung unterscheiden wir

- *univariate reellwertige Funktionen* der Form

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x), \tag{1}$$

- *multivariate reellwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n), \tag{2}$$

- *multivariate vektorwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}^m, x \mapsto f(x) = \begin{pmatrix} f_1(x_1, ..., x_n) \\ \vdots \\ f_m(x_1, ..., x_n) \end{pmatrix}. \tag{3}$$

Bemerkung

- In der Physik werden multivariate reellwertige Funktionen auch *Skalarfelder* genannt.
- In der Physik werden multivariate vektorwertige Funktionen auch *Vektorfelder* genannt.

In diesem Abschnitt betrachten wir univariate reellwertige Funktionen.

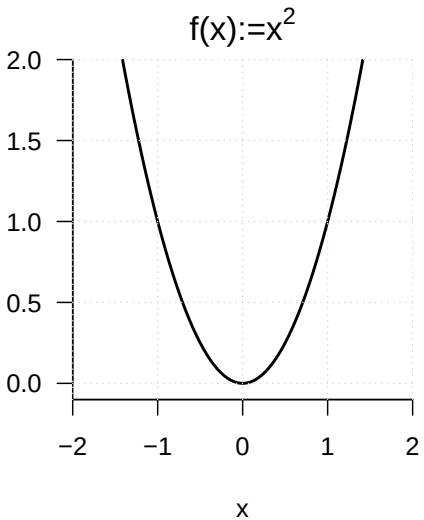
Beispiele

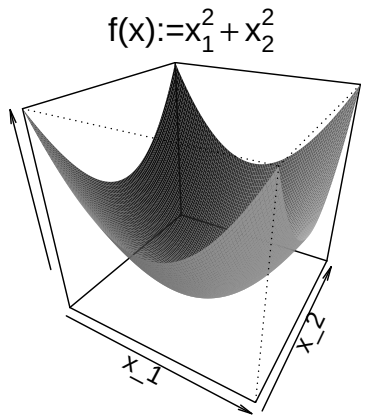
Univariate, reellwertige Funktion

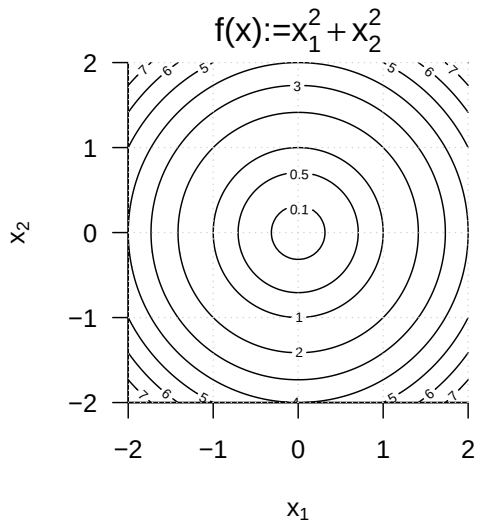
$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := x^2 \tag{4}$$

Multivariate (bivariate), reellwertige Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2 \tag{5}$$







Definition (Ableitung)

Es sei $I \subseteq \mathbb{R}$ ein Intervall und

$$f : I \to \mathbb{R}, x \mapsto f(x) \tag{6}$$

eine univariate reellwertige Funktion. f heißt in $a \in I$ *differenzierbar*, wenn der Grenzwert

$$f'(a) := \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} \tag{7}$$

existiert. $f'(a)$ heißt dann die *Ableitung von f an der Stelle a*. Ist f differenzierbar für alle $x \in I$, so heißt f *differenzierbar* und die Funktion

$$f' : I \to \mathbb{R}, x \mapsto f'(x) \tag{8}$$

heißt *Ableitung von f*

Bemerkungen

- Für $h > 0$ heißt $\frac{f(a+h)-f(a)}{h}$ *Differenzquotient*.
- Der Differenzquotient misst die Änderung $f(a+h) - f(a)$ von f pro Strecke h.
- Für $h \to 0$ misst der Differenzquotient die Änderungsrate von f in a.
- $f'(a)$ ist eine Zahl, f' ist eine Funktion.
- Wir werden keine Grenzwertbildung zur Berechnung von Ableitungen benötigen.

Definition (Notation für Ableitungen univariater reellwertiger Funktionen)

Es sei f eine univariate reellwertige Funktion. Äquivalente Schreibweisen für die Ableitung von f und die Ableitung von f an einer Stelle x sind

(1) die *Lagrange-Notation* f' und $f'(x)$,

(2) die *Newton-Notation* $\dot{f}$ und $\dot{f}(x)$,

(3) die *Leibniz-Notation* $\frac{df}{dx}$ und $\frac{df(x)}{dx}$ und

(4) die *Euler-Notation* Df und $Df(x)$.

Bemerkungen

- Für univariate reellwertige Funktionen benutzen wir f' und $f'(x)$ als Bezeichner.
- In Berechnungen benutzen wir auch die "Operator-Schreibweise" $\frac{d}{dx} f(x)$.
- Wir verstehen $\frac{d}{dx} f(x)$ als den Auftrag, die Ableitung von f zu berechnen.

Definition (Höhere Ableitungen)

Es sei f eine univariate reellwertige Funktion und

$$f^{(1)} := f' \tag{9}$$

sei die Ableitung von f. Die k-te Ableitung von f ist rekursiv definiert durch

$$f^{(k)} := \left(f^{(k-1)}\right)' \text{ für } k \geq 0, \tag{10}$$

unter der Annahme, dass $f^{(k-1)}$ differenzierbar ist. Insbesondere ist die *zweite Ableitung von* f definiert durch die Ableitung von f', also

$$f'' := (f')'. \tag{11}$$

Bemerkungen

- Wir schreiben auch $\frac{d^2}{dx^2} f(x)$ für den Auftrag, die zweite Ableitung von f zu bestimmen.
- Die nullte Ableitung $f^{(0)}$ von f ist f selbst.
- Üblicherweise schreibt man für $k < 4$ f', f'', f''' statt $f^{(1)}$, $f^{(2)}$, $f^{(3)}$.
- Im Allgemeinen benötigen wir nur f' und f''.

Theorem (Rechenregeln für Ableitungen)

Für $i = 1, \ldots, n$ seien g_i reellwertige univariate differenzierbare Funktionen. Dann gelten folgende Rechenregeln:

(1) Summenregel

$$\text{Für } f(x) := \sum_{i=1}^{n} g_i(x) \text{ gilt } f'(x) = \sum_{i=1}^{n} g_i'(x). \tag{12}$$

(2) Produktregel

$$\text{Für } f(x) := g_1(x)g_2(x) \text{ gilt } f'(x) = g_1'(x)g_2(x) + g_1(x)g_2'(x). \tag{13}$$

(3) Quotientenregel

$$\text{Für } f(x) := \frac{g_1(x)}{g_2(x)} \text{ gilt } f'(x) = \frac{g_1'(x)g_2(x) - g_1(x)g_2'(x)}{g_2^2(x)}. \tag{14}$$

(4) Kettenregel

$$\text{Für } f(x) := g_1(g_2(x)) \text{ gilt } f'(x) = g_1'(g_2(x))g_2'(x). \tag{15}$$

Bemerkung

- Für Beweise der Rechenregeln wird auf die einschlägige Literatur verwiesen.

