Maschinelles Lernen und Datenanalyse

In der Mess- und Prüftechnik

PD Stefan Bosse

Universität Bremen - FB Mathematik und Informatik

Daten und Sensoren

- Metriken von Daten
- Metriken von Aussagen
- Sensoren als Datenquellen
- Messverfahren und Sensorsysteme

Daten

- Daten sind die Grundlage für die Modellbildung und Modelltestung
- Daten können aus einer Vielzahl von Quellen stammen
  - Experiment
  - Simulation
  - Feldstudie
  - Abgeleitet aus anderen Datensätzen:
    MapAndReduce(D): $D \rightarrow D'$

Daten

- Allgemein kann man Daten und deren Werte unterteilen in:
  - Skalare Werte, wie Temperatur, Alter, usw.
  - Serien von Skalaren Werten, wie Zeitserien
  - Vektorielle Werte wie Bilder
  - Zusammengesetzte Daten, also Datenstrukturen (Records)
- Daten haben daher eine Dimensionalität $\mathbb{X}^N$, wobei die Wertemenge $\mathbb{X}$ einer Dimension aus den ganzen $\mathbb{N}$, reelen $\mathbb{R}$, der Zeit $\mathbb{T}$ oder kategorischen Wertemengen $\mathbb{S}$ bestehen kann (oder Untermengen davon).

Datenreduktion

- Ziel der Datenanalyse ist die Reduktion von Eingabedaten bezüglich Größe und Dimensionalität:

$$P(X^N) : X^N \rightarrow Y^M$$
$$|Y| < |X|, M < N$$

- Materialwissenschaften und Messtechnik:
  - Häufig metrische Eingabevariablen
  - Häufig metrische oder kategorische Ausgabevariablen (inkl. Boolescher Variablen)

```
function isRaining(temp,sunrad,moisture) =
  temp < 0 ? → false
  temp > 40 ? → false
  (sunrad-moisture) > 30? → false
  true
```

Bsp. 1. Beispiel aus der Messtechnik mit einer Datenreduktionsfunktion $\mathbb{R}^3 \rightarrow \mathbb{B}$

Datenreduktion

- Sozioalwissenschaften:
  - Häufig kategorische und metrische Eingabevariablen
  - Häufig kategorische Ausgabevariablen (inkl. Boolescher Variablen)

```
function isStrong(age,weight,length) =
  age < 10 ? → false
  weight > 200 ? → false
  (weight/length) > 30? → false
  true
```

Bsp. 2. Beispiel einer Datenreduktionsfunktion $\mathbb{R}^3 \rightarrow \mathbb{B}$

Datenklassen

Numerische und Metrische Werte

Das sind Werte die abzählbar sind und wo man Relationen (wie kleiner oder größer) sinnvoll definieren kann, also alle reellen und ganzen Zahlen.

- Beispiele: Temperatur, Länge, Dichte, Porengröße, Dehnung, Kraft, Ort, Zeit

Kategorische Werte

Das sind symbolische Werte für die entweder keine (sinnvolle) Ordnungsrelation existiert oder wo sich wenigstens keine Differenzen bilden lassen.

- Beispiele: Staatsangehörigkeit, Farbennamen (rot < gelb???), Schadenstyp, charakteristisches Merkmal (Anomalie?)

Skalierung der numerischen Werte

Intervallskaliert

Für diese Art von Attributen sind nur Unterschiede (Addition oder Subtraktion) sinnvoll. Beispielsweise wird die in °C oder °F gemessene Temperatur intervallskaliert. Wenn es 20 °C an einem Tag und 10 °C am folgenden Tag ist, ist es sinnvoll, über einen Temperaturabfall von 10 °C zu sprechen, aber es ist nicht sinnvoll zu sagen, dass es doppelt so kalt ist wie am Vortag.

Verhältnisskaliert

Hier kann man sowohl Differenzen als auch Verhältnisse zwischen Werten berechnen. Zum Beispiel kann man für das Alter sagen, dass jemand, der 20 Jahre alt ist, doppelt so alt ist wie jemand, der 10 Jahre alt ist.

