ALG5I - Übung 01

Alen Kocaj

26. Oktober 2017

1 Bayesian Decision Theory

a.) Erklären Sie in eigenen Worten sowie unter Verwendung von mathematischen Definitionen die grundlegenden Aspekte der Entscheidungstheorie nach Bayes

Die Bayesian Decision Theory repräsentiert ein statistisches Entscheidungsverfahren, welches sich auf der Annahme stützt, dass Entscheidungsprobleme statistisch als Wahrscheinlichkeiten dargestellt werden können. Die Theorie besagt, dass unter der Verwendung von relevanten Faktoren, das Entscheidungsproblem optimal und mit minimaler Fehlerwahrscheinlichkeit gelöst werden kann. Sie wird auch als Basis für simple Klassifizierer genommen.

Die Theorie basiert auf dem Satz von Bayes zur Berechnung von bedingten Wahrscheinlichkeiten wie z.b $P(\omega|x)$ - wie wahrscheinlich ist das Eintreten oder das Sein von ω gegeben x.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} \qquad \text{Satz von Bayes}$$

Die zwei wichtigen Faktoren, auf welchen die Entscheidungstheorie nach Bayes aufbaut, sind die:

$$\textit{a priori probability} = P(\omega) \qquad \text{und die}$$
$$\textit{class-conditional probability density oder likelihood} = p(x \mid \omega)$$

Dabei wird ω als *state of nature* − als eine Klasse oder Kategorie bezeichnet, zu der man klassifizieren kann. Würde man das Wetter klassifizieren wollen, wären z.b

$$\omega_1 = \text{sonnig}$$
$$\omega_2 = \text{bewölkt}$$
$$\omega_3 = \text{regnerisch}$$

Hingegen beschreibt x ein Feature, welches wir in Betracht nehmen, um unseren Klassifizierer zu verbessern. In Bezug zu dem Wetterbeispiel von vorhin wäre x z.b die Luftfeuchtigkeit. Nachdem jeder Tag eine andere Luftfeuchtigkeit besitzt, wird x statistisch als kontinuierliche Zufallvariable dargestellt, die eine bestimmte Verteilung hat. Welche Art der Verteilung hängt von der der jeweiligen Klasse ω ab = $p(x|\omega)$.

Zusammengefasst und unter der Verwendung des Satzes von Bayes lässt sich die Entscheidungstheorie darstellen als:

$$P(\omega \mid x) = \frac{P(\omega)P(x \mid \omega)}{P(x)} \tag{1}$$

wo *P(ω|x)* die *a posteriori probability* − also die resultierende Wahrscheinlichkeit der Klasse ω gegeben Feature x darstellt. Als Parallele zu unserem Wetterbeispiel bezeichnet $P(\text{bewölkt}|0.67)$ die *a posteriori* Wahrscheinlichkeit, dass der Tag bewölkt ist, gegeben einer Luftfeuchtigkeit von 67%.

Manch einer wird sich wundern über die Verwendung von *P(x)*, obwohl zuvor von nur zwei entscheidenden Faktoren gesprochen wurde. Tatsächlich stellt *P(x)* einen sogenannten Evidenzfaktor da, der lediglich dafür sorgt, dass die *a posteriori probabilities* aller Klassen ω_1 ... ω_n − also *P(*ω_1*| x)* ... *P(*ω_n*| x)* − unter Betrachtung des Features x summiert 1 ergibt, so wie alle guten Wahrscheinlichkeiten es tun sollten:

$$p(x) = \sum_{i=1}^{n} P(\omega_j)P(x \mid \omega_j)$$

Dieser Faktor kann in Entscheidungsfragen weggelassen werden. Dadurch ergibt sich die Entschiedungstheorie:

$$\text{Wähle } \omega_1 \text{ wenn } P(\omega_1)P(x \mid \omega_1) > P(\omega_2)P(x \mid \omega_2); \qquad \text{ansonsten wähle } \omega_2 \tag{2}$$

Weiters kann man unter Verwendung von englischen Ausdrücken die Entschiedungstheorie nach Bayes auch darstellen als:

$$\text{a posteriori} = \frac{\text{a priori x likelihood}}{\text{evidence}} \tag{3}$$

