Data Mining für Technische Anwendungen – Datenvorverarbeitung

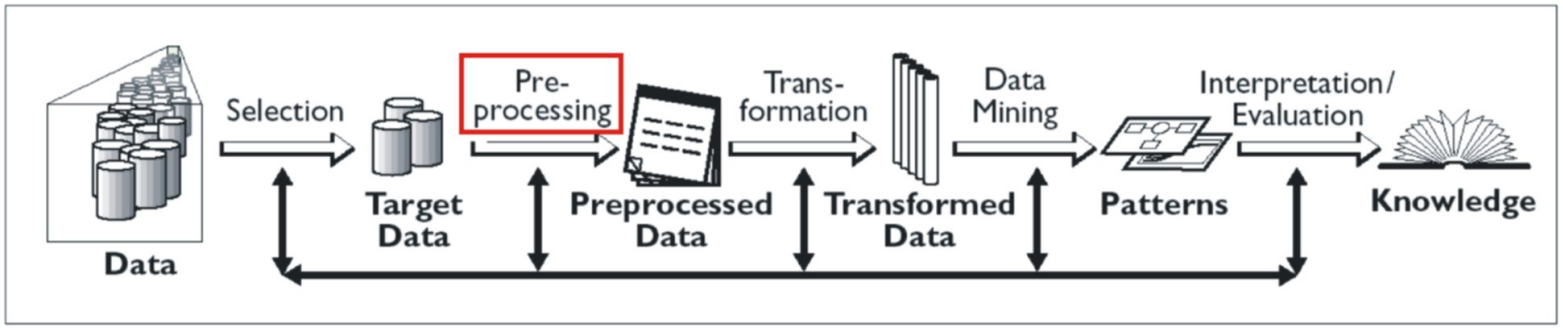
PD Dr.-Ing. habil. Sven Tomforde
Prof. Dr. Bernhard Sick

Universität Kassel
Fachbereich Elektrotechnik / Informatik
Fachgebiet „Intelligent Embedded Systems“

WS 2017/2018

Intelligent Embedded Systems

Agenda – 1



Datenvorverarbeitung

Aus der Definition von KDD: Aufbereitung der Daten durch Sichtung und Behandlung fehlerbehafteten oder fehlenden Datenmaterials (Identifizierung und Eliminierung von Ausreißern bzw. Rauschen), Entscheidung über Datenrepräsentation (z. B. Variablentypen, Darstellung von fehlenden bzw. unbekannten Daten)

Intelligent Embedded Systems

Agenda – 2

- Beziehungen zwischen Attributen
- Beziehungen zwischen Mustern
- Datenskalierung
- Ausreißer
- Datenkodierung

Intelligent Embedded Systems

Aufgaben

- Zusammenhänge zwischen zwei Attributen quantifizieren
- Zusammenhänge zwischen zwei Mustern quantifizieren
- Wertebereiche von Daten für das Data Mining geeignet transformieren
- Ausreißer in Datensätzen erkennen und ggf. behandeln
- nicht-numerische Daten geeignet kodieren, falls der DM-Algorithmus dies erfordert

Intelligent Embedded Systems

Erste Schritte des KDD ...

Intelligent Embedded Systems

Beziehungen zwischen Attributen

Datensatz

Jeder numerische Datensatz kann als Menge

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$$

geschrieben werden. Dabei ist D die Menge der Attribute, jeder Vektor $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nD})^{\mathrm{T}}$ ist ein Muster.

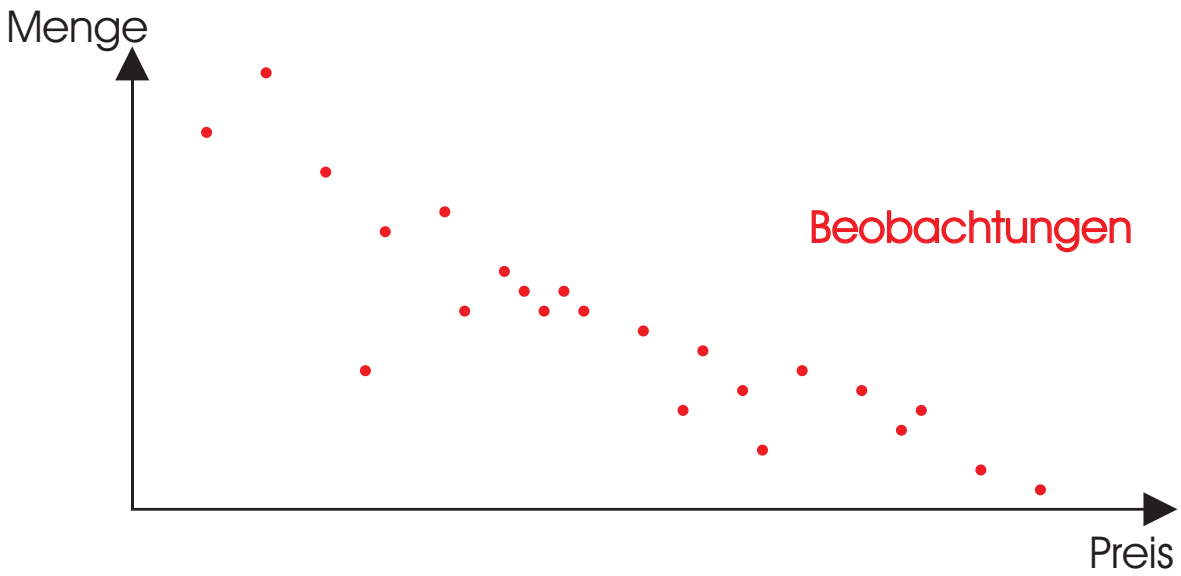
Attribute werden auch als Merkmale oder Features, Muster als Beobachtung oder Samples bezeichnet.
Ein Datensatz kann also beispielsweise durch Auflistung der Elemente angegeben werden (sogenannte Urliste).
Auswertung: Prinzipiell kann ein Muster auch eine komplette Zeitreihe (z. B. Sensorsignal), ein Text, ein Bild usw. sein.

