Maschinelles Lernen und Datenanalyse

In der Mess- und Prüftechnik PD Stefan Bosse

Universität Bremen - FB Mathematik und Informatik

Datenanalyse und Eigenschaftsselektion

Häufig sind die rohen sensorischen Daten(variablen) zu hochdimensional und abhängig voneinander

Reduktion auf wesentliche Merkmale kann ML Qualität deutlich verbessern

Häufig besitzen einzelne Sensorvariablen keine oder nur geringe Aussagekraft (geringe Entscheidbarkeitsqualität)

Datenqualität

- Die Daten $\boldsymbol{D}$ werden durch vier wesentliche Eigenschaften beschrieben, die auch mit statistischer Analyse quantifiziert werden können:

Rauschen. Rauschen ist die Verzerrung der Daten. Diese Verzerrung muss entfernt oder Ihre nachteiligen Auswirkungen vermindert werden, bevor ML Algorithmen ausgeführt werden, da die Leistung und Qualität der Algorithmen beeinträchtigen kwerden ann.

Es gibt eine Vielzahl von Filteralgorithmen um den Einfluß von Rauschen auf das eigentliche Sensorsignal zu vermindern.

Ausreißer. Ausreißer sind Instanzen, die sich erheblich von anderen Instanzen im Datensatz unterscheiden.

- Beispiel: Durchschnittliche Anzahl der Follower von Nutzern auf Twitter.
- Eine Berühmtheit mit vielen Followern kann die durchschnittliche Anzahl von Followern pro Person leicht verzerren. Da die Prominenten Ausreißer sind, müssen Sie aus der Gruppe der Personen entfernt werden, um die durchschnittliche Anzahl der Follower genau zu Messen.

Aber: Ausreißer können in besonderen Fällen nützliche Muster darstellen und die Entscheidung, sie zu entfernen, hängt vom Kontext und Fragestellung ab.

Fehlende Werte. Fehlende Werte sind Funktionswerte, die in Instanzen fehlen.

- Zum Beispiel, Einzelpersonen können es vermeiden, Profilinformationen auf social-media-Websites zu melden, wie Ihr Alter, Standort, oder Hobbys.
- Um dieses Problem zu lösen, können wir

  1. Instanzen mit fehlenden Werten entfernen;
  2. Fehlende Werte schätzen (Z. B. durch den gängigsten Wert ersetzen); oder
  3. Fehlende Werte ignorieren, wenn Data Mining Algorithmen ausgeführt werden.

Duplikate. Doppelte Daten treten auf, wenn mehrere Instanzen mit genau denselben Funktionswerten vorhanden sind.

- Doppelte blog-posts, doppelte tweets oder Profile auf Social-media-Websites mit doppelten Informationen sind Beispiele für dieses Phänomen.
- Je nach Kontext können diese Instanzen entweder entfernt oder beibehalten werden. Wenn Instanzen beispielsweise eindeutig sein müssen, sollten doppelte Instanzen entfernt werden.

Statistische Analyse

- Statistische Analysen von Mess- und Sensordaten können neue Datenvariablen erzeugen und Informationen über die Daten liefern:
  - Eigenschaftsselektion (Feature Selection) für ML und Informationsgewinnung
  - Variablentransformation mit Datenreduktion
- Statistische Analyse liefert eine Reihe von Kennzahlen über Datenvariablen, das können Eigenschaften für die Weiterverabeitung sein:

$$stat(\vec{x}) : \vec{x} \rightarrow \vec{p},$$
$$\vec{p} = \{mean, \sigma, ..\}$$

Welche statistische Größen gibt es? Was können statistische Größen über Daten aussagen?

Statistische Funktionen

Peak amplitude (y_{peak})	$y_{peak} = \max \lvert y_i \rvert$	[2:173]
Mean ($\overline{y}$)	$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$	
Mean square ($\overline{y}_{sq}$)	$\overline{y}_{sq} = \frac{1}{n} \sum_{i=1}^{n} (y_i)^2$	
Root-mean-square (rms)	$rms = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2}$	
Variance $(\sigma^2)^a$	$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$	

Statistische Funktionen

[2:173]