Die *a priori probability P(ω)* bezeichnet das Wissen, welches wir im Vorhinein über das Eintreten einer bestimmten Kategorie oder Klasse ω. In unserem Beispiel zum Wetter könnte man z.b aus den Daten entnehmen, dass *P(*ω_1*)* = 0.7, also die Wahrscheinlichkeit, dass der Tag sonnig ist. Das heißt für zukünftige Klassifizierungen würden wir uns zum Einen auf diesen Wahrscheinlichkeitswert stützen. Übrigens, es ist genauso legitim, die *a priori probability* zu schätzen bzw. anzunehmen. Es gibt Fälle, bei denen wir uns theoretisch nur auf die *a priori probability* verlassen könnten, ohne andere Wahrscheinlichkeiten zu betrachten. Dies passiert, wenn wir, gegeben den Daten, die Klassifizierung nur mit unserem Vorwissen vornehmen können. D.h wir würden ohne Betrachtung von Features oder Eigenschaften den Tag nach der höchsten *a priori* Wahrscheinlichkeit *P(ω)* einstufen. Als Beispiel würde unsere Klassifikiation immer die Klasse *sonnig* bevorzugen, wenn $\text{P}(\omega_1 = \text{sonnig}) = 0.7 > \text{P}(\omega_2 = \text{bewölkt}) = 0.2 > \text{P}(\omega_3 = \text{regnerisch}) = 0.1$. Natürlichweise fällt die Kurzsichtigkeit dieser Methode sofort auf. Wir würden jedes Mal den Tag als sonnig klassifizieren, selbst der zu betrachtende Tag ein regnerischer Tag wäre.

Möchte man ein Feature in die Entschiedungshilfe miteinfliesen lassen, so betrachtet man alle Samples dieses Features basierend auf den zu klassifizierenden Kategorien. Dies ergibt eine Verteilung des Features pro Klasse. Daher wird diese Wahrscheinlichkeit auch als *class-conditional probability density* function *p(x|ω)* bezeichnet, welche die *likelihood* von ω in Respekt zu x darstellt. Der Term *likelihood* lässt sich ausdrücken als "Unter der Prämisse, alle anderen Wahrscheinlichkeiten wären gleich, wie richtig ist die Kategorie ω für die Verteilung der Zufallsvariable x *p(x|ω)* geeignet?". Die Verteilung der Zufallsvariable wird entweder aus den Daten entnommen oder, wie oft in der Praxis geschätzt. Aus unserem Beispiel von vorhin wäre dies die Luftfeuchtigkeit.

Abschließend muss noch erwähnt werden, dass die Entscheidungstheorie nach Bayes, so wie zuvor simpel beschrieben, lediglich die mathematische Definitionen für ein Feature x beschreibt. Zur Verwendung von mehreren Features benötigt es weitere Definitionen, welche hier nicht behandelt werden.

b.) Zeigen Sie anhand eines selbst entworfenen Beispiels den Einsatz der Entscheidungstheorie nach Bayes

Man nehme an Sportler unterziehen sich einen Drogentest um herauszufinden, ob diese Dopingsubstanzen genommen haben oder nicht. Dabei betrachtet man für jeden Sportler das Resultat seines/ihres Drogentests, dessen Eigenschaft sich als diskrete Zufallsvariable x darstellen lässt (positiv + oder negativ −). Man möchte nun basierend auf dem Drogentest herausfinden, ob ein beliebiger Sportler Dopingsubstanzen konsumiert oder nicht gegeben dass sein Drogentest positiv ausgefallen ist. Sportler lassen sich in zwei Kategorien eingliedern: In jene, welche Dopingsubstanzen konsumieren und jene, die es nicht tun. Aus den Daten erschließen wir, dass 0.5% der Sportler Dopingmittel konsumieren, während 99.5% es nicht tun. D.h unsere *a priori* Wahrscheinlichkeiten sind:

$$P(\omega_1 = \text{Konsumiert Dopingmittel}) = 0.5\% = 0.005$$
$$P(\omega_2 = \text{Konsumiert keine Dopingmittel}) = 99.5\% = 0.995$$