Ordnungsrelationen

Nominal

Die Attributwerte in der Domäne sind ungeordnet und somit nur Gleichheitsvergleiche sinnvoll. Das heißt, wir können nur überprüfen, ob der Wert des Attributs für zwei bestimmte Instanzen gleich ist oder nicht. Zum Beispiel ist Geschlecht ein nominales Attribut.

Ordinal

Die Attributwerte sind geordnet und somit Gleichheitsvergleiche (ist ein Wert gleich einem anderen?) und relationale Vergleiche (ist ein Wert kleiner oder größer als ein anderer?) sind erlaubt, obwohl es möglicherweise nicht möglich ist, die Differenz zwischen den Werten zu quantifizieren!

Datenklassen (longitudinal)

- Sensor- und Messdatenvariablen (sowohl kategorisch wie auch metrisch) können weiter unterschieden werden in:

Statisch

Die Variable s ist zeitlich nicht veränderlich bzw. ist in einem wesentlichen Zeitintervall $t \in [t_0, t_1]$ als stationär (unveränderlich) anzusehen.

Dynamisch

Zeitlich veränderliche Variable $s(t)$ ist zeitabhängig und bildet eine Datenserie (oder Zeitvektor) $s(t)=\{s_0,s_1,..s_t\}$ bei disketer Erfassung, d.h., wir sprechen von longitudinalen Daten.

Ein Sensorsignal ist zeitlich immer diskret, aber die physikalische Variable die der Sensor misst ist zeitlich kontinuierlich (Sampingtheorem beachten)

Daten

Datensätze als Matrizen

- Ein Menge von Daten kann in **Matrizenform** als Matrix ***D*** dargestellt werden (Analogie zur Tabellenform) [1]:

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- Der Vektor ***X*** ist die Menge aller Variablen (Sensoren) und die Spalten der Matrix ***D***:

$$\vec{X} = (x_1, x_2, .., x_d)$$

- Jede Zeile $\boldsymbol{x}_j$ ist ein Rekord der Variablenmenge $\{X_i|i=1,d\}$ mit konkreten Werten und geben als d-stelliges Tupel je nach Anwendung und Zielsetzung einzelne Beispiele, Instanzen, Experimente, Entitäten, Objekte, und Eigenschaftsvektoren wieder:

$$\vec{d}_j = \vec{x}_j = (x_{j,1}, x_{j,2}, .., x_{j,d})$$

```
// JavaScript
type row = { x1:number|string, x2:number|string, ..,
             xd:number|string }
type table = row array;
```

Eingabe- und Ausgabevariablen

- Die Variablenmenge setzt sich aus Ein- und Ausgabevariablen zusammen
- Sensoren sind typischerweise Eingabevariablen *x*
- Aussagen sind Ausgabevariablen *y*, also Ergebnisse die sich aus den Eingangsvariablen ableiten lassen können (durch eine Funktion *F*):

$$\vec{X}_{xy} = (X_1, X_2, .., X_u, Y_1, Y_2, .., Y_v)$$
$$\vec{X} = (X_1, X_2, .., X_u)$$
$$\vec{Y} = (Y_1, Y_2, .., Y_v)$$
$$\vec{d}_j = (x_{j,1}, x_{j,2}, .., x_{j,u}, y_{j,1}, y_{j,2}, .., y_{j,v})$$
$$F(\vec{X}) : \vec{X} \rightarrow \vec{Y},$$

mit $u+v=d$.

Merkmale und Eigenschaften (Features)

Wir unterscheiden zwei Arten von Merkmalen:

Eingabemerkmale Ft_{i}

Das sind Merkmale der Eingabedaten $\boldsymbol{x}$. Das können z.B. statistische Eigenschaften wie der Mittelwert oder Frequenzen eines Zeitsignals sein. Die Merkmale sollen möglichst die Zielmerkmale verstärken, also eine signfikante (wenn auch noch nicht bekannte) Abhängigkeit $\boldsymbol{Ft}_{\mathrm{o}}(\boldsymbol{Ft}_{\mathrm{i}})$ besitzen

Ziel- und Ausgabemerkamel Ft_{o}

Das sind die Ergebnisse der Datenanalyse, z.B. die Antwort auf die Frage Schaden Ja/Nein?, oder eine Schadensposition, eine Überlebenswahrscheinlichkeit. Die Eingabemerkmale sind die starken Variablen für die Modellfunktion M, die wir suchen.