Theorem (Ableitungen elementarer Funktionen)

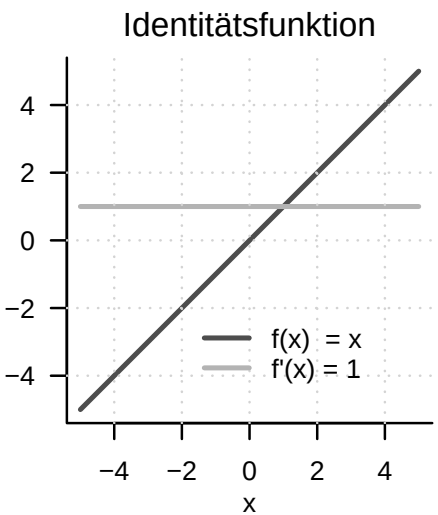
Für einige elementare Funktionen der Datenanalyse ergeben sich folgende Ableitungen:

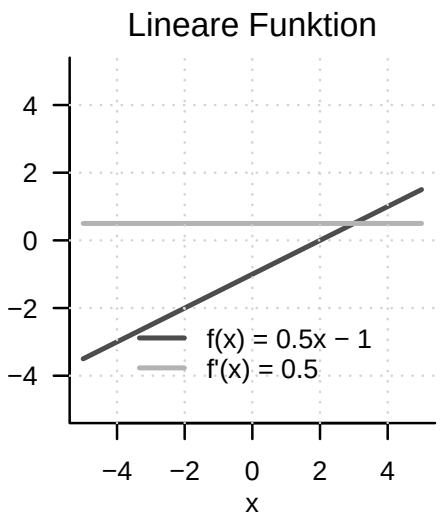
Name	Definition	Ableitung
Polynomfunktionen	$f(x) := \sum_{i=0}^{n} a_i x^i$	$f(x) = \sum_{i=1}^{n} i a_i x^{i-1}$
Konstante Funktion	$f(x) := a$	$f'(x) = 0$
Identitätsfunktion	$f(x) := x$	$f'(x) = 1$
Lineare Funktion	$f(x) := ax + b$	$f'(x) = a$
Quadratfunktion	$f(x) := x^2$	$f'(x) = 2x$
Exponentialfunktion	$f(x) := \exp(x)$	$f'(x) = \exp(x)$
Logarithmusfunktion	$f(x) := \ln(x)$	$f'(x) = \frac{1}{x}$

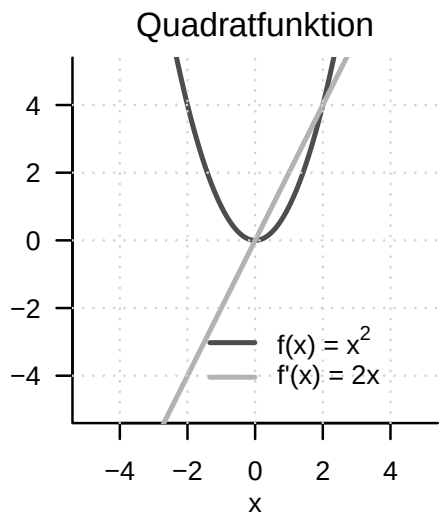
Bemerkung

- Für Beweise wird auf die einschlägige Literatur verwiesen.

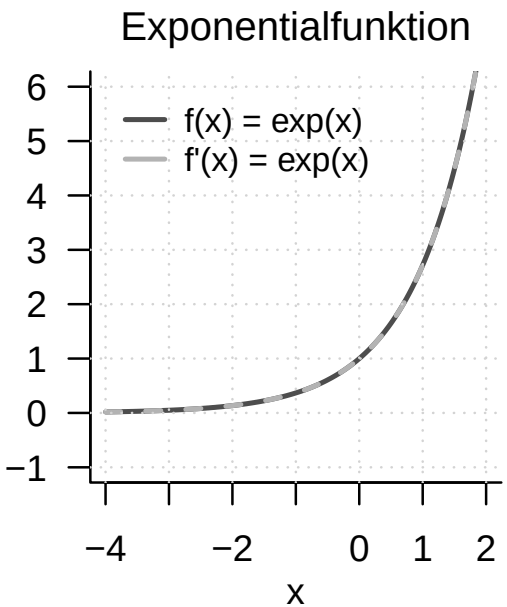
Ableitungen elementarer Funktionen

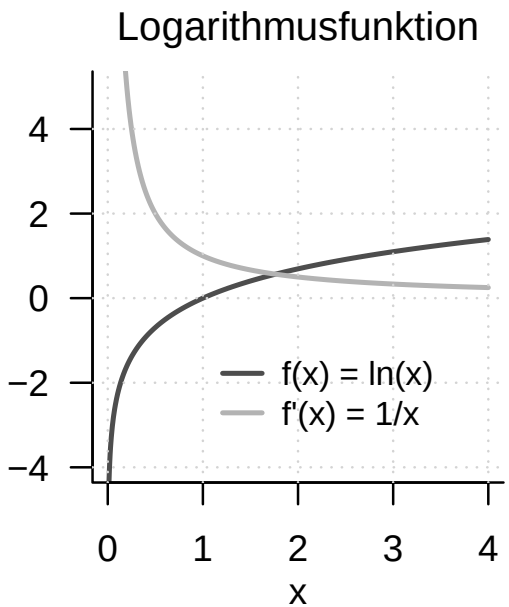






Ableitungen elementarer Funktionen





Definition (Extremstellen und Extremwerte)

Es sei $U \subseteq \mathbb{R}$ und $f : U \to \mathbb{R}$ eine univariate reellwertige Funktion. Dann hat f an der Stelle $x_0 \in U$

- ein *lokales Minimum*, wenn es ein Intervall $I :=]a, b[$ gibt mit $x_0 \in]a, b[$ und
$$f(x_0) \leq f(x) \text{ für alle } x \in I \cap U, \tag{16}$$
- ein *globales Minimum*, wenn gilt, dass
$$f(x_0) \leq f(x) \text{ für alle } x \in U, \tag{17}$$
- ein *lokales Maximum*, wenn es ein Intervall $I :=]a, b[$ gibt mit $x_0 \in]a, b[$ und
$$f(x_0) \geq f(x) \text{ für alle } x \in I \cap U, \tag{18}$$
- *lokales Maximum*, wenn gilt, dass
$$f(x_0) \geq f(x) \text{ für alle } x \in U. \tag{19}$$

Der Wert $x_0 \in U$ der Definitionsmenge von f heißt entsprechend *lokale* oder *globale Minimalstelle* oder *Maximalstelle*, der Funktionswert $f(x_0) \in \mathbb{R}$ heißt entsprechend *lokales* oder *globales Minimum* oder *Maximum*. Generell heißt der Wert $x_0 \in U$ *Extremstelle* und der Funktionswert $f(x_0) \in \mathbb{R}$ *Extremwert*.

Bemerkungen

- Extremstellen werden auch mit $\arg\min_{x \in I \cap U} f(x)$ oder $\arg\max_{x \in I \cap U} f(x)$ bezeichnet.
- Extremwerte werden auch mit $\min_{x \in I \cap U} f(x)$ oder $\max_{x \in I \cap U} f(x)$ bezeichnet.

Definition (Notwendige Bedingung für Extrema)

f sei eine univariate reellwertige Funktion. Dann gilt

$$x_0 \text{ ist Extremstelle von } f \Rightarrow f'(x_0) = 0. \tag{20}$$

Bemerkungen

- Wenn x_0 eine Extremstelle von f ist, dann ist die erste Ableitung von f in x_0 null.
- Sei zum Beispiel x_0 eine lokale Maximalstelle von f. Dann gilt
  - Links von x_0 steigt f an, rechts von x_0 fällt f ab.
  - In x_0 steigt f weder an, noch fällt f ab, also ist $f'(x_0) = 0$.

Definition (Hinreichende Bedingungen für lokale Extrema)

f sei eine zweimal differenzierbare univariate reellwertige Funktion.

- Wenn für $x_0 \in U \subseteq \mathbb{R}$

$$f'(x_0) = 0 \text{ und } f''(x_0) > 0 \tag{21}$$

  gilt, dann hat f an der Stelle x_0 ein Minimum.

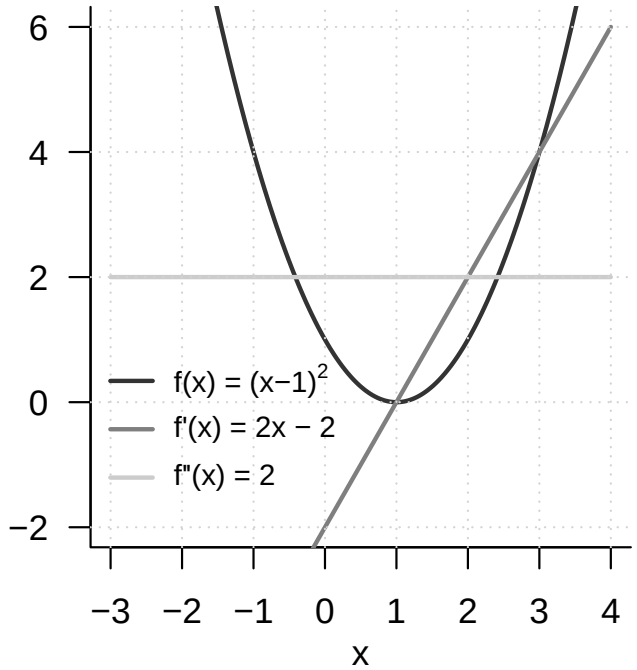
- Wenn für $x_0 \in U \subseteq \mathbb{R}$

$$f'(x_0) = 0 \text{ und } f''(x_0) < 0 \tag{22}$$

  gilt, dann hat f an der Stelle x_0 ein Maximum.