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 1

Beziehungen zweier numerischer Attribute kann man sich oft in einem *Streudiagramm* (*scatter plot*) veranschaulichen.

Ein Streudiagramm ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale. Diese Wertepaare werden in ein Koordinatensystem eingetragen.



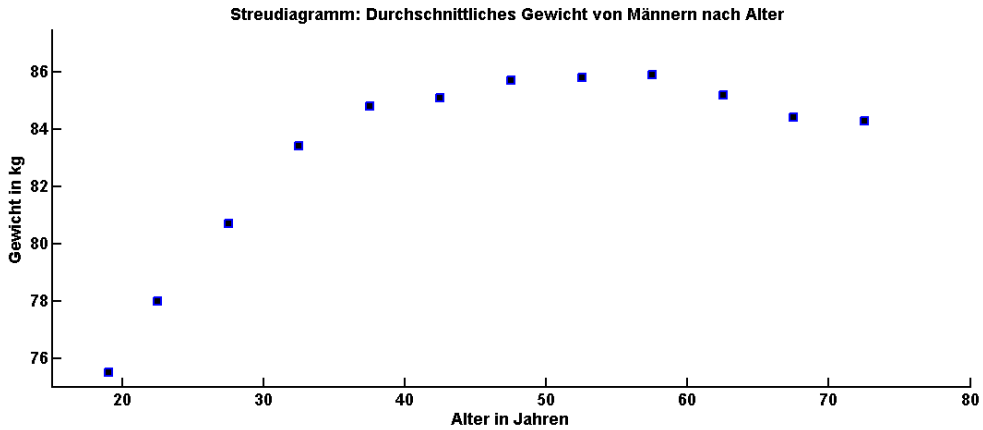
Bei einem höheren Preis werden offensichtlich geringere Mengen verkauft!

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 2

Ein weiteres Beispiel: durchschnittliche Gewichte von Männern nach Altersklassen (Ergebnisse des Mikrozensus 2009 durch das statistische Bundesamt). Für das Streudiagramm der Gewichte nach Alter wurden die Altersklassen durch die Klassenmitten ersetzt.

Alters-klasse	Klassen-mitte "Alter"	Durch-schnitts-gewicht (kg)	Durch-schnitts-größe (cm)
18 – 20	19	75,5	181
20 – 25	22,5	78,0	181
25 – 30	27,5	80,7	180
30 – 35	32,5	83,4	180
35 – 40	37,5	84,8	180
40 – 45	42,5	85,1	180
45 – 50	47,5	85,7	179
50 – 55	52,5	85,8	178
55 – 60	57,5	85,9	177
60 – 65	62,5	85,2	176
65 – 70	67,5	84,4	176
70 – 75	72,5	83,3	174



Auch hier hängen die beiden Attribute offensichtlich voneinander ab! Wie kann man den Zusammenhang messen?

Quelle: `https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/GesundheitszustandRelevantesVerhalten/Tabellen/Koerpermasse.html` Stand: 06.11.2017

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 3

Um die Beziehungen zweier Attribute i und j zu bewerten, wird zunächst das *arithmetische Mittel* gebildet:

$$\mu_i := \frac{1}{N}\sum_{n=1}^{N} x_{ni} \text{ und } \mu_j := \frac{1}{N}\sum_{n=1}^{N} x_{nj}.$$

Dann benötigt man die *empirischen Varianzen*:

$$\sigma_i^2 := \frac{1}{N-1}\sum_{n=1}^{N}(x_{ni}-\mu_i)^2 \text{ und } \sigma_j^2 := \frac{1}{N-1}\sum_{n=1}^{N}(x_{nj}-\mu_j)^2.$$

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 4

Die Kovarianz der beiden Merkmale wird so bestimmt:

$$s_{ij} := \frac{1}{N-1} \sum_{n=1}^{N} (x_{ni} - \mu_i) \cdot (x_{nj} - \mu_j)$$

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 5

Eigenschaften der Kovarianz:

- im Gegensatz zur (empirischen) Varianz kann sie negativ sein
- sie ist nicht normiert und kann beliebige reelle Werte annehmen
- ist sie positiv, haben beide Attribute gleiche Tendenz,
  ist sie negativ, haben beide entgegengesetzte Tendenz
- es gilt $s_{ij} = s_{ji}$ (Symmetrie)
- sie ist lage-invariant und linear, d. h., für
  $\hat{x}_{ni} = a \cdot x_{ni} + b$ und $\hat{x}_{nj} = c \cdot x_{nj} + d$ (mit $a, b, c, d \in \mathbb{R}$)
  gilt $\hat{s}_{ij} = a \cdot c \cdot s_{ij}$
- s_{ii} ist die Varianz

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 6

Der Korrelationskoeffizient ergibt sich aus der Normierung der Kovarianz mit den empirischen Standardabweichungen:

$$r_{ij} := \frac{s_{ij}}{\sigma_i \cdot \sigma_j}$$

Der Korrelationskoeffizient ist ein Maß für den Grad eines statistischen linearen Zusammenhangs zwischen zwei Merkmalen.

Er gibt an, wie sehr zwei Datensätze miteinander korrelieren, das bedeutet, inwieweit sie voneinander abhängig sind.

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 7

Für $r_{ij} = 1$ besteht ein exakter positiver linearer Zusammenhang, also z. B. hängt die Zahl der verwendeten Fahrradreifen linear von der Zahl produzierter Fahrräder ab.

Für $r_{ij} = -1$ besteht ein exakter negativer linearer Zusammenhang, also z. B.: je mehr Füchse desto weniger Hasen.

Je näher r an Null liegt, desto weniger besteht ein linearer Zusammenhang.

Das heißt aber nicht, dass gar kein Zusammenhang besteht, denn es kann auch ein anderer (z. B. nichtlinearer) Zusammenhang existieren, der sich nicht mit dem Korrelationskoeffizienten messen lässt.

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 8

Der Korrelationskoeffizient liefert umgekehrt noch keinen Nachweis eines ursächlichen Zusammenhangs (Kausalität): Der Rückgang der Besiedlung durch Störche in Deutschland korreliert zwar seit Jahren mit dem Geburtenrückgang, doch das bedeutet noch nicht, dass ein ursächlicher Zusammenhang besteht.

Die Interpretation eines Korrelationskoeffizienten als hoch oder niedrig hängt stark von der Art der korrelierenden Daten ab. Manchmal werden Werte bis etwa 0.3 als schwache Korrelation angesehen, während man ab 0.8 von einer sehr hohen Korrelation ausgeht.

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 9

Beispiel:

In ländlichen Regionen gibt es nur wenige Arbeitsplätze. Daher ist zu vermuten, dass viele Menschen aus ländlich geprägten Gemeinden pendeln, um in nahe gelegenen Städten einer Arbeit nachzugehen. Wenn jedoch die Land- und Forstwirtschaft einen traditionell hohen Stellenwert besitzt und viele Arbeitsplätze bietet, ist die Notwendigkeit des Pendelns nicht gegeben.

Fragestellung:

Gibt es einen Zusammenhang zwischen dem Anteil der Arbeitsplätze in der Land- und Forstwirtschaft in einer Gemeinde und dem Anteil der Berufspendler?

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 10

Beispieldatensatz:

Anteil Erwerbstätiger in der Landwirtschaft in %	Anteil der Berufspendler in %
78	10
23	62
60	18
74	12
32	48
12	69
16	63
65	28
43	35
70	17

für verschiedene Gemeinden

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 11

Anteil der Erwerbspersonen in der Ldw. (%)	Anteil der Auspendler an Erwerbspersonen (%)	μ_i	μ_j	$\mu_i \cdot \mu_j$	μ_i^2	μ_j^2
78	10	30,7	-26,2	-804,34	942,49	686,44
23	62	-24,3	25,8	-626,94	590,49	665,64
60	18	12,7	-18,2	-231,14	161,29	331,24
74	12	26,7	-24,2	-646,14	712,89	585,64
32	48	-15,3	11,8	-180,54	234,09	139,24
12	69	-35,3	32,8	-1157,84	1246,09	1075,84
16	63	-31,3	26,8	-838,84	979,69	718,24
65	28	17,7	-8,2	-145,14	313,29	67,24
43	35	-4,3	-1,2	5,16	18,49	1,44
70	17	22,7	-19,2	-435,84	515,29	368,64
47,3	**36,2**		**Summe:**	**-5061,6**	**5714,1**	**4639,6**

Intelligent Embedded Systems

Beziehungen zwischen Attributen – 12

Ergebnis:

$$r_{ij} = -0.9830$$

Eine hohe negative Korrelation. Es ist ein ausgeprägter gegenläufiger Zusammenhang zwischen den Variablen festzustellen.

Schlussfolgerung:

Je höher der Anteil der Erwerbspersonen in der Land- und Forstwirtschaft, desto niedriger der Pendleranteil.

Beziehungen zwischen Attributen – 13

Für weitere Maße siehe Literatur, z. B. Autokorrelation zur Messung von Zusammenhängen zwischen den beobachteten Werten zu verschiedenen Zeitpunkten einer Messreihe oder Rangordnungskoeffizient nach Spearman für ordinale Daten u. v. m.