- Die Merkmalsselektion ist also die Vorstufe und Datenvorverarbeitung, selten werd mit Rohdaten direkt gearbeitet
- Es muss eine Merkmalsselektionsfunktion *MF* geben, die automatisch die Merkmale aus den den Daten ableitet:

$$M(\vec{x}) : \vec{x} \rightarrow \vec{y} \Rightarrow$$
$$Ft_o \Leftrightarrow y$$
$$MF(\vec{x}) : \vec{x} \rightarrow \vec{F}t_i$$
$$M_t\left(\vec{F}t_i\right) : \vec{F}t_i \rightarrow \vec{F}t_o$$

Beispiel einer Datenmatrix

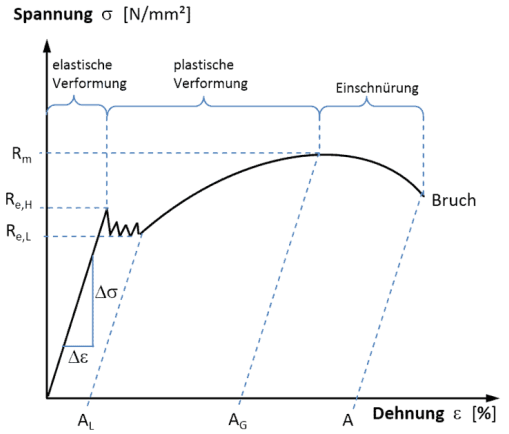
- Botanischer Datensatz mit geometrischen (numerischen) Eigenschaften einer Pflanze und kategorischer Klassifikation:

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

[1]

- Messdatensatz

Berechnetes Dehnungs-Spannungsdiagramm



[www.precifast.de/elastizitaetsmodul-e-modul]

Messdaten aus Dehnversuch

Dehnung [mm]	Kraft [kN]
0	0
0.1	0.2
0.2	0.7
0.3	1.5
0.4	1.7
0.5	1.9
0.6	2.0
0.7	0.2
0.8	-0.5

Attribute

- Die gemessenen Variablen X_1 bis X_4 sind metrische Datenvariablen, die Variable $X_5{=}y$ ist eine kategorische Variable!
- Die gemessenen Variablen X_1 bis X_4 (also Sensoren) nennt man **Attribute**, da sie Eigenschaften und beschreibende Variablen der Zielvariablen y sind

Sensoren

Welche Sensoren und Messdaten kennt ihr:

Sensoren

- Messtechnik
  - Physikalische Größen wie Temperatur, Dehnung, Spannung, Zeit
  - Fusionierte Umfragevariablen (z.B. Ensemblemittelwerte)
- Bei der Messung mit Sensoren unterscheidet man:
  - Einmalige bzw. einzelne Messungen (single shot)
  - Wiederholte Messungen der gleichen physikalischen Größe (Mittelwertbildung..)
  - Serien von Messwerten, vor allem zeitaufgelöste Datenserien: $\mathbf{D} = \{d_1, d_2, .., d_n\}$, wobei i.A. $\Delta t(d_i, d_{i+1}) = \text{constant}$

Sensoren

- Soziotechnische Systeme, Umfragen
  - Umfragevariablen (Antworten auf Fragen) sind Sensoren von einzelnen Menschen
  - Fusionierte Umfragevariablen (z.B. Ensemblemittelwerte) sind Sensoren von Menschengruppen
- Allgemein verfügbare Daten
  - Soziale Netzwerke und soziale Medien
  - Datenbanken von Behörden usw.