Bemerkung

- Eine Intuition vermittelt nachfolgende Abbildung.



Hier ist offenbar $x_0 = 1$ eine lokale Minimalstelle von $f(x) = (x-1)^2$. Man erkennt:

- Links von x_0 fällt f ab, rechts von x_0 steigt f an.
- In x_0 steigt f weder an, noch fällt f ab, also ist $f'(x_0) = 0$.
- Links und rechts von x_0 und in x_0 ist die Änderung f'' von f' positiv.
- Links von x_0 schwächt sich die Negativität von f' zu 0 ab.
- Rechts von x_0 verstärkt sich die Positivität von f'.

Definition (Standardverfahren der analytischen Optimierung)

f sei eine univariate reellwertige Funktion. Lokale Extremstellen von f können mit folgendem *Standardverfahren der analytischen Optimierung* identifiziert werden:

(1) Berechnen der ersten und zweiten Ableitung von f.

(2) Bestimmen von Nullstellen x^* von f' durch Auflösen von $f'(x^*) = 0$ nach x^*.

$\Rightarrow$ Nullstellen von f' sind Kandidaten für Extremstellen von f.

(3) Evaluation von $f''(x^*)$.

$\Rightarrow$ Wenn $f''(x^*) > 0$, dann ist x^* lokale Minimumstelle von f.

$\Rightarrow$ Wenn $f''(x^*) < 0$, dann ist x^* lokale Maximumstelle von f.

$\Rightarrow$ Wenn $f''(x^*) = 0$, dann ist x^* keine Extremstelle von f.

Beispiel

Wir betrachten die Funktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := (x-1)^2. \tag{23}$$

Die erste Ableitung von f ergibt sich mit der Kettenregel zu

$$f'(x) = \frac{d}{dx}\left((x-1)^2\right) = 2(x-1) \cdot \frac{d}{dx}(x-1) = 2x - 2. \tag{24}$$

Die zweite Ableitung von f ergibt sich zu

$$f''(x) = \frac{d}{dx} f'(x) = \frac{d}{dx}(2x-2) = 2 > 0 \text{ für alle } x \in \mathbb{R}. \tag{25}$$

Auflösen von $f'(x^*) = 0$ nach x^* ergibt

$$f'(x^*) = 0 \Leftrightarrow 2x^* - 2 = 0 \Leftrightarrow 2x^* = 2 \Leftrightarrow x^* = 1. \tag{26}$$

$x^* = 1$ ist folglich eine Minimalstelle von f mit zugehörigen Minimalwert $f(1) = 0$.

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Definition (Funktionenarten)

In der statistischen Anwendung unterscheiden wir

- *univariate reellwertige Funktionen* der Form

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x), \tag{27}$$

- *multivariate reellwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n), \tag{28}$$

- *multivariate vektorwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}^m, x \mapsto f(x) = \begin{pmatrix} f_1(x_1, ..., x_n) \\ \vdots \\ f_m(x_1, ..., x_n) \end{pmatrix}. \tag{29}$$

In diesem Abschnitt betrachten wir multivariate reellwertige Funktionen.

Beim Parameterlernen der Prädiktiven Modellierung ist

$x \in \mathbb{R}^n$ **ein Vektor von Parameterwerten und** $f : \mathbb{R}^n \to \mathbb{R}$ **eine Zielfunktion**.

Definition (Partielle Ableitung)

Es sei $D \subseteq \mathbb{R}^n$ eine Menge und

$$f : D \to \mathbb{R}, x \mapsto f(x) \tag{30}$$

eine multivariate reellwertige Funktion. f heißt in $x \in D$ *nach* x_i *partiell differenzierbar*, wenn der Grenzwert

$$\frac{\partial}{\partial x_i} f(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{31}$$

existiert. $\frac{\partial}{\partial x_i} f(x)$ heißt dann die *partielle Ableitung von* f *nach* x_i *an der Stelle* x. Wenn f für alle $x \in D$, nach x_i partiell differenzierbar ist, dann heißt f *nach* x_i *partiell differenzierbar* und die Funktion

$$\frac{\partial}{\partial x_i} f : D \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_i} f(x) \tag{32}$$

heißt *partielle Ableitung von* f *nach* x_i.

f heißt *partiell differenzierbar in* $x \in D$, wenn f für alle $i = 1, \ldots, n$ in $x \in D$ nach x_i partiell differenzierbar ist, und f heißt partiell differenzierbar, wenn f für alle $i = 1, \ldots, n$ in allen $x \in D$ nach x_i partiell differenzierbar ist.

Bemerkungen

- $e_i \in \mathbb{R}^n$ bezeichnet den $i$ten Einheitsvektor.
- $\frac{f(x+he_i)-f(x)}{h}$ misst die Änderung $f(x + he_i) - f(x)$ von f pro Strecke h in Richtung e_i.
- Für $h \to 0$ misst der Differenzquotient die Änderungsrate von f in x in Richtung e_i.
- $\frac{\partial}{\partial x_i} f(x)$ ist eine Zahl, $\frac{\partial}{\partial x_i} f$ ist eine Funktion.
- Praktisch berechnet man $\frac{\partial}{\partial x_i} f$ als die (einfache) Ableitung

$$\frac{d}{dx_i} \tilde{f}_{x_1,\dots x_{i-1},x_{i+1},\dots,x_n}(x_i) \tag{33}$$

der univariaten reellwertigen Funktion

$$\tilde{f} : \mathbb{R} \to \mathbb{R}, x_i \mapsto \tilde{f}_{x_1,\dots x_{i-1},x_{i+1},\dots,x_n}(x_i) := f(x_1, ..., x_i, ..., x_n). \tag{34}$$

- Man betrachtet alle x_j mit $j \neq i$ also als Konstanten.

Beispiel (1)

Wir betrachten die Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{35}$$

Weil die Definitionsmenge dieser Funktion zweidimensional ist, kann man zwei partielle Ableitungen berechnen

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) \text{ und } \frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x). \tag{36}$$

Um die erste dieser partiellen Ableitungen zu berechnen, betrachtet man die Funktion

$$f_{x_2} : \mathbb{R} \to \mathbb{R}, x_1 \mapsto f_{x_2}(x_1) := x_1^2 + x_2^2, \tag{37}$$

wobei x_2 hier die Rolle einer Konstanten einnimmt. Um explizit zu machen, dass x_2 kein Argument der Funktion ist, die Funktion aber weiterhin von x_2 abhängt haben wir die Subskriptnotation $f_{x_2}(x_1)$ verwendet. Um nun die partielle Ableitung zu berechnen, berechnen wir die (einfache) Ableitung von f_{x_2},

$$f'_{x_2}(x) = 2x_1. \tag{38}$$

Es ergibt sich also

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) = \frac{\partial}{\partial x_1}(x_1^2 + x_2^2) = f'_{x_2}(x) = 2x_1. \tag{39}$$

Analog gilt mit der entsprechenden Formulierung von f_{x_1}, dass

$$\frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x) = \frac{\partial}{\partial x_2}(x_1^2 + x_2^2) = f'_{x_1}(x) = 2x_2. \tag{40}$$

Definition (Zweite partielle Ableitungen)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion und $\frac{\partial}{\partial x_i} f$ sei die partielle Ableitung von f nach x_i. Dann ist die zweite partielle Ableitung von f nach x_i und x_j definiert als

$$\frac{\partial^2}{\partial x_j x_i} f(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial}{\partial x_i} f \right) \tag{41}$$

Bemerkungen

- Wie die zweite Ableitung ist auch die zweite partielle Ableitung rekursiv definiert.
- Zu jeder partiellen Ableitung $\frac{\partial}{\partial x_i} f$ gibt es n zweite partiellen Ableitungen $\frac{\partial^2}{\partial x_j \partial x_i} f, j = 1, \ldots, n$.