Weiters wissen wir, dass diese Drogentests sehr zuverlässig sind, was ihre Vorhersage angeht. So sagen sie bei Sportler, welche Dopingmittel konsumieren zu 99% vorraus, dass diese positiv sind und zu 1%, dass diese negativ ausfallen. Ebenso bei nicht dopenden Sportler. Da liegt die Wahrscheinlichkeit der *likelihood*, dass der Test negativ ausfällt bei 99% und dass er positiv ausfällt bei 1%. Also eine ziemlich gute True Positiv Rate und eine True Negativ Rate.

$$P(x = positiv \mid \text{Konsumiert Drogenmittel}) = 99\% = 0.99$$
$$P(x = positiv \mid \text{Konsumiert keine Drogenmittel}) = 1\% = 0.01$$

und

$$P(x = negativ \mid \text{Konsumiert keine Drogenmittel}) = 99\% = 0.99$$
$$P(x = negativ \mid \text{Konsumiert Drogenmittel}) = 1\% = 0.01$$

Der Evidenzfaktor − also Wahrscheinlichkeit, dass der Test des/der Sportlers/Sportlerin positiv ausfallen kann, lässt sich akkumuliert ausrechnen.

$$P(\text{positiv}) = P(\text{positiv} \mid \omega_1)P(\omega_1) + P(\text{positiv} \mid \omega_2)P(\omega_2)$$
$$P(\text{positiv}) = 0.99 * 0.005 + 0.01 * 0.995 = 0.0150 = 1.5\%$$

Lassen wir uns nun die *a posteriori* Wahrscheinlichkeiten ausrechnen, wie wahrscheinlich ein gegebener beliebiger Sportler Dopingmittel konsumiert gegeben, dass sein Drogentest positiv ausfiel.

$$P(\omega_1 \mid \text{positiv}) = \frac{P(\text{positiv} \mid \omega_1)P(\omega_1)}{P(\text{positiv})} = \frac{0.99 * 0.005}{0.015} = 0.33 = 33\%$$

Das Ergebnis überrascht, da ja der Drogentest bei Spotlern, welche Dopingmittel konsumieren, die Positivität zu 99% vorhersagt. Diese Annahme täuscht jedoch. Ausschlagebend ist die Anzahl der tatsächlichen Sportler/Sportlerinnen, welche Dopingmittel konsumieren in unserem Datenset. Da die *a priori* Wahrscheinlichkeit sehr gering zu sein scheint mit $P(\omega_1)$ = 0.5%, überwiegt das Gewicht der Falschpositiv klassifizierten Sportler aus der Menge an nicht doping-konsumierenden Sportler.

Als Gegenprobe zum Beweis von guten Wahrscheinlichkeiten berechnen wir uns nun die Wahrscheinlichkeit, dass ein Sportler keine Dopingmittel konsumiert gegeben, dass sein Drogentest positiv ausfällt. Diese Wahrscheinlichkeit sollte diegleiche sein, wie 1 − P(ω_1 | positiv) = 67% unter der Annahme, dass P(ω_1 | positiv) + P(ω_2 | positiv) = 1.

$$P(\omega_2 \mid \text{positiv}) = \frac{P(\text{positiv} \mid \omega_2)P(\omega_2)}{P(\text{positiv})} = \frac{0.01 * 0.995}{0.015} = 0.6633 \sim= 67\%$$

Da die Wahrscheinlichkeit, dass ein beliebiger Sportler unter der Prämisse, sein Drogentest sei positiv, dennoch keine Dopingmittel verwendet höher ist, als dass er welche verwendet, wird unser Klassifizierer den Sportler dementsprechend auch so kategorisieren.

2 Maximum Likelihood

a.) Erklären Sie in eigenen Worten sowie unter Verwendung von mathematischen Definitionen die grundlegenden Aspekte der Maximum Likelihood Methode zur Schätzung von Parametern

Wie zuvor in der Entscheidungstheorie von Bayes beschrieben, kann man einen optimalen Klassifizierer erstellen, dessen Fehlerrate minimal ist unter Verwendung von der *a priori probability*: *P(ω)* und der *likelihood* einer Klasse ω zu einem Feature x, dargestellt durch die *class-condition probability density* function: *p(x|ω)*.