Beziehungen zwischen Mustern

Datensatz

Oft lassen sich Zusammenhänge zwischen Paaren von Mustern durch eine *Relationsmatrix* angeben:

$$\mathbf{X} := \begin{pmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & \ddots & \vdots \\ r_{N1} & \cdots & r_{NN} \end{pmatrix} \in \mathbb{R}^{N \times N}.$$

Jeder Eintrag r_{ij} in $\mathbf{X}$ beschreibt die Ähnlichkeit oder Unähnlichkeit, die Kompatibilität oder Inkompatibilität, die Nähe oder den Abstand zweier Muster $\mathbf{x}_i$ und $\mathbf{x}_j$.

Im Allgemeinen gilt $r_{ij} = r_{ji}$, d. h., $\mathbf{X}$ ist symmetrisch.
Eine solche Matrix kann auf einer Metrik basieren, muss aber nicht.

Intelligent Embedded Systems

Norm und Metrik

Definition Norm

Eine Abbildung $\|.\| : \mathbb{R}^D \longrightarrow \mathbb{R}^+$ heißt *Norm*, genau dann, wenn

$$\begin{aligned} \|\mathbf{x}\| &\geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^D, \\ \|\mathbf{x}\| = 0 &\iff \mathbf{x} = (0, \ldots, 0)^{\mathrm{T}}, \\ \|a \cdot \mathbf{x}\| &= |a| \cdot \|\mathbf{x}\|, \quad \forall a \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^D, \\ \|\mathbf{x} + \mathbf{y}\| &\leq \|\mathbf{x}\| + \|\mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^D \end{aligned}$$

Eine Norm induziert durch $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$ eine *Metrik*.

Normen bzw. Metriken werden verwendet, um Relationen auf reellwertigen Datensätzen zu definieren.

Intelligent Embedded Systems

Abstände von Mustern – 1

Minkowsky-Normen sind definiert durch

$$\|\mathbf{x} - \mathbf{y}\|_q := \left(\sum_{j=1}^{D} |x_j - y_j|^q \right)^{\frac{1}{q}},$$

wobei x_j, $y_j \in \mathbb{R}$ die j-ten Elemente der Muster $\mathbf{x}$ und $\mathbf{y}$ darstellen.