Theorem (Satz von Schwarz)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine partiell differenzierbare multivariate reellwertige Funktion. Dann gilt

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \text{ für alle } 1 \leq i, j \leq n. \tag{42}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem von Schwarz besagt, dass die Reihenfolge des partiellen Ableitens irrelevant ist
- Das Theorem erleichtert die Berechnung von zweiten partiellen Ableitungen.
- Das Theorem hilft, Fehler bei der Berechnung zweiter partieller Ableitungen aufzudecken.

Beispiel (1) (fortgeführt)

Wir wollen die partiellen Ableitungen zweiter Ordnung der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{43}$$

berechnen. Mit den Ergebnissen für die partiellen Ableitungen erster Ordnung dieser Funktion ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 x_1} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_1} (2x_1) = 2 \\
\frac{\partial^2}{\partial x_1 x_2} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_1} (2x_2) = 0 \\
\frac{\partial^2}{\partial x_2 x_1} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_2} (2x_1) = 0 \\
\frac{\partial^2}{\partial x_2 x_2} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_2} (2x_2) = 2
\end{aligned} \tag{44}$$

Offenbar gilt

$$\frac{\partial^2}{\partial x_1 x_2} f(x) = \frac{\partial^2}{\partial x_2 x_1} f(x). \tag{45}$$

Beispiel (2)

Wir wollen die partiellen Ableitungen erster und zweiter Ordnung der Funktion

$$f : \mathbb{R}^3 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_1 x_2 + x_2 \sqrt{x_3}. \tag{46}$$

berechnen.

Mit den Rechenregeln für Ableitungen ergibt sich für die partiellen Ableitungen erster Ordnung

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x) &= \frac{\partial}{\partial x_1} \left(x_1^2 + x_1 x_2 + x_2 \sqrt{x_3} \right) = 2x_1 + x_2, \\
\frac{\partial}{\partial x_2} f(x) &= \frac{\partial}{\partial x_2} \left(x_1^2 + x_1 x_2 + x_2 \sqrt{x_3} \right) = x_1 + \sqrt{x_3}, \\
\frac{\partial}{\partial x_3} f(x) &= \frac{\partial}{\partial x_3} \left(x_1^2 + x_1 x_2 + x_2 \sqrt{x_3} \right) = \frac{x_2}{2\sqrt{x_3}}.
\end{aligned} \tag{47}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_1 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_1}(2x_1 + x_2) = 2, \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_2}(2x_1 + x_2) = 1, \\
\frac{\partial^2}{\partial x_3 \partial x_1} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_3}(2x_1 + x_2) = 0.
\end{aligned} \tag{48}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_2 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_1}(x_1 + \sqrt{x_3}) = 1, \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_2}(x_1 + \sqrt{x_3}) = 0, \\
\frac{\partial^2}{\partial x_3 \partial x_2} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_3}(x_1 + \sqrt{x_3}) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{49}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_3 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_1}\left(\frac{x_2}{2}\sqrt{x_3}\right) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_2}\left(\frac{x_2}{2\sqrt{x_3}}\right) = \frac{1}{2\sqrt{x_3}}, \\
\frac{\partial^2}{\partial x_3 \partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_3}\left(x_2 \frac{1}{2} x_3^{-\frac{1}{2}}\right) = -\frac{1}{4} x_2 x_3^{-\frac{3}{2}}.
\end{aligned} \tag{50}$$

Weiterhin erkennt man, dass die Reihenfolge der partiellen Ableitungen irrelevant ist, denn es gilt

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(x) = 1, \\
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_1} f(x) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_2} f(x) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{51}$$

Definition (Gradient)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist der *Gradient* $\nabla f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^n. \tag{52}$$

Bemerkung

- $\nabla f(x)$ fasst die partiellen Ableitungen von f an der Stelle $x \in \mathbb{R}^n$ in einem Vektor zusammen.
- Gradienten sind multivariate vektorwertige Abbildungen der Form $\nabla f : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto \nabla f(x)$.
- Wir zeigen später, dass $-\nabla f(x)$ die Richtung des steilsten Abstiegs von f in $\mathbb{R}^n$ anzeigt.
- Für $n = 1$ gilt $\nabla f(x) = f'(x)$.

Beispiele

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \in \mathbb{R}^2. \tag{53}$$

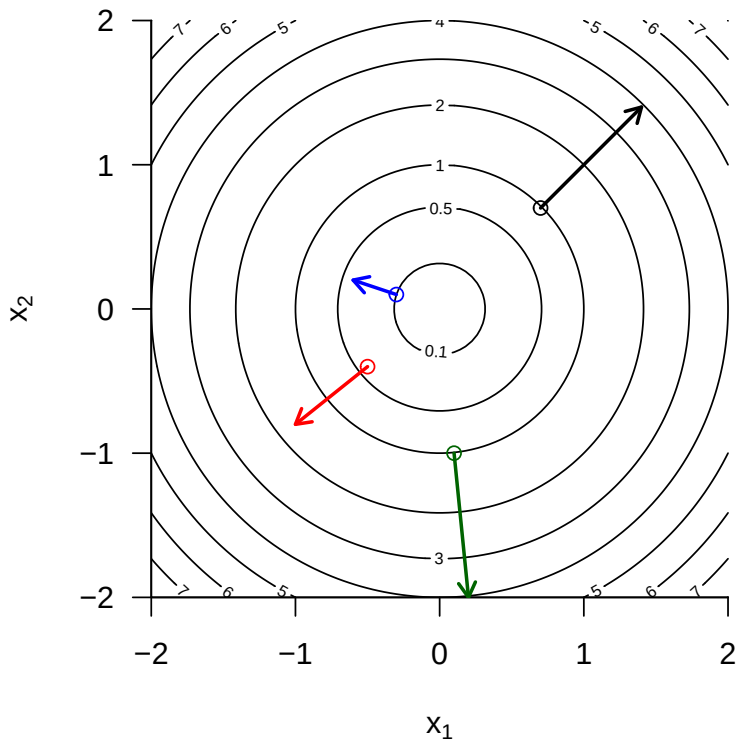
Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + \sqrt{x_3} \\ \frac{x_2}{2\sqrt{x_3}} \end{pmatrix} \in \mathbb{R}^3. \tag{54}$$

Beispiel (1) (fortgeführt)

Gradienten von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$ bei

$$x = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} \qquad x = \begin{pmatrix} -0.3 \\ 0.1 \end{pmatrix} \qquad x = \begin{pmatrix} -0.5 \\ -0.4 \end{pmatrix} \qquad x = \begin{pmatrix} 0.1 \\ -1.0 \end{pmatrix}$$



Definition (Hesse-Matrix)

$f : \mathbb{R}^n \to \mathbb{R}$ sei ein multivariate reellwertige Funktion. Dann ist die *Hesse-Matrix* $\nabla^2 f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 x_n} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n x_1} f(x) & \frac{\partial^2}{\partial x_n x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n x_n} f(x) \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{55}$$

Bemerkung

- $\nabla^2 f(x)$ fasst die partiellen Ableitungen zweiter Ordnung von f in einer Matrix zusammen.
- Hesse-Matrizen sind multivariate matrixwertige Abbildungen der Form $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}, x \mapsto \nabla^2 f(x)$.
- Für $n = 1$ gilt $\nabla^2 f(x) = f''(x)$.
- Mit $\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x)$ für $1 \leq i, j \leq n$ folgt, dass $\left(\nabla^2 f(x)\right)^T = \nabla^2 f(x)$.