Leider haben wir in der Realität nicht immer alles Informationen zur Verfügung. Meist stehen uns nur eine limitierte Anzahl an Trainingsdaten bestehend aus Samples zur Verfügung, aus denen wir die Information annährend müssen, um unseren Klassifizierer zu bauen. Unter Betrachtung von Überwachtem Lernen, wo die Daten sich durch Klassen markiert und bezeichnet sind, scheint die Erhebung der *a priori probability* kein sonderlich großes Problem darzustellen (meist nimmt man die prozentuelle Verteilung der jeweiligen Klasse ω über alle Datensamples). Schwieriges gestaltet sich die **Schätzung** der *class-condition probability density* function. Besondern dann, wenn die Anzahl der Samples knapp ist, während der Featurevektor zu groß ist (n << d; n = Anzahl der Sampels, d = Anzahl der Features).

Eine Abhilfe kann man sich schaffen, in dem man nicht die Funktion *p(x|ω)*, sondern deren Parameter schätzt. Wenn man sich *p(x|ω)* als normalverteilte Funktion *p(x|ω)* $\sim$ *N(μ, σ)* vorstellt, braucht man nur mehr die Parameter μ und σ schätzen. Die Parameter μ und σ können gesammelt als Parametervektor $\theta := [\mu, \sigma]$ bezeichnet werden. So lässt sich das Problem der Parameterschätzung mathematisch definieren als: Für alle erhobenen Samples, finde einen geeigneten Parametervektor θ für die Wahrscheinlichkeitsdichte *p(x|ω, θ)* oder kurz *p(x|θ)*, welcher bestmöglichst (most likely) zu unseren Samples und rückschließend zu unseren Realdaten passt bzw. dessen Verteilung am Besten darstellt. Hier bezeichnet x ein Sample des Features x. Das heißt man geht alle möglichen Erklärungsmodelle / Verteilungen durch, schätzt den jeweiligen Parametervektor θ (z.b.: für X $\sim$ N = $[\mu, \sigma]$; für X $\sim$ Poisson = $[\lambda]$) und sucht sich den Paramatervektor zur der Verteilung heraus, die zu unseren Daten am Besten passt. Man sucht mit *p(x|θ)* die Wahrscheinlichkeit für die Daten, gegeben das Modell, dargestellt durch den Parametervektor θ. Nachdem meist mehrere Samples unabhängig erhoben werden, die jeweils ihrer eigene Verteilung darstellen, nutzen wir diese um für das gesamte Sammlung an Samples θ zu schätzen. Die resultierende *likelihood* Funktion lässt darstellen als:

$$L(\theta) = \prod_{k=1}^{N} p(x_k \mid \theta) \tag{4}$$

Hier bezeichnet *L(θ)* die *likelihood* von θ unter Berücksichtigung (oder in Respekt) zu allen Samples. Zur Maximierung des *likelihood* schreibt man jenes $\hat{\theta}$ als das *maximum likelihood estimate* von θ, welches *L(θ)* maximiert. Das ist auch genau jener Parameterverktor des jeweiligen Erklärungsmodells / Verteilung, welcher - wie im oberen Absatz beschrieben - am Besten zu unseren Daten passt.

Nachdem Wahrscheinlichkeiten immer einen Kommawert unter Eins darstellen, führt die Multiplikation von vielen Wahrscheinlichkeiten zu sehr geringen Kommawerten. Zum Beispiel würden zehn Wahrscheinlichkeiten mit jeweils einem Wert von 0.1 multipliziert $1 * 10^{-10}$ ergeben. Zu geringe Werte um vernünftig mit ihnen zu arbeiten. Zudem würde heutige Computer durch so geringe Floatwerte an ihre rechnerischen Grenzen getrieben werden. Helfen wird uns hier die Logarithmusfunktion, die ihrerseits eine kontinuierliche steig-steigende Funktion darstellt, welche die größenverhältnisse unserer Wahrscheinlichkeiten nicht verändert. Wir können also für alle Werte der Likelihood Funktion ebenso den