Standard deviation $(\sigma)^a$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$
Skewness (dimensionless) $(\gamma)^a$	$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{\sigma^3}$
Kurtosis (dimensionless) $(\kappa)^a$	$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{\sigma^4}$
Crest factor (X_{cf})	$X_{CF} = y_{peak}/rms$
K-factor (X_k)	$X_{CF} = (y_{peak})(rms)$

WorkBook Live: Statistische_Analyse

CLEAR LOAD + - dataan1 dataan2 dat

Korrelation von Datenvariablen

- Variablen **X** sollten möglichst (linear) unabängig sein,
  - Um eine geeignete, robuste und genaue Modellsynthese (also ML) zu ermöglichen, d.h.
  - Es sollte möglichst keine Zusammenhänge der Form $\exists correlation(X_i,X_j)$ geben!
  - Um den Modellsyntheseprozess zu beschleunigen (also das Training); Rechenzeit reduzieren
  - Um Modelle klein und kompakt zu halten
  - Um Modellspezialisierung zu vermeiden und Varianz zu erhöhen

Abhängige Variablen sollten identifiziert und in unabhängige "transformiert" werden!

Beispiel Prozessanalyse

- Eine Datentabelle $\boldsymbol{D}$ mit experimentellen Messgrößen und Fertigungsparametern (Prozessparameter) von additiv gefertigten Bauteilen hatte zunächst 7 Variablen (numerisch):
  - X_1: Hatchabstand [mm]
  - X_2: Scangeschwindigkeit [mm/s]
  - X_3: Laserleistung [W]
  - X_4: Schichtstärke [mm]
  - X_5: Volumenenergiedichte [J/mm ^{3}]
  - X_6: Bauplatten Position x
  - X_7: Bauplatten Position y
  - Y_1: Dichte (%)

- $\boldsymbol{D}$ bestand aus 61 Experimenten mit unterschiedlichen Fertigungsserien
- Mit einer Principle Component Analysis (PCA) konnte die ganze Tabelle auf die Variablen PC_1 und Y_1 reduziert werden!
  - Die Genauigkeit der mit ML synthetisierten Funktion $M(\boldsymbol{X})$: $\boldsymbol{X} \rightarrow \textit{Dichte}$ konnte ohne signifikanten Genauigkeitsverlust nur aus PC_1 abgeleitet werden, d.h. $M(PC_1)$: $PC_1 \rightarrow \textit{Dichte}$
  - Aber: Für die Inferenz (Applikation) von M muss die PCA für die Eingabedaten $\boldsymbol{X}$ wiederholt werden bzw. die Datentransformation durchgeführt werden!

Übung

Übung: Erstelle eine WorkBook um die Daten des botanischer Iris Datensatzes aus der SQL Datenbank einlesen kann. Dann sollen einfache statistische Analysen erstellt werden.

Welche Eigenschaften besitzen die Variabel X_1 bis X_4? Welche Verteilung besitzt die Variable Y?

Analyse Kategorischer Variablen

- Die Analyse von kategorischen Variablen vereint die Konzepte:
  - Mengenlehre
  - Kodierung/Dekodierung
  - Verteilung (Wahrscheinlichkeit des Auftretens)

[5]

Attribut	Wertemenge
Aussicht	sonnig, regnerisch, bewölkt
Temperatur	kalt, mild, heiß
Luftfeuchtigkeit	hoch, normal
Windig?	ja, nein

Abb. 1. Beispiel von rein kategorischen Attributen einer Datentabelle ***D***

Messdaten

[5:53]

Beispiel	Aussicht	Temperatur	Luftfeuchtigk.	Windig?	Klasse
1	sonnig	heiß	hoch	nein	N
2	sonnig	heiß	hoch	ja	N
3	bewölkt	heiß	hoch	nein	P
4	regnerisch	mild	hoch	nein	P
5	regnerisch	kalt	normal	nein	P
6	regnerisch	kalt	normal	ja	N
7	bewölkt	kalt	normal	ja	P
8	sonnig	mild	hoch	nein	N
9	sonnig	kalt	normal	nein	P
10	regnerisch	mild	normal	nein	P
11	sonnig	mild	normal	ja	P
12	bewölkt	mild	hoch	ja	P
13	bewölkt	heiß	normal	nein	P
14	regnerisch	mild	hoch	ja	N