Beispiel

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\begin{aligned} \nabla^2 f(x) &:= \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) & \frac{\partial^2}{\partial x_1 x_3} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) & \frac{\partial^2}{\partial x_2 x_3} f(x) \\ \frac{\partial^2}{\partial x_3 x_1} f(x) & \frac{\partial^2}{\partial x_3 x_2} f(x) & \frac{\partial^2}{\partial x_3 x_3} f(x) \end{pmatrix} \\ &:= \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{4} x_2 x_3^{-3/2} \end{pmatrix} \end{aligned}$$

Definition (Glatte multivariate reellwertige Funktion)

Eine multivariate reellwertige Funktion

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) \tag{56}$$

heißt *glatt*, wenn ihr Gradient und ihre Hesse-Matrix existieren und für alle $x \in \mathbb{R}^n$ stetig sind.

Bemerkungen

- Der Gradient und die Hesse-Matrix einer glatten Funktion könnten überall in $\mathbb{R}^n$ berechnet werden.

Theorem (Multivariater Mittelwertsatz erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

$$f(x + p) = f(x) + \nabla f(x + tp)^T p. \tag{57}$$

Theorem (Multivariater Mittelwertsatz zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

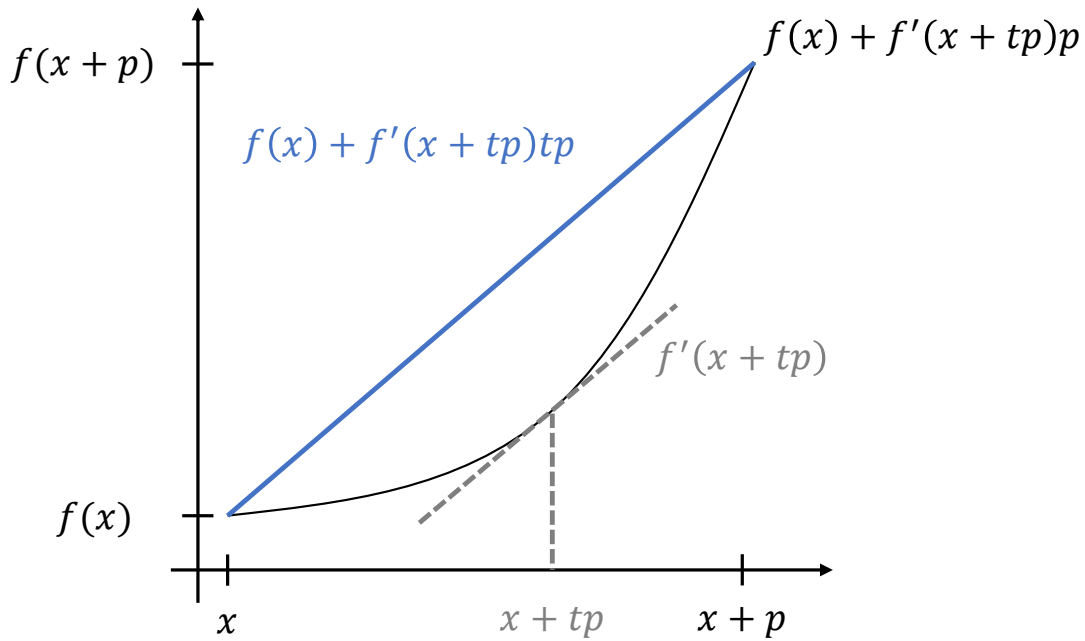
$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p. \tag{58}$$

Bemerkung

- Wir verzichten auf Beweise.
- Nocedal and Wright (2006) bezeichnen die Theoreme als “Taylor’s Theorem”, das ist ein wenig misleading.
- ∇f und $\nabla^2 f$ werden an einer Stelle zwischen x und $x + p$ evaluiert.

Univariate visuelle Intuition zum Mittelwertsatz 1. Ordnung

$$\text{Es gibt ein } t \in]0,1[\text{ mit } f(x+p) = f(x) + f'(x+tp)p \tag{59}$$



Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Definition (Optimierungsproblem)

Ein *Optimierungsproblem* hat die allgemeine Form

$$\min_x f(x), \tag{60}$$

wobei $x \in \mathbb{R}^n$ und $f : \mathbb{R}^n \to \mathbb{R}$ eine glatte multivariate reellwertige Funktion ist. Weil gilt, dass

$$\max_x f(x) = \min_x -f(x) \tag{61}$$

genügt es, sich mit Minimierungsproblemen zu befassen.

Definition (Globale und lokale Minimierer, globale und lokale Minima)

$f : \mathbb{R}^n \to \mathbb{R}$ sei ein multivariate reellwertige Funktion.

- $x^* \in \mathbb{R}^n$ heißt globaler Minimalstelle von f, wenn $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$ gilt. $f(x^*) \in \mathbb{R}$ heißt das globale Minimum von f.

- $x^* \in \mathbb{R}^n$ heißt lokale Minimalstelle von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) \leq f(x)$ für alle $x \in N \subset \mathbb{R}^n$. In diesem Fall heißt $f(x^*) \in \mathbb{R}$ lokales Minimum von f.

- $x^* \in \mathbb{R}^n$ heißt strikte lokale Minimalstelle von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) < f(x)$ für alle $x \in N \subset \mathbb{R}^n$. In diesem Fall heißt $f(x^*) \in \mathbb{R}$ striktes lokales Minimum von f.

Bemerkung

- Eine Umgebung von $x \in \mathbb{R}^n$ ist eine offene Menge, die x enthält.

Theorem (Notwendige Bedingung erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann gilt

$$\nabla f(x^*) = 0_n\,. \tag{62}$$

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Dazu nehmen wir an, dass x^* zwar eine lokale Minimalstelle von f ist, aber $\nabla f(x^*) \neq 0_n$ ist. Dazu definieren wir zunächst $p := -\nabla f(x^*)$. Dann gilt, dass

$$p^T \nabla f(x^*) = -\nabla f(x^*)^T \nabla f(x^*) = -||\nabla f(x^*)|| < 0. \tag{63}$$

Weil ∇f in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass auch

$$p^T \nabla f(x^* + tp) < 0 \text{ für alle } t \in [0, T]. \tag{64}$$

gilt. Nun gilt für $\tilde{t} \in]0, T]$ aber mit dem Mittelwertssatz erster Ordnung, dass

$$f(x^* + \tilde{t}p) = f(x^*) + \nabla f(x^* + tp)^T \tilde{t}p = f(x^*) + \tilde{t}p^T \nabla f(x^* + tp) \text{ für ein } t \in]0, \tilde{t}[. \tag{65}$$

Also folgt $f(x^* + \tilde{t}p) < f(x^*)$ für alle $\tilde{t} \in]0, T]$. Wir haben also eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla f(x^*) \neq 0_n$ gilt. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla f(x^*) \neq 0_n$ gilt. Also muss $\nabla f(x^*) = 0_n$ gelten, wenn x^* eine lokale Minimalstelle ist.

Theorem (Notwendige Bedingung zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann ist $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ ist positiv semidefinit.

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Wir haben schon gesehen, dass $\nabla f(x^*) = 0_n$ ist, wenn x^* eine lokale Minimalstelle von f. ist. Für einen Widerspruchsbeweis nehmen wir nun an, dass x^* zwar eine lokale Minimalstelle von f ist, aber dass $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dann ist es möglichen einen Vektor p zu finden, so dass gilt

$$p^T \nabla^2 f(x^*) p < 0. \tag{66}$$

Weil $\nabla^2 f(x^*)$ in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass

$$p^T \nabla^2 f(x^* + tp) p < 0 \text{ für alle } t \in [0, T]. \tag{67}$$

gilt. Mithilfe des Mittelwertsatzes zweiter Ordnung gilt dann für alle $\bar{t} \in]0, T]$ und ein $t \in]0, \bar{t}[$, dass

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x^*) + \frac{1}{2} \bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*). \tag{68}$$

Wir haben also wieder eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Also muss $\nabla^2 f(x^*)$ positiv semidefinit sein, wenn x^* eine lokale Minimalstelle ist.

Theorem (Hinreichende Bedingungen zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es seien $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ positiv definit. Dann ist x^* eine strikte Minimalstelle von f.