Sensormodell

- Ein Sensor ist ein Messwandler, auch in der Soziologie (Indikator für eine Eigenschaft die nicht direkt messbar ist)
- Ein Sensor bildet daher eine i.A. physikalische Größe *x* auf eine andere Größe *y* ab:

$$S(x) : x \rightarrow y, K : correct(x \rightarrow y)$$

- Es gibt i.A. eine Kalibrierungsfunktion $K(f, x, y)$
- Beispiele:
  - Druck → Spannung, Strahlung → Strom, usw.
  - Soziale Vernetzung → Numerischer Radiuswert, Wählerstimmen → Politik, d.h., **Zuordnung von Zahlen zu Objekten oder Ereignissen nach festgelegten Regeln**

Sensordaten

- Sensoren *S* sind Datenquellen *d* von physikalischen, soziologischen oder sonstigen natürlichen nicht direkt erfassbaren Größen *x*
- Die Datenwerte (numerisch) werden in einem definierbaren Intervall liegen
  - Die Kenntnis des Werteintervalls ist wichtig für spätere Datenverarbeitung, Analyse, und Maschinelles Lernen!
  - Kategorische Werte werden ebenfalls durch eine Menge definiert

$$S(x) : x \rightarrow d$$
$$d \in [a,b] \Rightarrow \{v_0, v_1, .., v_i\}$$

Mess- und Sensorische Systeme

Der Ursprung der Daten für Analyse und Maschinelles Lernen!

Ein Sensor kommt selten allein.

Messverfahren

Man unterscheidet zwei verschiedene Messverfahren:

Passives Messverfahren

Die sensorischen Werte sind Ergebnis einer intrinsischen Eigenschaft (Dichte) oder bereits existierender externer Größen (Temperatur). Der Stimulus der Messung ist das Bauteil, der Mensch, die Umwelt.

Aktive Messverfahren

Es gibt einen aktiven Stimulus dessen Antwortsignal durch den Sensor erfasst wird. Beispiel ist das Ultraschallmessverfahren mit geführten Wellen. Das Sensorsignal ist immer abhängig vom Stimulus. In der Soziologie ist der Stimulus z.B. ein Fragenkatalog in einer Umfrage, die Antworten sind die Sensorvariablen.

Sensoraggregation

Sensorklassen

Physische Sensoren

Physische Sensoren messen direkt eine Größe mit einem Messinstrument (kann auch die Auswertung einer Frage in einem Fragebogen sein), Smartphone

Virtuelle Sensoren

Verwenden Daten (von physischen und anderen virtuellen Sensoren) um neue sensorische Werte zu berechnen (kein Messinstrument) → **Aggregatoren**!!

Sensoraggregation

Schichtenmodell von Sensorischen Systemen

In sensorischen Systemen werden Sensordaten in verschiedenen Ebenen verarbeitet:

- **Vertikale Ebenen** repräsentieren die sensorischen Domainen und die Sensorklassen;
- **Horizontale Ebenen** repräsentieren die Datenverarbeitung.

Sensoraggregation

Vertikale Ebenen

Perzeption
: Hier findet die Akquisition der rohen Sensordaten statt. Die Sensoren sind räumlich verteilt und werden lokal vorverarbeitet.

Aggregation
: Einzelne Sensordaten werden zeitlich und räumlich zusammengeführt und gesammelt (Sensorfusion)

Applikation
: Die gesammelten Daten werden nutzbar gemacht: Weitere Datenverarbeitung, Aufbereitung, Eigenschaftsselektion, Informationsgewinnung, Visualisierung

Sensoraggregation

Horizontale Ebenen

- Die horizontalen Ebenen durchziehen alle vertikalen Ebenen:
  1. Sicherheit
  2. Datenverarbeitung
  3. Kommunikation
  4. Datenspeicherung
  5. Nachrichtenvermittlung
  6. Management

Sensoraggregation

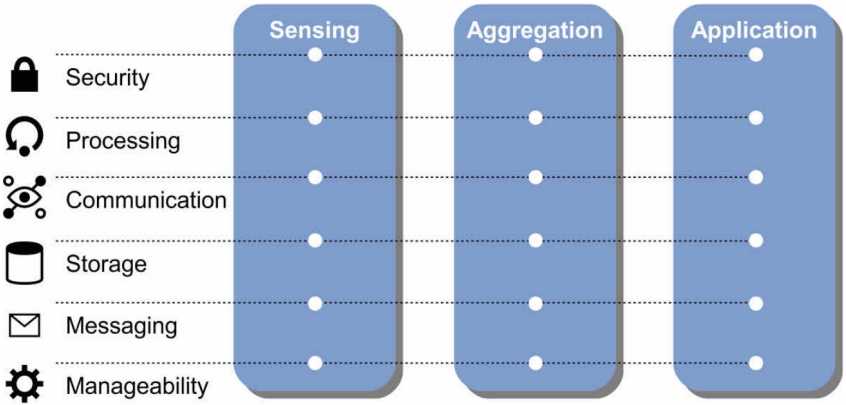


Abb. 1. Grundlegender Zusammenhang der horizontalen und vertikalen Ebenen in Sensorischen Systemen