Abb. 2. Beispiel einer rein kategorischen Datentabelle ***D***. Die Zielvariable *Klasse* mit den Werten {N,P} ist ebenfalls kategorisch, z.B. Klasse=P ⇒ Sportliche Aktivität

Gemischte Variablenklassen

[7:47]

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	85	85	False	5
Sunny	80	90	True	0
Overcast	83	86	False	55
Rainy	70	96	False	40
Rainy	68	80	False	65
Rainy	65	70	True	45
Overcast	64	65	True	60
Sunny	72	95	False	0
Sunny	69	70	False	70
Rainy	75	80	False	45
Sunny	75	70	True	50
Overcast	72	90	True	55
Overcast	81	75	False	75
Rainy	71	91	True	10

Abb. 3. Einige Datenvariablen wurden mit numerischen/metrischen Werten ersetzt (Klasse → Play-time)

Kann aus der vorherigen Datentabelle mit numerischen Variablen noch ein Zusammenhang aus $\boldsymbol{X}$ zu Y hergestellt werden?

Reicht die Anzahl der Experimente im Vergleich zu der rein kategorischen Datentabelle?

Wo liegen die Probleme?

Weiteres Beispiel

[7:7]

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	Hard
Prepresbyopic	Myope	No	Reduced	None
Prepresbyopic	Myope	No	Normal	Soft
Prepresbyopic	Myope	Yes	Reduced	None
Prepresbyopic	Myope	Yes	Normal	Hard
Prepresbyopic	Hypermetrope	No	Reduced	None
Prepresbyopic	Hypermetrope	No	Normal	Soft
Prepresbyopic	Hypermetrope	Yes	Reduced	None
Prepresbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

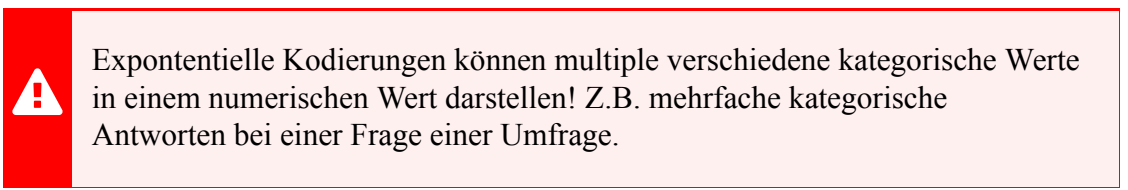
Kodierung

> Kategorische Variablen (sowohl Attribute als auch Zielvariablen) können von einer Vielzahl von numerisch basierten ML Verfahren nicht verarbeitet werden (wie neuronale Netze)

- Eine Lösung ist die Abbildung von kategorischen Werten (also Mengen von Symbolen) auf numerische Werte → **Kodierung**
- Kodierte Werte sind aber i.A. weder intervall- noch verhältnisskalierbar!

Kodierungsformate

- *Linear* und nicht akkummulativ (skalar), d.h.
  $\{\alpha,\beta,\gamma,..\} \rightarrow \{\delta,2\delta,3\delta,..\}$
- *Exponentiell* (z.B. zur Basis B=2) und akkummulativ (skalar), d.h.
  $\{\alpha,\beta,\gamma,..\} \rightarrow \{2^0,2^1,2^2,..\}$
- *One-hot* und evtl. akkummulativ (vektoriell), d.h.
  $\{\alpha,\beta,\gamma,..\} \rightarrow \{[1,0,0,...],[0,1,0,..],[0,0,1,..],..\}$



Beispiele

```
{sonnig,bewölkt,regnerisch} → {3,2,1}
{ja,nein} → {1,0}
{Schaden A, Schaden B, Schaden C} → { 1,2,3 }
{rot,grün,blau,braun,weiß} → {1,2,4,8,16}
{Sport, Kino, Theater, Musik} → {1,2,4,8}
{heiß,kalt} → {[1,0],[0,1]}
```

- Numerische/metrische Werte können auf kategorische durch Intervallkodierung reduziert werden:

$$cat(x) : x \rightarrow \{\alpha_1, \alpha_2, .., \alpha_n\}, x \in \mathbb{R}/\mathbb{N}$$
$$\alpha_i \leftrightarrow x = [x_0 + i\delta, x_0 + (i+1)\delta]$$

- Verwendung der `code=Math.code(val,codes)` und `val=Math.decode(code,codes)` Funktionen