Beweis

Wir halten zunächst fest, dass weil die Hesse-Matrix stetig und positiv definit in x^* ist, wir ein $r > 0$ wählen können, so dass $\nabla^2 f(x)$ positiv definit für alle x in

$$D = \{x | ||x - x^*|| < r\} \tag{69}$$

ist. Für einen Vektor p mit $||p|| > 0$ und $||p|| < r$ gilt $x^* + p \in D$. Für ein $t \in]0, 1[$ gilt dann mit dem Mittelwertssatz zweiter Ordnung, dass

$$\begin{aligned} f(x^* + p) &= f(x^*) + \nabla f(x^*) p^T + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p. \end{aligned} \tag{70}$$

Weil aber $x^* + tp \in D$ ist, gilt, dass $p^T \nabla^2 f(x^* + tp) p > 0$ ist und somit $f(x^* + p) > f(x^*)$. In jeder Richtung p von x^* weg erhöht sich also der Wert von f und damit ist x^* eine strikte Minimalstelle.

Theorem (Minmalstellen konvexer Funktionen)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine *konvexe Funktion*, das heißt, für alle $x, y \in \mathbb{R}^n$ gelte

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \text{ für alle } \lambda \in [0,1]. \tag{71}$$

Dann ist eine lokale Minimalstelle x^* von f auch die globale Minimalstelle von f.

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Nehmen wir dazu an, x^* sei eine lokale, aber keine globale Minimalstelle. Dann können wir ein $z \in \mathbb{R}^n$ mit $f(z) < f(x^*)$ finden. Wir betrachten nun die Strecke, die x^* und z in $\mathbb{R}^n$ verbindet, also

$$x = \lambda z + (1-\lambda) x^* \text{ mit } \lambda \in]0,1] \tag{72}$$

Mit der Konvexität von f folgt dann aber, dass

$$f(x) \leq \lambda f(z) + (1-\lambda) f(x^*) < f(x^*) \tag{73}$$

Jede Umgebung N von x^* enthält ein Stück dieser Strecke, also gibt es immer Punkte $x \in N$ mit $f(x) < f(x^*)$. Also ist $x*$ keine lokale Minimalstelle und wir haben einen Widerspruch.

Zusammenfassung

Optimierungsproblem

$$\min_x f(x) = \max_x -f(x) \text{ für } f : \mathbb{R}^n \to \mathbb{R}$$

Lokale Minimalstelle

$$x^* = \arg\min_x f(x), x^* \in \mathbb{R}^n \Leftrightarrow f(x^*) \leq f(x) \text{ für alle } x \in N \subset \mathbb{R}^n$$

Notwendige Bedingung erster Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n$$

Notwendige Bedingung zweiter Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n \text{ und } \nabla^2 f(x) \text{ positiv semidefinit}$$

Hinreichende Bedingung zweiter Ordnung

$$\nabla f(x^*) = 0 \text{ und } \nabla^2 f(x) \text{ positiv definit } \Rightarrow x^* = \arg\min_x f(x)$$

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Allgemeine Form von Optimierungsalgorithmen

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, \ldots$

1. Berechnung von x_{k+1} basierend auf Information über f an der Stelle x_k.
2. STOP, wenn Minimalstelle gefunden ist oder kein Fortschritt mehr erzielt wird.

Gradientenverfahren

Initialisierung

0. Wahl von Startpunkt $x_0 \in \mathbb{R}^n$, Lernrate $\alpha > 0$, Konvergenzkriteriums $\delta > 0$.

Iterationen

Für $k = 0, 1, 2, \ldots$

1. Setze $x_{k+1} := x_k - \alpha \nabla f(x_k)$.
2. STOP, wenn $||\nabla f(x_{k+1})|| < \delta$, ansonsten gehe zu 1.

Theorem (Gradientenverfahren)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $x_k \in \mathbb{R}^n$. Dann ist die Gradientenrichtung

$$p_k^G := -\nabla f(x_k) \tag{74}$$

die Richtung des steilsten Abstiegs von f in x_k.

Bemerkungen

- Es gibt unendliche viele mögliche Richtungen p in x_k.
- $\nabla f(x) \in \mathbb{R}^n$ ist eine Richtung in der Definitionsmenge von f (Parameterraum).
- Die Gradientenrichtung ist davon die Richtung, in der die Zielfunktion f am schnellsten abnimmt.
- Zum Vergleich von Richtungen genügt es, Richtungen der Länge $||p|| = 1$ zu vergleichen.

Beweis

Mit dem Mittelwertsatz zweiter Ordnung gilt für jede Richtung p und Schrittlängenparameter α, dass

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f(x_k) + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p \text{ für ein } t \in]0, \alpha[. \tag{75}$$

Die Änderungsrate von f in Richtung p in x_k ist also der Koeffizient von α, also $p^T \nabla f(x_k)$ (man denke an $x = tv$ für einen Ort x, eine Geschwindigkeit v und eine Zeit t). Also gilt, dass die Richtung des steilsten Abstiegs p in x_k mit Länge 1 die Lösung des Optimierungsproblems

$$\min_p p^T \nabla f(x_k) \text{ mit der Nebenbedingung } ||p|| = 1. \tag{76}$$

ist. Wir erinnern nun zunächst daran, dass für $x, y \in \mathbb{R}^n$ gilt der Kosinus des Winkel zwischen x und y durch

$$\cos\alpha = \frac{\langle x, y\rangle}{||x||||y||} = \frac{x^T y}{||x||||y||} \tag{77}$$

gegeben ist. Damit aber gilt, dass

$$p^T \nabla f(x_k) = ||p|| \cdot ||\nabla f(x_k)|| \cos\theta = 1 \cdot ||\nabla f(x_k)|| \cos\theta = ||\nabla f(x_k)|| \cos\theta \tag{78}$$

und somit liegt hier bei $\cos\theta = -1$ eine Minimalstelle vor. Dies bedeutet aber, dass die minimierende Länge p exakt antiparallel zu $\nabla f(x_k)$ und von Länge 1 sein muss. Also ist die Minimalstelle des Optimierungsproblems

$$p = \frac{-\nabla f(x_k)}{||\nabla f(x_k)||}. \tag{79}$$

Damit ist $p_k^G := -\nabla f(x_k)$ aber der Richtungsvektor beliebiger Länge in der die Abnahme von f maximal ist.

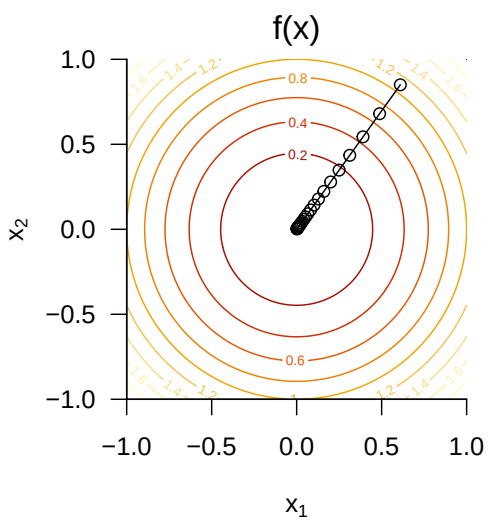
Beispiel

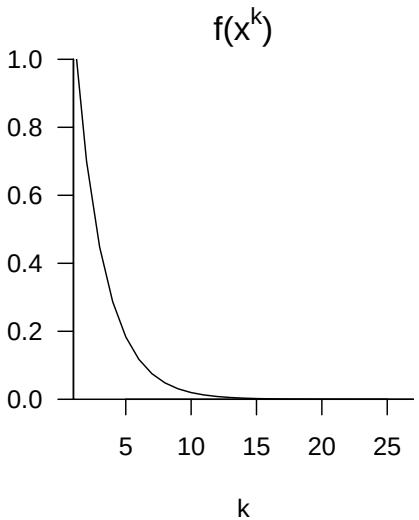
Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$