Sensoraggregation

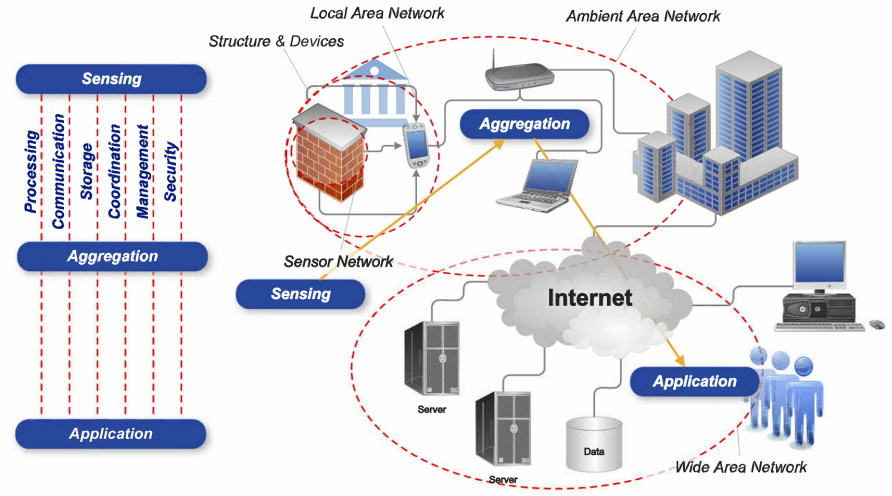


Abb. 2. Räumliche Abbildung der vertikalen Ebenen auf Cloud Computing

Sensoren in den Ebenen

Erfassung

Vorwiegend physische Sensoren

Aggregation

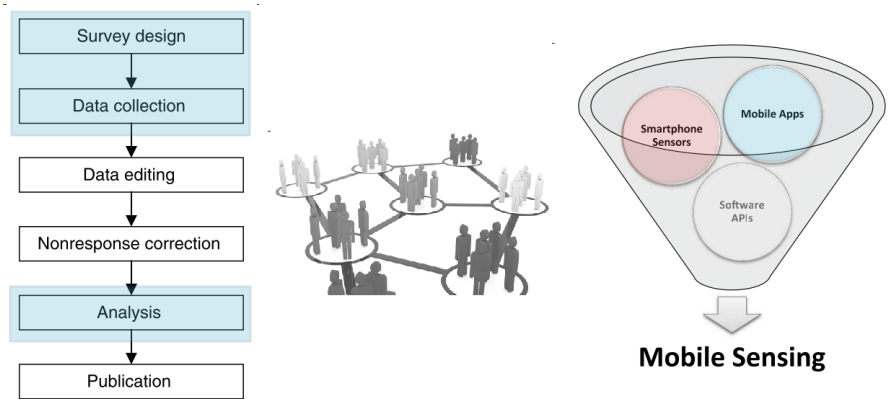
Virtuelle Sensoren, Datenreduktion (Größe und Dimensionalität)

Applikation

Datenanalyse und Modellbildung, Inferenz von Information, Maschinelles Lernen

Umfragen und Crowd Sensing

- Menschen sind Sensoren



[8]

Abb. 3. Von klassischen Umfragen zu mobilen Crowd Sensing mit Smartphones

Messfehler und Vertrauen

- Die Messgrößen können statisch (zeitlich konstant) oder dynamisch (zeitlich veränderlich) sein. Die Wandlung dieser Messgrößen ergeben dann entsprechend Gleich- und Wechselsignale.
- Auch eine prinzipiell zeitlich unveränderliche Messgröße (bezogen auf die Messung in einem vorgegeben Zeitinterval τ) erzeugt kein konstantes Signal. Ursache: Rauschen
- Wiederholt man daher eine Messung N-mal unter gleichen Bedingungen, so wird man eine Reihe von verschiedenen Messwerten $\{s_1,s_2,...,s_n\}$ erhalten.
- Es gibt systematische und zufällige Fehler bei der Messung, die sich überlagern.