WorkBook Live: Kodierung

CLEAR LOAD + - code1 code2

Entropie und Informationsgehalt

- Sensorvariablen können unterschiedlichen *Informationsgehalt* besitzen
  - Nur auf den Dateninhalt (Werte) der Variable X_i bezogen (inherenter Informationsgehalt)
  - Oder zusätzlich bezogen auf die Zielvariable Y (abhängiger Informationsgehalt)
- Der *Informationsgehalt* einer Menge X aus Elementen der Menge C wird durch die *Entropie* $E(X)$ gegeben:

$$E(X) = -\sum_{i=1,k} p_i \log_2(p_i), p_i = \frac{count(c_i, X)}{N}, X = \{c | c \in C\}$$

- Dabie ist k die Anzahl der unterscheidbaren Elemente/Klassen $Val(X) \subseteq C$ in der Datenmenge X (z.B. die Spalte einer Tabelle) und p_i die Häufigkeit des Auftretens eines Elements $c_\mathrm{i} \in C$ in X.

- Beispiele:

```
X1={A, C, B, C, B, C} → Val(X1)=C={A, B, C},N=6
E(X1)=-(1/6)log(1/6)-(2/6)log(2/6)-(3/6)log(3/6)=1.46
X2={A, B, C} → Val(X2)=C={A, B, C},N=3
E(X2)=-(1/3)log(1/3)-(1/3)log(1/3)-(1/3)log(1/3)=1.58
X3={A, A, A, A, B, B} → Val(X3)={A,B} ⊂ C={A, B, C},N=6
E(X3)=-(4/6)log(4/6)-(2/6)log(2/6)-(0/6)log(0/6)=0.92
X4={A, A, A, A, B, B} → Val(X4)=C={A, B},N=6
E(X4)=-(4/6)log(4/6)-(2/6)log(2/6)=0.92
```

- Die Entropie ist Null wenn die Datenmenge X "rein" ist, d.h., nur Elemente einer einzigen Attributklasse $c_1 \in C$ enthält, z.B. X={A,A,A}.
- Die Entropie ist $log_2(|C|)$ wenn alle Werte gleichverteilt vorkommen, wenn nicht dann kleiner (nicht gleichverteilt).
- Die Entropie reicht allein zur Bewertung des Informationsgehaltes nicht aus:

X1	X2	Y
A	C	P
B	C	P
A	D	N
B	D	N

- E(X1)=1, E(X2)=1 !! Welche Variable X ist für die Entscheidung der Zielvariable Y geeignet?

WorkBook Live: Entropie

CLEAR LOAD + - DataSports Analysis

Informationsgewinn (Gain)

- Ansatz: Die Datenmenge *Y* wird nach den möglichen Werten von *X* partitioniert, also je eine Partition pro $c_\mathrm{i} \in Val(X)$.

$$G(Y|X) = E(Y) - \sum_{v \in Val(X)} \frac{|Y_v|}{|Y|} E(Y_v)$$

- Die Menge Y_v enthält nur Werte für die X=v ist!

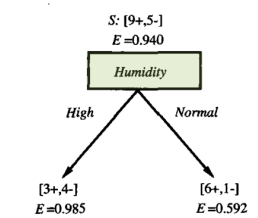
- Ein **Verteilungsvektor** ist dann *Dist*(*X*)=[$|v_1|$,$|v_2|$,..] und bedeutet wie häufig der bestimmte Wert $v_\mathrm{i} \in Val(X)$ in *X* auftaucht!

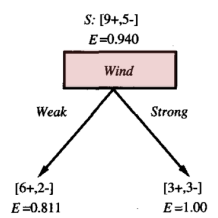
- Ein **Verteilungsvektor** ist dann $Dist(Y_v|X)=[|u_1|,|u_2|,..]$ und bedeutet wie häufig der bestimmte Wert $u \in Val(Y)$ in Y_v auftaucht!