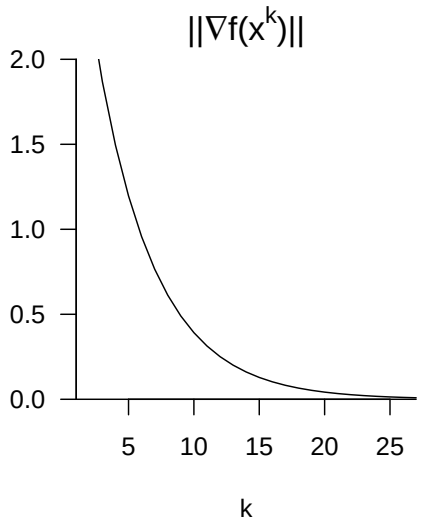
```
# Funktionsdefinitionen
# ------------------------------------------------------------------------------
# Zielfunktion
f = function(x) {
  return(x[1]^2 + x[2]^2)                  # f(x) := x_1^2 + x_2^2
}
# Gradient der Zielfunktion
nabla_f = function(x) {
  return(matrix(c(2*x[1],2*x[2]),          # \nabla f(x) := (2x_1, 2x_2)^T
                nrow = 2))
}
# Gradientenverfahren
# ------------------------------------------------------------------------------
# Parameter
n      = 2                               # Dimension
alpha  = 1e-1                            # Lernrate
delta  = 1e-2                            # Konvergenzkriterium

# Initialisierung
x_k    = matrix(c(.61,.85), nrow = 2)    # Zufälliger Startpunkt in [0,1]^2
x      = x_k                             # Initialisierung Iteranden
fx     = f(x_k)                          # Initialisierung Funktionswerte
crt    = norm(nabla_f(x_k))              # Initialisierung Kriterium

# Iterationen
while(norm(nabla_f(x_k)) > delta){

  # Argumentupdate
  x_k    = x_k - alpha*nabla_f(x_k)

  # Dokumentation
  x      = cbind(x,x_k)
  fx     = c(fx,f(x_k))
  crt    = c(crt, norm(nabla_f(x_k)))
}
```

Beispiel

Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$







Liniensuchverfahren als generalisierte Gradientenverfahren

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, \ldots$

1. Wahl einer Abstiegsrichtung p_k
2. Wahl eines Lernparameters $\alpha_k \approx \min_\alpha f(x_k + \alpha p_k)$.
3. Setze $x_{k+1} := x_k + \alpha_k p_k$.
4. Konvergenztest.

$\Rightarrow$ Die Wahl sinnvoller Lernraten α_k ist für eine gute Performanz entscheidend!

(vgl. Ostwald and Starke (2016))

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Definition (Optimierungsproblem mit Nebenbedingungen)

Ein *Optimierungsproblem mit Nebenbedingungen* hat die allgemeine Form

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I, \tag{80}$$

wobei $f : \mathbb{R}^n \to \mathbb{R}$ und $c_i : \mathbb{R}^n \to \mathbb{R}, i \in E \cup I$ glatte multivariate reellwertige Funktions und E, I endliche Indexmengen sind. f heißt *Zielfunktion*, die $c_i, i \in E$ heißen *Gleichungsnebenbedingungen* und die $c_i, i \in I$ heißen *Ungleichungsnebenbedingungen*. Die Menge

$$\mathcal{X} := \{x \in \mathbb{R}^n | c_i(x) = 0, i \in E \text{ und } c_i(x) \geq 0, i \in I\} \tag{81}$$

heißt *feasible set*.

Bemerkung

- Die notwendigen Bedingungen für Minimalstellen bei Optimierungsproblem ohne Nebenbedingungen sind für $n = 1$: $f'(x^*) = 0$ und für $n > 1$: $\nabla f(x^*) = 0_n$. Im Folgenden führen wir analoge notwendige Bedingungen erster Ordnung für Minimalstellen bei Optimierungsproblemen mit Nebenbedingungen ein.

Beispiel

Definition (Quadratisches Programm)

Ein *Quadratisches Programm* ist das konvexe Optimierungsproblem mit den Nebenbedingungen

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T P x + q^T x \text{ u.d.N. } Ax = b \text{ und } -Gx + h \geq 0, \tag{82}$$

wobei

- $P \in \mathbb{R}^{n \times n}$ eine positiv definite Matrix ist,
- $q \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$ sind und
- $G \in \mathbb{R}^{m \times n}$, und $h \in \mathbb{R}^m$ sind.

Bemerkungen

- Quadratische Programme sind Optimierungsprobleme mit Nebenbedingungen.
- Parameterlernen bei Support Vektor Maschinen führt auf ein Quadratisches Programm.
- Optimierungstoolboxen enhalten Funktionen zur Lösung Quadratischer Programme.
- In R bietet sich das Paket `quadprog` an.

Definition (Lagrange Funktion, Lagrange Multiplikatoren)

Es sei

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I, \tag{83}$$

ein Optimierungsproblem mit Nebenbedingungen. Dann ist die *Lagrange Funktion* dieses Problems definiert als

$$L : \mathbb{R}^n \times \mathbb{R}^{|E \cup I|} \to \mathbb{R}, (x, \lambda) \mapsto L(x, \lambda) := f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x). \tag{84}$$

Hierbei wird $\lambda \in \mathbb{R}^{|E \cup I|}$ *Lagrange-Multiplikatoren Vektor* genannt und die einzelnen $\lambda_i \in \mathbb{R}$ mit $i \in E \cup I$ werden *Lagrange Multiplikatoren*genannt.

Bemerkung

- Die Lagrange Funktion und die Lagrange Multiplikatoren nehmen in den notwendigen Bedingungen der Optimierung mit Nebenbedingungen eine zentrale Rolle ein.

Definition (Notwendige Bedingungen erster Ordnung)

x^* sei eine lokale Lösung des Optimierungsproblems

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c_i(x) = 0, i \in E, c_i(x) \geq 0, i \in I. \tag{85}$$

Dann gibt es einen Lagrange-Multiplikatoren Vektor $\lambda^* \in \mathbb{R}^{|E \cup I|}$ mit den Komponenten $\lambda_i^*, i \in E \cup I$, so dass die folgenden Bedingungen an der Stelle $(x^*, \lambda^*) \in \mathbb{R}^{n+|E \cup I|}$ gelten

$$\begin{aligned}
\nabla_x L(x^*, \lambda^*) &= 0 \\
c_i(x^*) &= 0 \text{ für alle } i \in E \\
c_i(x^*) &\geq 0 \text{ für alle } i \in I \\
\lambda_i^* &\geq 0 \text{ für alle } i \in I \\
\lambda_i^* c_i(x^*) &= 0 \text{ für alle } i \in E \cup I
\end{aligned}$$

Bemerkungen

- Die Bedingungen werden auch *Karush-Kuhn-Tucker (KKT)* Bedingungen genannt.
- Für einen Beweis und Regularitätsbedingungen, siehe Nocedal and Wright (2006) Section 12.4.
- Die letzte Bedingung impliziert $\lambda_i^* > 0 \Rightarrow c_i(x^*) = 0$.

Definition (Duales Problem)

Es sei

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c(x) \geq 0, \tag{86}$$

ein Optimierungsproblem ohne Gleichungsnebenbedingungen, $c(x) := (c_1(x), c_2(x), ..., c_m(x))^T$ sei die multivariate vektorwertige Funktion der Ungleichungsnebenbedingungen und die zugehörige Lagrange Funktion und der Lagrange Multiplikatoren Vektoren $\lambda \in \mathbb{R}^m$ seien durch

$$L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, (x, \lambda) \mapsto L(x, \lambda) := f(x) - \lambda^T c(x). \tag{87}$$

gegeben. Dann ist die *duale Zielfunktion* (auch *duale Lagrange Funktion genannt*) definiert als

$$q : \mathbb{R}^m \to \mathbb{R}, \lambda \mapsto q(\lambda) := \min_x L(x, \lambda), \tag{88}$$

und das *duale Problem* ist definiert als

$$\max_{\lambda \in \mathbb{R}^m} q(\lambda) \text{ u.d.N. } \lambda \geq 0. \tag{89}$$

Bemerkung

- Duale Probleme sind manchmal einfacher zu lösen als die (primären) Ausgangsprobleme.
- Duale Probleme sind für das Parameterlernen von Support Vektor Maschinen zentral.