Messfehler und Vertrauen

Systematische Abweichung (systematischer Fehler)

- Abweichung wird durch den Sensor verursacht
- z.B.: falsche Eichung, dauernd vorhandene Störungen wie Reibung
- lässt sich nur durch sorgfältiges Untersuchen der Fehlerquelle beseitigen

Zufällige Abweichung (zufälliger oder statistischer Fehler)

- Abweichung wird durch unvermeidbare, regellose Störungen verursacht
- bei wiederholter Messung weichen Einzelergebnisse voneinander ab
- Einzelergebnisse schwanken um einen Mittelwert

Messfehler und Vertrauen

[9]

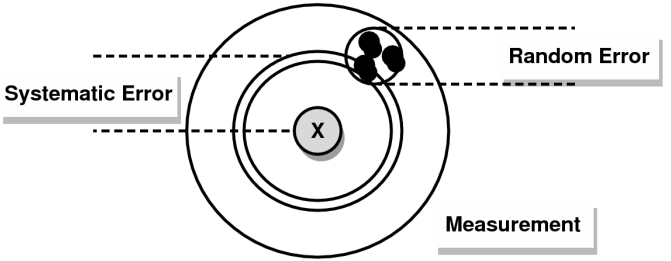


Abb. 4. Offset und Präzision bei der Messung einer Variable *X*

Messfehler und Vertrauen

Systematische Fehler

- Eine Messgröße X ist meistens durch störende Messgrößen Y,Z,... usw. überlagert:

$$K(X,Y,Z) \; : \; X \times Y \times Z \rightarrow S, K(x,y,z) \approx \sum_{n=0}^{m} a_n x^n + \sum_{n=0}^{m} b_n y^n + \sum_{n=0}^{m} c_n z^n$$

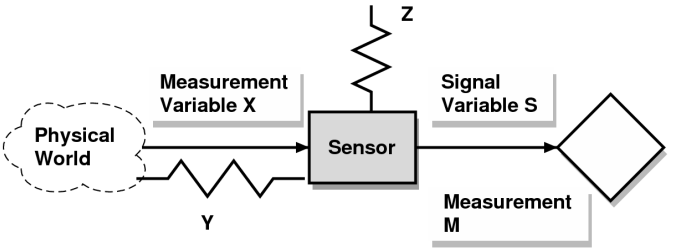
So kann z.B. bei einer Messung einer Kraft oder einer Dehnung die umgebende Temperatur T oder Strukturschwingungen Einfluss auf den Sensor und dessen Übertragungsfunktion und somit auf das Messsignal S haben.

Messfehler und Vertrauen

Systematische Fehler

So kann z.B. bei einer Messung von sozialpsychologischen Parametern der Wohnort und die Lebensumgebung Einfluss auf den Sensor und dessen "Übertragungsfunktion" und somit auf das "Messsignal" S haben.

Messfehler und Vertrauen



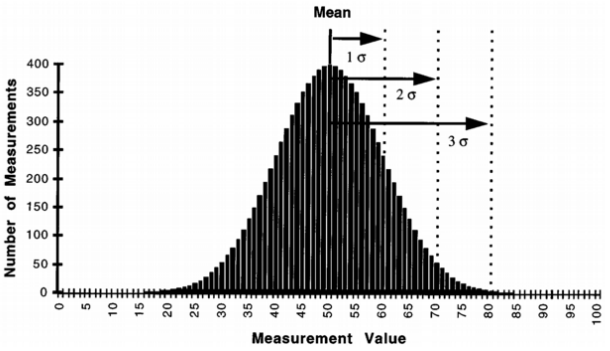
[9]

- Systematische Fehler verfälschen die Kalibrierungsfunktion (z. B. bei Geraden den Offset und Steigung). Sind sie bekannt, können sie kompensiert (rausgerechnet) werden.
- Systematische Fehler können aber auch während der Signalverarbeitung entstehen, so z.B.
  - Offsetspannungen und zeitlicher Drift von Parametern (Verstärkungsfaktor); durch
  - Rundungsfehler oder Verwendung von Funktionsmodellen außerhalb ihres Spezifikationsbereiches.