Logarithmus verwenden, daher spricht man von der *log-likelihood* von θ:

$$l(\theta) = \ln\, L(\theta)$$
$$l(\theta) = \ln \prod_{k=1}^{N} p(x_k \mid \theta)$$

welches unter Verwendung der Produktregel des Logarithmus:

$$l(\theta) = \ln \sum_{k=1}^{N} p(x_k \mid \theta)$$
$$l(\theta) = \sum_{k=1}^{N} \ln\, p(x_k \mid \theta)$$

ergibt. Würde man z.b.: die Parameter für eine eindimensionale Zufallsvariable $x \sim N(\mu, \sigma)$ schätzen, ergibt das

$$l(\theta) = \sum_{k=1}^{N} \ln\, \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}$$

mit $\theta := [\mu, \sigma]$. Um nun die eigentlichen Parameter μ und σ zu schätzen, bedarf es zwei weiterer Schritte:

1. Leite $l(\theta)$ nach den jeweiligen Parametern aus θ ab
2. Setze die Ableitung auf 0 und löse die Gleichung nach den jeweiligen Parametern

Die ergibt auch durchaus Sinn. Wenn man generell den Maximalwert eine beliebigen Funktion bestimmen möchte, leitet man diese ab und setz die Gleichung auf 0.

b.) Zeigen Sie unter Verwendung der Maximum Likelihood Methode, warum es eine gute Wahl zu sein scheint, bei der Schätzung der Parameter einer Normalverteilung folgendermaßen vorzugehen

$$\hat{\mu} = \frac{1}{N} \sum_{1}^{j} x_j \quad \text{und} \quad \hat{\sigma} = \sqrt{\frac{\sum_1^j (x_j - \mu)^2}{N}}$$

Zunächst nehmen wir als Basis die Log-likelihood Funktion unter Verwendnung der Normalverteilung als unser Erklärungsmodell für unsere Daten.

$$l(\theta) = \sum_{k=1}^{N} \ln\, \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right) \qquad e^x \equiv \exp x$$

Nun nehmen wir die Eigenschaft des Logarithmus für Multiplikationen und teilen die beiden Ausdrücke.

$$l(\theta) = \sum_{k=1}^{N} \ln\, \frac{1}{\sqrt{2\pi\sigma^2}} + \ln\, \exp\left(-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}\right) \qquad \ln(\exp) = 1$$
$$l(\theta) = \sum_{k=1}^{N} \ln\, \frac{1}{\sigma\sqrt{2\pi}} + -\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}$$

Nun heben wir den Ausdruck heraus, in dem das durch die Summe definierte k nicht vorkommt. Nachdem dieser Ausdruck aber dennoch N-mal vorkommt, multiplizieren wir ihn mit N. Gleichzeitig verschieben wir das $-\frac{1}{2}$, damit es besser positioniert ist.

$$l(\theta) = \mathrm{N} \ln \frac{1}{\sigma\sqrt{2\pi}} - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

Der Bruch mit $\frac{1}{\sigma\sqrt{2\pi}}$ kann in zwei seperate Brüche geteilt werden.

$$l(\theta) = \mathrm{N} \ln \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2} \qquad \ln(a * b) = \ln a + \ln b$$

$$l(\theta) = \mathrm{N} \ln \frac{1}{\sigma} + \ln \frac{1}{\sqrt{2\pi}} - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2} \qquad \ln \frac{1}{a} = -\ln a$$

$$l(\theta) = \mathrm{N} \left(-\ln \sigma - \ln \sqrt{2\pi}\right) - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

Nun haben wir unsere Ausgangsformel von der wir nach μ und σ ableiten.

Ableitung nach μ

Möchte man die Ableitung von der Funktion *l(θ)* notieren, nutzt man den Differentialoperator ∂.

$$\frac{\partial l(\theta)}{\partial \mu} = \mathrm{N} \left(-\ln \sigma - \ln \sqrt{2\pi}\right) - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

bedeutet "Leite $l(\theta)$ nach μ ab". Nachdem der erste Term in der Gleichung nach μ nicht abgeleitet werden kann - da er eine Konstante darstellt - fällt dieser weg.