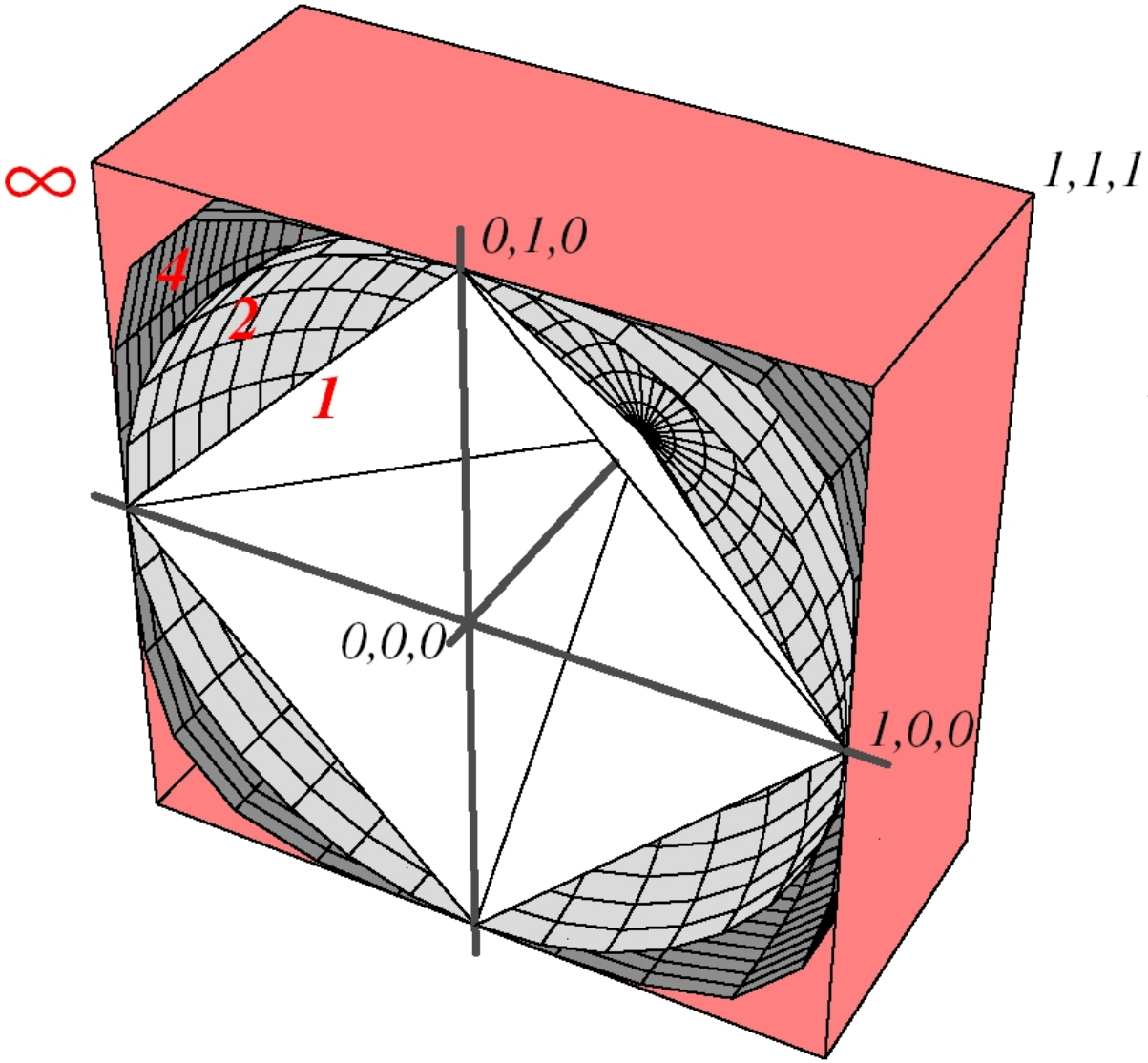
Abstände von Mustern – 2

Es gilt:

1. $q = 1$: *Manhattan-Abstand* $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_{j=1}^{D} |x_j - y_j|$,
2. $q = 2$: *Euklidischer Abstand* $\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{j=1}^{D} (x_j - y_j)^2}$,
3. $\lim_{q \longrightarrow \infty}$: *Supremums-Abstand* (*Maximums-Abstand*)
   $\|\mathbf{x} - \mathbf{y}\|_\infty = \max_{j=1,\ldots,D} \{|x_j - y_j|\}$.

Intelligent Embedded Systems

Abstände von Mustern – 3



Darstellung des Minkowsky-Abstands für verschiedene Werte von q im $\mathbb{R}^3$.

Die gefärbten Flächen haben jeweils den Abstand 1.0 zum Ursprung.

Quelle: [Duda, Hart, Stork, 2001]

Intelligent Embedded Systems

Abstände von Mustern – 4

Eine *Matrixnorm* ist definiert durch

$$\|\mathbf{x}-\mathbf{y}\|_{\mathbf{M}} := \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{y})}$$

für $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ und $\mathbf{M} \in \mathbb{R}^{D \times D}$.

Intelligent Embedded Systems

Abstände von Mustern – 5

Für

$$\mathbf{M} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

erhält man wieder die *Euklidische Norm* als Spezialfall.

Abstände von Mustern – 6

Allgemeiner spricht man bei

$$\mathbf{M} := \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_D \end{pmatrix}$$

mit beliebigen reellwertigen Diagonalelementen von einer *Diagonalnorm*.

Intelligent Embedded Systems

Abstände von Mustern – 7

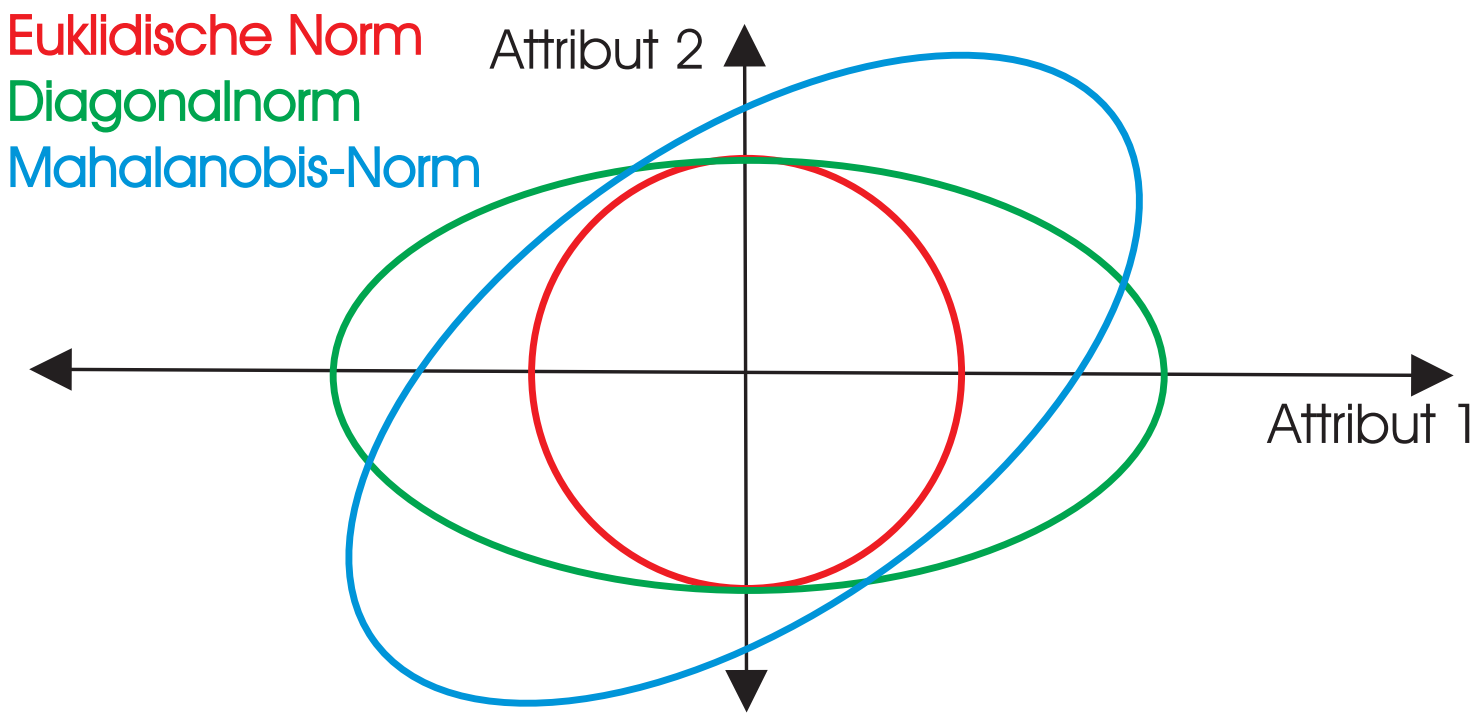
Über die Inverse der Kovarianzmatrix des Datensatzes ist die *Mahalanobis-Norm* definiert:

$$\mathbf{M} := \left(\frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}} \right)^{-1}$$

mit dem Mittelwert $\mu := \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$.
Achtung: Für einen Vektor $\mathbf{x} \in \mathbb{R}^D$ gilt $\mathbf{x}^{\mathrm{T}} \in \mathbb{R}$ und $\mathbf{x}\mathbf{x}^{\mathrm{T}} \in \mathbb{R}^{D \times D}$.

Intelligent Embedded Systems

Abstände von Mustern – 8



Die Punkte auf Kreis bzw. Ellipse haben bzgl. der gewählten Norm jeweils gleichen Abstand zum Ursprung.

(Datensätze sind hier nicht gezeigt!)

Intelligent Embedded Systems

Abstände von Mustern – 9

Weitere Beispiele für *Abstandsmaße* (nicht notwendigerweise Metriken!):

- **Cosinusdistanz:** normiertes Standardskalarprodukt zweier Vektoren:

$$d(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}$$

  (Cosinus des Winkels)

- **Hamming Abstand:**

$$\|\mathbf{x} - \mathbf{y}\|_{Hamm} := \sum_{j=1}^{D} \delta(x_j, y_j),$$

  wobei x_j, $y_j \in \mathbb{R}$ die j-ten Elemente der Muster $\mathbf{x}$ und $\mathbf{y}$ darstellen

  und $\delta(x_j, y_j) := \begin{cases} 0 & \text{falls } x_j = y_j \\ 1 & \text{sonst} \end{cases}$

  (z. B. bei Abständen von Wörtern)

Intelligent Embedded Systems

Abstände von Mustern – 10

- **Hyperbelabstand:**

$$\|\mathbf{x} - \mathbf{y}\|_{Hyp} := \prod_{j=1}^{D} |x_j - y_j| \,,$$

  wobei x_j, $y_j \in \mathbb{R}$ die j-ten Elemente der Muster $\mathbf{x}$ und $\mathbf{y}$ darstellen.
- **Tanimoto-Abstand:** Abstand zweier Mengen $\mathcal{S}_1$ und $\mathcal{S}_2$:

$$d(\mathcal{S}_1, \mathcal{S}_2) := \frac{|\mathcal{S}_1| + |\mathcal{S}_2| - 2 \cdot |\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1| + |\mathcal{S}_2| - |\mathcal{S}_1 \cap \mathcal{S}_2|}$$

- **Ordinale Attribute:** Nummerierung der Werte entsprechend ihrer Ordnung (Rang) und Verwendung des Betrages der Differenz der Ordnungszahlen.

Intelligent Embedded Systems

Abstände von Mustern – 11

Beispiel: Konstruktion eines Ähnlichkeitsmaßes aus einem Unähnlichkeitsmaß

Euklidischer Abstand $\|\mathbf{x} - \mathbf{y}\|$ zweier reellwertiger Vektoren (eigentlich ein Maß für Unähnlichkeit)

eingesetzt in Gauß-Funktion: $\exp\left(-\left(\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)\right)$ mit $\sigma \in \mathbb{R}^+$ ergibt ein Maß für Ähnlichkeit.

Intelligent Embedded Systems

Abstände von Mustern – 12

Dazu gibt es Verfahren der Attributberechnung, die z. B. reellwertige oder ganzzahlige Attribute bestimmen.

Beispiele:

- **Texte:** statistische Attribute wie Auftreten oder Häufigkeit von bestimmten Schlüsselwörtern; Verwendung von Kontextwissen (z. B. Apfel ist eine Frucht, Berücksichtigung der Ähnlichkeit von Buchstaben, z. B. ss und ß), ...
- **Bilder:** Farbhistogramme, zweidimensionale Fourier- oder Wavelettransformation, ...
- **Spracherkennung:** Hidden-Markov-Modelle, ...
- usw.

Intelligent Embedded Systems

Abstände von Mustern – 13

Fazit:

Es gibt Abstandsmaße, die Strukturen der Daten berücksichtigen (z. B. Mahalanobis-Abstand), und solche, die davon unabhängig sind (z. B. Euklidischer Abstand).