Messfehler und Vertrauen

Zufällige Fehler - Streuung

- Zufällige Fehler beeinflussen die Genauigkeit einer Messung (Rauschen).
- Rauschen beeinflußt die Berechung von Eingabedaten- und Zieleigenschaften (ML Ausgabe)!
- Wiederholt man eine Messung einer Größe X die durch reine zufälligen Fehler verfälscht wird, so ist die Häufigkeitsverteilung der Messwerte $S=\{s_1,s_2,...,s_n\}$ um einen Mittelwert $\bar{S}$ durch eine Gaussverteilung gegeben (dabei muss die Anzahl der Messungen N groß sein).



[9]

Abb. 5. Häufigkeitsverteilung nach Gauss von Messwerten um einen Mittelwert

Messfehler und Vertrauen

- Der Mittelwert S repräsentiert die Abschätzung des wahren/wirklichen Wertes Σ der Messgröße X (oder S):

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} s_i$$

- Die Standardabweichung ist ein Maß für die Zuverlässigkeit (Präzision) der einzelnen Messwerte einer Messreihe $\{s_1, s_2, \ldots, s_n\}$:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (s_i - \bar{S})^2}$$

Eine Vergrößerung der Anzahl N der Messungen (unter gleichen Bedingungen!) führt zu einer Verbesserung des Mittelwertes $\bar{S}$ (Grenzfall $N \rightarrow \infty$), nicht aber zu einer wesentlichen Verkleinerung der Standardabweichung σ, da die Genauigkeit nicht steigt!

Messfehler und Vertrauen

- Der wirkliche Mittelwert Σ ist nicht bekannt (nur im Grenzfall $N \rightarrow \infty$ ist $\bar{S}=\Sigma$) - Es gibt aber ein Vertrauensintervall mit einer Wahrscheinlichkeit *P* dass dieser darin enthalten ist:

  $\Sigma \in [\bar{S}-\sigma,\bar{S}+\sigma]$ mit 68.3%

  $\Sigma \in [\bar{S}-2\sigma,\bar{S}+2\sigma]$ mit 95.4%

  $\Sigma \in [\bar{S}-3\sigma,\bar{S}+3\sigma]$ mit 99.73%

Messfehler und Vertrauen

- Auch in der Soziologie!

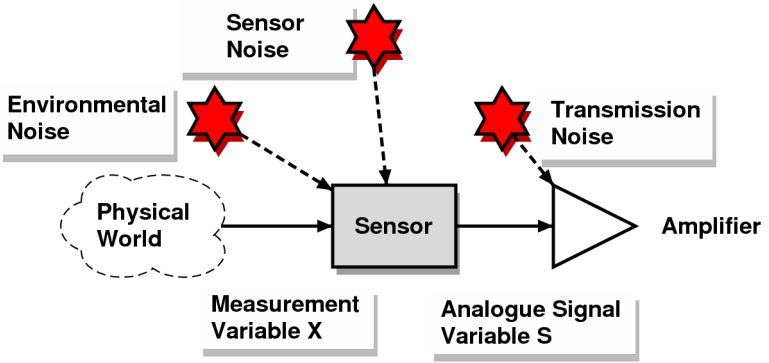


Abb. 6. Rauschquellen bei einer Messung

Beispiele: Statistische Analyse

Web WorkShell Live

CLEAR LOAD + - Plot Analysis

Zusammenfassung

- Daten können klassifiziert werden in:
  - Kategorische Variablen und Werte
  - Metrische Variablen und Werte
  - Zeitlich statische Variablen
  - Zeitl dynamische Variablen (Zeitserien)
- Alle Sensorvariablen unterliegen Messfehlern:
  - Rauschen
  - Verzerrung
  - Verschiebung (Bias)
  - Problem der Reproduzierbarkeit und systematische Fehler (Umgebung!)
- Eine (statistische) Datenanalyse ist häufig erster Schritt im ML Workflow