Theorem (Schwache Dualität)

Für jede Lösung $\bar{x}$ von

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c(x) \geq 0, \tag{90}$$

und jedes $\bar{\lambda} \geq 0$ gilt, dass

$$q(\bar{\lambda}) \leq f(\bar{x}). \tag{91}$$

Beweis

Mit den Definitionen von q, $\bar{\lambda} \geq 0$, und $c(\bar{x}) \geq 0$, gilt, dass

$$q(\bar{\lambda}) = \min_x f(x) - \bar{\lambda}^T c(x) \leq f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) \leq f(\bar{x}). \tag{92}$$

□

Bemerkung

- Das Theorem besagt, dass der optimierte Wert des dualen Problems eine untere Grenze für den optimalen Wert der Zielfunktion des Ausgangsproblems ist.

Theorem (Starke Dualität)

Gegeben seien das Optimierungsproblem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ u.d.N. } c(x) \geq 0 \tag{93}$$

und seine zugehörigen notwendigen Bedingungen erster Ordnung

$$\begin{aligned} \nabla f(\bar{x}) - \nabla c(\bar{x})\bar{\lambda} &= 0, \\ c(\bar{x}) &\geq 0, \\ \bar{\lambda} &\geq 0, \\ \bar{\lambda}_i c_i(\bar{x}) &= 0, i = 1, 2, \ldots, m, \end{aligned} \tag{94}$$

mit $\nabla c(x) = (\nabla c_1(x), \nabla c_2(x), \ldots, \nabla c_m(x)) \in \mathbb{R}^{n \times m}$. $\bar{x}$ sei eine Lösung des Ausgangsproblems und f sowie $-c_i, i = 1, 2, \ldots, m$ konvexe Funktionen auf $\mathbb{R}^n$, die in $\bar{x}$ differenzierbar sind. Dann ist jedes $\bar{\lambda}$, für das $(\bar{x}, \bar{\lambda})$ die notwendigen Bedingungen des Ausgangsproblem erfüllt, eine Lösung des dualen Problems

Bemerkungen

- Die optimalen Lagrange Multiplikatoren des Ausgangsproblems sind Lösungen des dualen Problems.
- SVM Training als Quadratisches Programm benötigt das Konzept der starken Dualität.

Beweis

Wir nehmen an, dass $(\bar{x}, \bar{\lambda})$ die notwendigen Bedingungen erster Ordnung für ein Minimum des Ausgangsproblem erfüllen und dass $L(\cdot, \bar{\lambda})$ konvex und differenzierbar ist. Dann gilt für jedes $x \in \mathbb{R}^n$, dass

$$L(x, \bar{\lambda}) \geq L(\bar{x}, \bar{\lambda}) + \nabla_x L(\bar{x}, \bar{\lambda})(x - \bar{x}) = L(\bar{x}, \bar{\lambda}), \tag{95}$$

weil $\nabla_x L(\bar{x}, \bar{\lambda}) = 0$. Also gilt für die duale Zielfunktion

$$q(\bar{\lambda}) = \inf_x L(x, \bar{\lambda}) = L(\bar{x}, \bar{\lambda}). \tag{96}$$

Mit der letzten der notwendigen Bedingungen erster Ordnung folgt weiterhin

$$q(\bar{\lambda}) = L(\bar{x}, \bar{\lambda}) = f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) = f(\bar{x}) \tag{97}$$

Schließlich gilt mit dem Theorem zur Schwachen Dualität, dass $q(\lambda) \leq f(\bar{x})$ für alle $\lambda \geq 0$. Also folgt mit $q(\bar{\lambda}) = f(\bar{x})$, dass $\bar{\lambda}$ eine Lösung des dualen Problems ist. □

Prädiktive Modellierung

Differentialrechnung und Analytische Optimierung

Multivariate Differentialrechnung

Grundlagen der Optimierung

Gradientenverfahren

Grundlagen der Optimierung mit Nebenbedingungen

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie das allgemeine datenanalytische Vorgehen im Rahmen der Prädiktiven Modellierung.
2. Nennen Sie drei in der Prädiktiven Modellierung typischerweise vewendete Verfahren.
3. Erläutern Sie den Begriff der k-fachen Kreuzvalidierung.
4. Erläutern Sie Gemeinsamkeiten und Unterschiede explanatorischer und prädiktiver Modellierung.
5. Warum ist die Kenntnis von Optimierungsprinzipien im Rahmen der Prädiktiven Modellierung wichtig?
6. Definieren Sie die Begriffe der univariat-und multivariat-reellwertigen Funktion.
7. Definieren Sie Begriff der multivariaten vektorwertigen Funktion.
8. Definieren Sie den Begriff der Ableitung $f'(a)$ einer Funktion f an einer Stelle a.
9. Definieren den Begriff der Ableitung f' einer Funktion f.
10. Erläutern Sie die Symbole $f'(x)$, $\dot{f}(x)$, $\frac{df(x)}{dx}$, und $\frac{d}{dx}f(x)$.
11. Definieren Sie den Begriff der zweiten Ableitung f'' einer Funktion f.
12. Geben Sie die Summenregel für Ableitungen wieder.
13. Geben Sie die Produktregel für Ableitungen wieder.
14. Geben Sie die Quotientenregel für Ableitungen wieder.
15. Geben Sie die Kettenregel für Ableitungen wieder.

Selbstkontrollfragen

16. Bestimmen Sie die Ableitung der Funktion $f(x) := 3x^2 + \exp\left(-x^2\right) - x\ln(x)$
17. Bestimmen Sie die Ableitung der Funktion $f(x) := \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$ für $\mu \in \mathbb{R}$.
18. Definieren Sie die Begriffe des globalen und lokalen Maximums/Minimums einer Funktion.
19. Geben Sie die notwendige Bedingung für ein Extremum einer Funktion wieder.
20. Geben Sie die hinreichende Bedingung für ein lokales Extremum einer Funktion wieder.
21. Geben Sie das Standardverfahren der analytischen Optimierung wieder.
22. Bestimmen Sie einen Extremwert von $f(x) := \exp\left(-\frac{1}{2}(x-\mu)^2\right)$ für $\mu \in \mathbb{R}$.
23. Berechnen Sie die (ersten) partiellen Ableitungen der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right) \tag{98}$$

24. Berechnen Sie die zweiten partiellen Ableitungen obiger Funktion f.
25. Geben Sie den Satz von Schwarz wieder.
26. Definieren Sie den Gradienten einer multivariaten reellwertigen Funktion.
27. Geben Sie den Gradienten obiger Funktion f an und werten Sie ihn in $x = (1,2)^T$ aus.

Selbstkontrollfragen

30. Definieren Sie die Hesse-Matrix einer multivariaten reellwertigen Funktion.
31. Geben Sie die Hesse-Matrix obiger Funktion f an und werten Sie sie in $x = (1, 2)^T$ aus.
32. Definieren Sie die allgemeine Form eines Optimierungsproblems.
33. Was sind x und f eines Optimierungsproblems in der Prädiktiven Modellierung?
34. Warum betrachtet man in der Theorie der Optimierung nur die Minimierung?
35. Definieren Sie die Begriffe der globalen und lokalen Minimalstellen einer multivariaten reellwertigen Funktion.
36. Geben Sie die notwendige Bedingung erster Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
37. Geben Sie die notwendige Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
38. Geben Sie die hinreichende Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
39. Geben Sie das Theorem zu Minimalstellen konvexer Funktionen an.
40. Formulieren Sie den Algorithmus des Gradientenverfahrens.
41. Geben Sie das Theorem zum Gradientenverfahren wieder.
42. Erläutern Sie die Bedeutung der Lernrate $\alpha > 0$ und des Konvergenzkriteriums $\delta > 0$ im Gradientenverfahren.

References

Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris. 2018. "Machine Learning Approaches for Clinical Psychology and Psychiatry." *Annual Review of Clinical Psychology* 14 (1): 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research. New York: Springer.

Ostwald, Dirk, and Ludger Starke. 2016. "Probabilistic Delay Differential Equation Modeling of Event-Related Potentials." *NeuroImage* 136 (August): 227–57. https://doi.org/10.1016/j.neuroimage.2016.04.025.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3). https://doi.org/10.1214/10-STS330.

Vapnik, Vladimir. 2010. *The Nature of Statistical Learning Theory*. 2., nd ed. Softcover version of original hardcover edition 2000. Information Science and Statistics. New York, NY: Springer New York.