$$\frac{\partial l(\theta)}{\partial \mu} = -\sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2} \qquad \text{Ziehe } \frac{1}{2\sigma} \text{ heraus}$$

$$\frac{\partial l(\theta)}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{k=1}^{N} (x_k - \mu)^2 \qquad \text{Leite nach } \mu \text{ ab}$$

$$\frac{\partial l(\theta)}{\partial \mu} = -\frac{1}{2\sigma^2}\, 2 \sum_{k=1}^{N} x_k - \mu \qquad \text{Innere Ableitung von } x_k - \mu$$

$$\frac{\partial l(\theta)}{\partial \mu} = +\frac{1}{2\sigma^2}\, 2 \sum_{k=1}^{N} x_k - \mu \qquad \text{Kürze 2}$$

$$\frac{\partial l(\theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{k=1}^{N} x_k - \mu$$

Maximieren von μ

Die Maximierung von μ zur Findung von $\hat{\mu}$ geschieht durch Gleichsetzung mit 0.

$$\frac{1}{\sigma^2} \sum_{k=1}^{N} x_k - \mu = 0 \qquad \text{Dividiere } \frac{1}{\sigma^2}$$

$$\sum_{k=1}^{N} x_k - \mu = 0 \qquad \text{Erweitere Summe}$$

$$\sum_{k=1}^{N} x_k - \sum_{k=1}^{N} \mu = 0 \qquad \text{Bringe auf andere Seite}$$

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \mu \qquad \text{Löse Summe, da } \mu \text{ konstant N mal vorkommt}$$

$$\sum_{k=1}^{N} x_k = N\mu \qquad \text{Dividiere durch N}$$

$$\hat{\mu} = \frac{\sum_{k=1}^{N} x_k}{N} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

Ableitung nach σ

$$\frac{\partial l(\theta)}{\partial \sigma} = \mathrm{N}\ (-\ln\ \sigma - \ln\ \sqrt{2\pi}) - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

bedeutet "Leite $l(\theta)$ nach σ ab". Zunächst fällt einmal $-\ln\ \sqrt{2\pi}$ - weil Konstante - weg.

$$\frac{\partial l(\theta)}{\partial \sigma} = -\mathrm{N}\ \ln\ \sigma - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2} \qquad \text{Leite ln } \sigma \text{ ab und ziehe } \frac{1}{2\sigma} \text{ heraus}$$

$$\frac{\partial l(\theta)}{\partial \sigma} = -\mathrm{N}\ \frac{1}{\sigma} - \frac{1}{2\sigma^2} \sum_{k=1}^{N} (x_k - \mu)^2 \qquad \text{Schreibe } \frac{1}{\sigma^2} \text{ als } \sigma^{-2}$$

$$\frac{\partial l(\theta)}{\partial \sigma} = -\mathrm{N}\ \frac{1}{\sigma} - \frac{1}{2}\ \sigma^{-2} \sum_{k=1}^{N} (x_k - \mu)^2 \qquad \text{Leite } \sigma^{-2} \text{ ab und kürze -2}$$

$$\frac{\partial l(\theta)}{\partial \sigma} = -\mathrm{N}\ \frac{1}{\sigma} + \sigma^{-3} \sum_{k=1}^{N} (x_k - \mu)^2 \qquad \text{Forme } \sigma^{-3} \text{ um}$$

$$\frac{\partial l(\theta)}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{k=1}^{N} (x_k - \mu)^2$$

Maximieren von σ

Die Maximierung von σ zur Findung von $\hat{\sigma}$ geschieht durch Gleichsetzung mit 0.

$$-\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{k=1}^{N} (x_k - \mu)^2 = 0 \qquad \text{Bringe } -\frac{N}{\sigma} \text{ auf die andere Seite}$$

$$\frac{1}{\sigma^3} \sum_{k=1}^{N} (x_k - \mu)^2 = \frac{N}{\sigma} \qquad \text{Multipliziere mit } \sigma^3$$

$$\sum_{k=1}^{N} (x_k - \mu)^2 = \frac{N\sigma^3}{\sigma} \qquad \text{Kürze } \sigma$$

$$\sum_{k=1}^{N} (x_k - \mu)^2 = N\sigma^2 \qquad \text{Dividiere durch N und ziehe die Wurzel}$$

$$\sigma = \sqrt{\frac{\sum_{k=1}^{N} (x_k - \mu)^2}{N}}$$