Beispiele

Beispiel	Aussicht	Temperatur	Luftfeuchtigk.	Windig?	Klasse
1	sonnig	heiß	hoch	nein	N
2	sonnig	heiß	hoch	ja	N
3	bewölkt	heiß	hoch	nein	P
4	regnerisch	mild	hoch	nein	P
5	regnerisch	kalt	normal	nein	P
6	regnerisch	kalt	normal	ja	N
7	bewölkt	kalt	normal	ja	P
8	sonnig	mild	hoch	nein	N
9	sonnig	kalt	normal	nein	P
10	regnerisch	mild	normal	nein	P
11	sonnig	mild	normal	ja	P
12	bewölkt	mild	hoch	ja	P
13	bewölkt	heiß	normal	nein	P
14	regnerisch	mild	hoch	ja	N

[12]





WorkBook Live: Gain

CLEAR LOAD + - DataSports Analysis

Principle Component Analysis

- PCA: Klassische Methode zur unüberwachten linearen Dimensionsreduktion → Analyse der Hauptkomponenten

PCA ist noch keine Reduktionsmethode. PCA liefert bei einem n-dimensionalen Vektor $\vec{X}$ genau n Vektoren der Dimensionalität n!

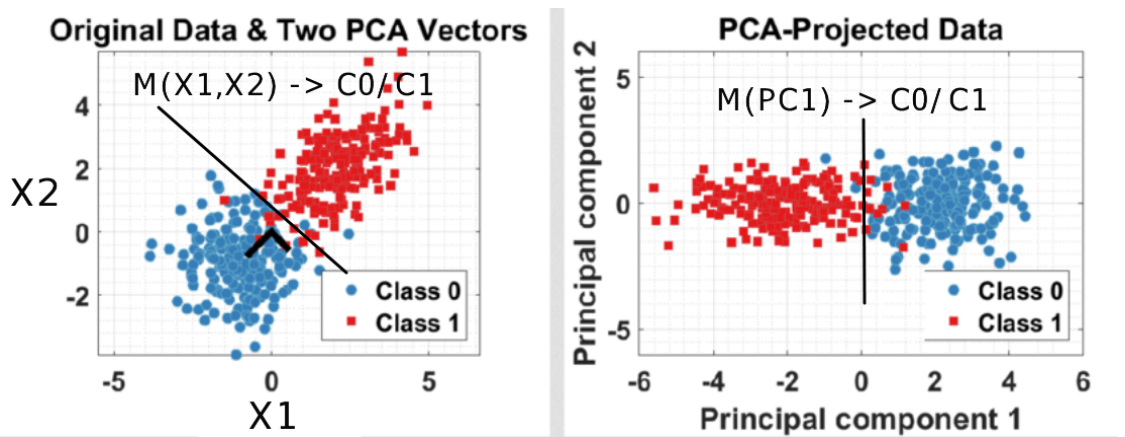
Aber: Reduktion von Redundanz in den Attributen $\boldsymbol{X}$ ist mit diesen Hauptkomponenten möglich.

- Bessere Trennung bei der Inferenz von kategorischen (und ggfs. auch numerischen) Zielvariablen
- Weitere Verfahren:
  - Lineare Diskriminanzanalyse (LDA)
  - Singuläre Wertzerlegung (SVD)

Beispiel

- $\boldsymbol{D}=[X_1,X_2,Y]$, mit $Val(Y)=\{\text{Class1},\text{Class2}\}$
- "Rotation" des zweidimensionalen Attributraums führt zu einer reduzierten Datentabelle $\boldsymbol{D'}=[PC_1,Y]$ (PC_2 kann weg gelassen werden)

[Czarnek, RG]



Amplitudenspektrum

$$Mag(X) = \frac{\sqrt{[re(DFT(X))]^2 + [im(DFT(X))]^2}}{N} \tag{1}$$

Phasenspektrum

$$Pha(X) = \arctan\left(\frac{im(DFT(X))}{re(DFT(X))}\right) \tag{2}$$

Leistungsspektrum

$$Power(X) = \frac{[re(DFT(X))]^2 + [im(DFT(X))]^2}{N} \tag{3}$$

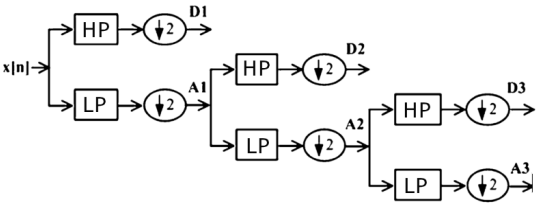
Waveletanalyse

- Diskrete Wavelet Transformation (DWT)
  - DFT liefert nur Informationen im Frequenzraum
  - DWT liefert Informationen aus Zeit- und Frequenzraum
  - Höherer Informationsgehalt

> DWT verwendet lange Zeitfenster für niedrige Frequenzen und kurze Zeitfenster für höhere Frequenzen, was zu einer guten zeitfrequenzanalyse führt.