Für weitere Abstandsmaße siehe Literatur.

Intelligent Embedded Systems

Datenskalierung

Skalierung von Daten – 1

Problem: Unterschiedliche Wertebereiche von Attributen

Beispiel: Attribute Größe und Gewicht eines Menschen

- Misst man z. B. Größe in cm und Gewicht in kg, so liegen die auftretenden Werte in etwa derselben Größenordnung; Berechnung von Abständen von Mustern macht Sinn.
- Misst man z. B. Größe in m und Gewicht in g, so sind die auftretenden Werte in verschiedenen Größenordnungen; Berechnung von Abständen von Mustern macht keinen Sinn, da das Gewicht die Größe dominiert.

Lösung: *Normalisierung* oder *Standardisierung* der Werte (für jedes Attribut getrennt!)

Intelligent Embedded Systems

Skalierung von Daten – 2

Normalisierung:

Liegen die Werte eines Attributs im Intervall $[a, b]$, so werden sie so linear transformiert, dass die transformierten Werte im Einheitsintervall $[0, 1]$ liegen:

$$x' = \frac{x - a}{b - a}.$$

wobei x der zu transformierende Wert ist und x' der transformierte Wert.

Die Werte von a und b können der minimale und der maximale in einem Datensatz für das Attribut auftretende Wert sein.

Intelligent Embedded Systems

Skalierung von Daten – 3

Problem der Normalisierung:

1. Bei neuen Daten (z. B. in der Anwendung) können Werte außerhalb des Intervalls $[a, b]$ auftreten.
2. Einzelne Ausreißerwerte können dazu führen, dass der zur Verfügung stehende Wertebereich $[0, 1]$ sehr schlecht ausgenutzt wird.

   Beispiel: Fast alle Kunden eines Supermarkts kaufen zwischen 0 und 10 Stück eines Produkts, nur ein einzelner Kunde (z. B. eine Firma) bezieht 10 000 Stück auf einmal.

   Lösung: Standardisierung, die diesen Ausreißereffekt vermeidet.

Intelligent Embedded Systems

Skalierung von Daten – 4

Standardisierung: (oder *Mahalanobis-Skalierung*)

Standardisierung transformiert die Daten so, dass sich ein Mittelwert von 0 und eine Streuung (empirische Standardabweichung) von 1 ergibt:

$$x' = \frac{x - \mu}{\sigma}.$$

Dabei ist μ der Mittelwert der Werte dieses Attributs und σ die empirische Standardabweichung.

Intelligent Embedded Systems

Skalierung von Daten – 5

Mittelwert μ_i der Werte $x_i n(k)$ eines Attributs i und *empirische Varianz* σ_i^2:

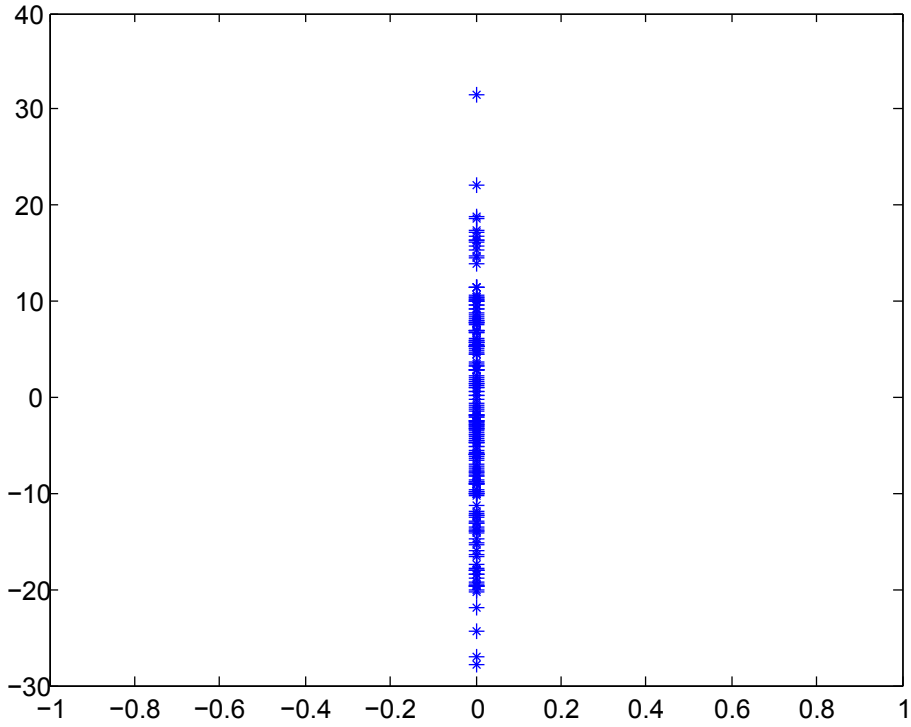
$$\mu_i := \frac{1}{N} \sum_{n=1}^{N} x_{in}$$