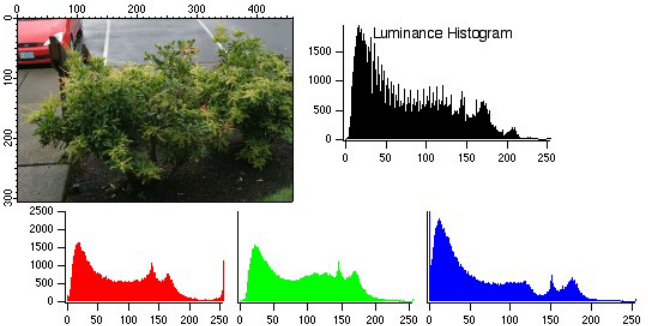
DWT kann mit digitalen Filterkaskaden aufgebaut werden:

- Jede Ebene der Filterkaskade besteht aus einem Hoch- und einem Tiefpassfilter
- Der Hochpassfilter liefert die Details, der Tiefpassfilter die Approximation der DWT auf der n-ten Ebene
- Die Approximation der Ebene *n* ist das Eingangssignal für die Ebene *n+1*
- In jeder Ebene wird der Eingangsvektor $|\mathbf{x}_i|=N$ auf $N/2$ reduziert, d.h. $|\mathbf{x}_{i+1}|=N/2$



Histogrammanalyse

- Ein Histogramm gibt die (intervallbasierte) Verteilung von unterscheidbaren Elementen in einer Datenmenge X an
  - Beispiel sind Histogramme von Bildern die die Verteilung der Farb-/Grauwerte mit den Werten == Kanälen {0,1,..,255} wiedergeben.



[wavemetrics]

Analyse von Datenserien

Welcher Sensoren können bei der Bauteilprüfung und Schadensüberwachung Zeit- oder Datenserien erzeugen..

Messdaten

[2:164]

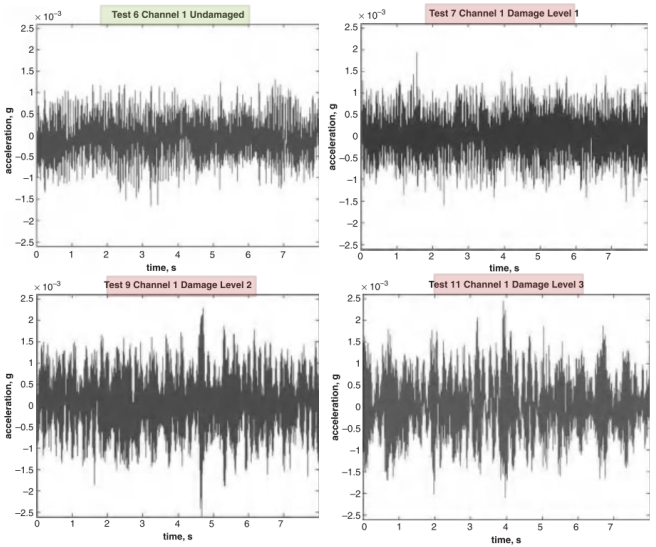
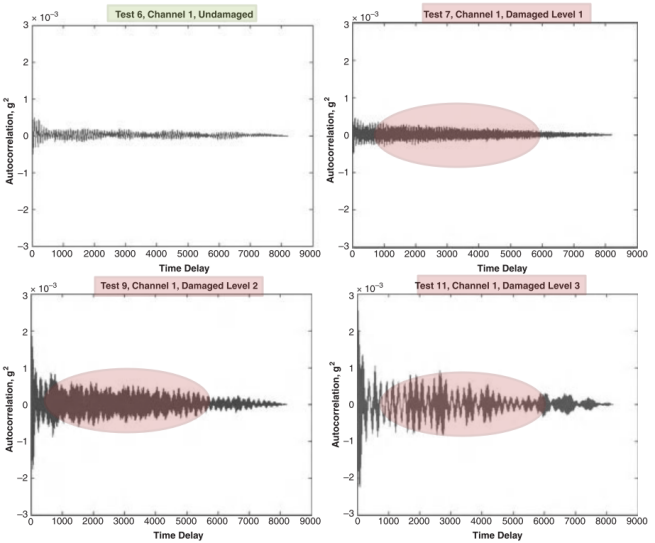


Abb. 4. Zeitaufgelöste Sensordaten $s(t)$ eines Beschleunigungssensors einer Maschine ohne und mit Schäden

Korrelationsanalyse



[2:164]

Abb. 5. Autokorrlelation der zeitaufgelösten Sensordaten $s(t)$ eines Beschleunigungssensors einer Maschine ohne und mit Schäden

Spektralanalyse

[2:167]

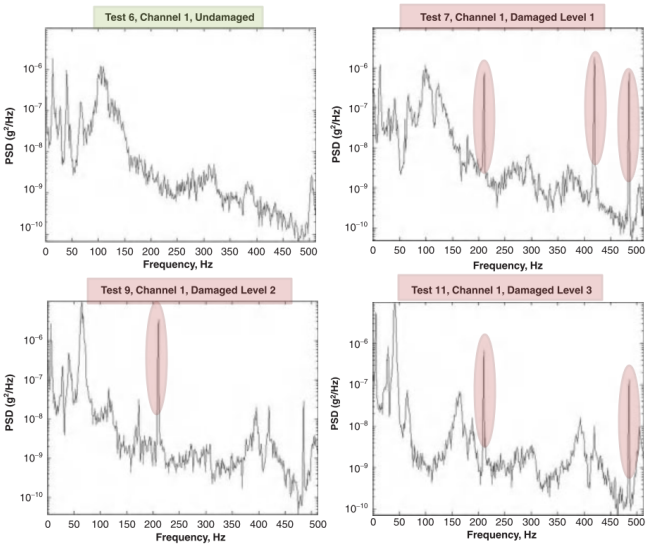


Abb. 6. Spektralanalyse der zeitaufgelösten Sensordaten $s(t)$ eines Beschleunigungssensors einer Maschine ohne und mit Schäden

Merkmalsselektion

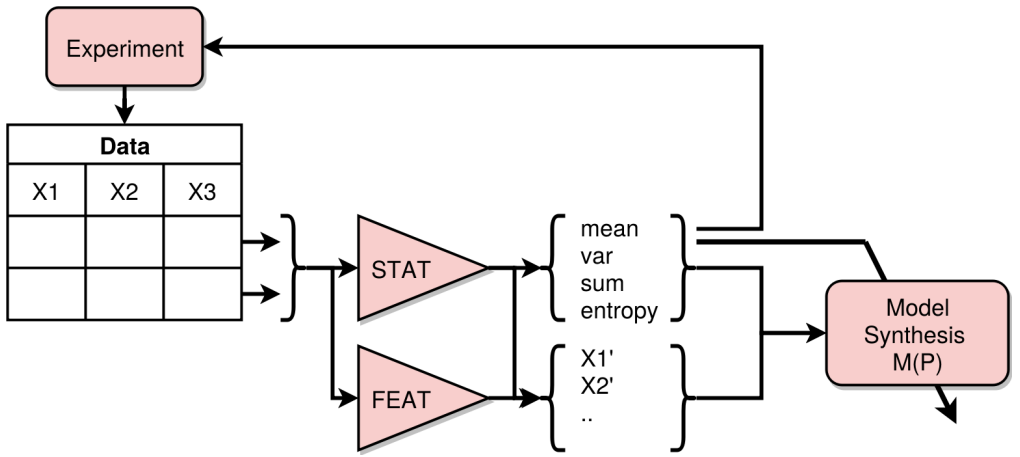


Abb. 7. Die statistische und weitere Analysen können die Eingabe für ML liefern, aber auch die Modellsynthese parametrisieren bzw. beeinflussen

Zusammenfassung

- Die statistische Analyse von Datentabellen liefert wichtige Informationen über die Qualität der Daten
- Die Merkmalsselektion transformiert die Rohdaten auf neue möglichst linear unabhängige Attribute
  - Datenreduktion → Dimensionalität
  - Datenreduktion → Datengröße
  - Datenqualitätserhöhung
- Es werden Verfahren für kategorische und numerische Datenvariablen unterschieden