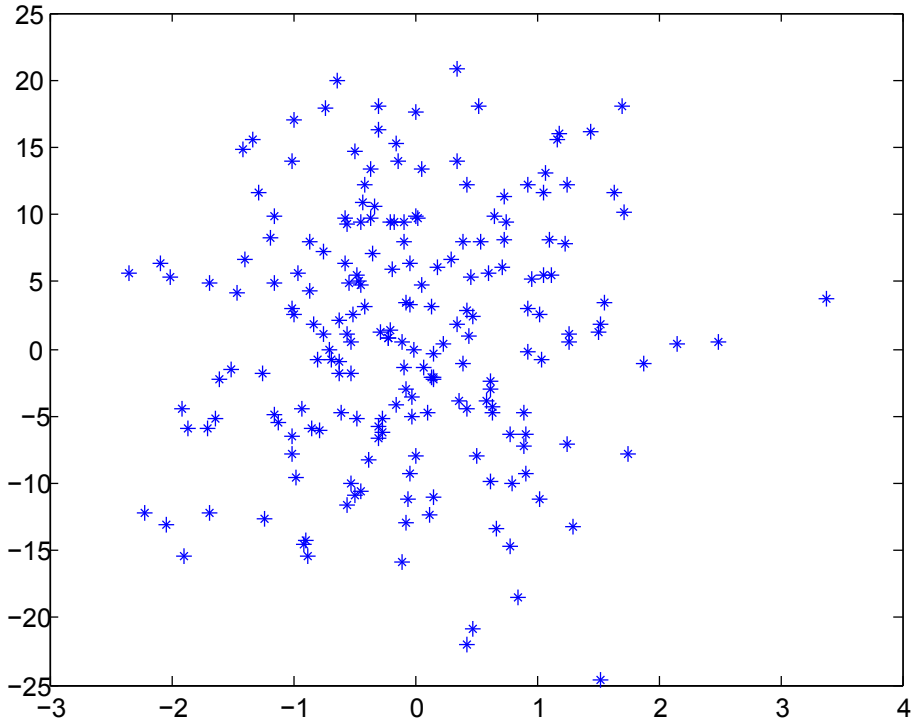
$$\sigma_i^2 := \frac{1}{N-1} \sum_{n=1}^{N} (x_{in} - \mu_i)^2$$

Die *empirische Standardabweichung* oder *Streuung* ist die Wurzel aus der empirischen Varianz.

Intelligent Embedded Systems

Skalierung von Daten – 6





Originaldatensatz (links): Gaussscher Zufallsprozess mit Mittelwert $(0, 0)$ und Standardabweichung $(0.1, 10)$.

Intelligent Embedded Systems

Skalierung von Daten – 7

Sofern kein kanonisches Ähnlichkeitsmaß oder Unähnlichkeitsmaß zu vorliegenden Daten bekannt ist, sollten alle Attribute

- normalisiert oder
- standardisiert

werden.

Intelligent Embedded Systems

Skalierung von Daten – 8

Wichtig: Sowohl Normalisierung als auch Standardisierung werden für jedes Attribut einzeln (d.h. unabhängig von den anderen) durchgeführt.
Weitere Skalierungsverfahren siehe Literatur, z. B.

- reziproke Skalierung (Kehrwert)
- logarithmische Skalierung
- Wurzelskalierung
- multidimensionale Skalierungsverfahren
- ...

Intelligent Embedded Systems

Ausreißer

Ausreißer – 1

Bei einigen Mustern können die Werte von Attributen ungenau, gestört, oder verfälscht sein (vgl. auch Missing Values).

Mögliche Ursachen:

- Sensorrauschen bei Messung physikalischer Größen
- Übertragungsfehler
- falsche Auskunft bei Interviews (z. B. Frage nach Alter oder Gewicht)
- ...

Solche *Ausreißer* (*outlier*) sollen erkannt und geeignet behandelt werden.

Ausreißer – 2

Erkennung von Ausreißern: ein Muster wird als Ausreißer identifiziert, wenn der Wert mindestens eines Attributs

- außerhalb eines zulässigen Wertebereichs liegt.
- um mehr als zwei oder drei mal die Standardabweichung vom Mittelwert abweicht (statistisches Maß).
- um mehr als einen vorgegebenen Betrag von einem mit einem geeigneten Modell geschätzten Wert abweicht.
- ...

Problem: Unterscheidung der Ausreißer von *Exoten* (korrekte, aber ungewöhnliche Daten, die jedoch wertvolle Informationen tragen).

Intelligent Embedded Systems

Ausreißer – 3

Behandlung von Ausreißern:

Verschiedene Möglichkeiten, abhängig davon, wie stark der Datensatz modifiziert wird:

- Markierung (nur für manche DM-Algorithmen geeignet, vgl. auch Missing Values)
- Entfernung des entsprechenden Musters oder Kennzeichnung des Ausreißers als „ungültig“
- Korrektur des Wertes

Intelligent Embedded Systems

Ausreißer – 4

Einfache Techniken zur Korrektur:

- Ersetzung durch Maximal- oder Minimalwert
- Ersetzung durch globalen Mittelwert
- lineare oder nichtlineare Interpolation bei Zeitreihen

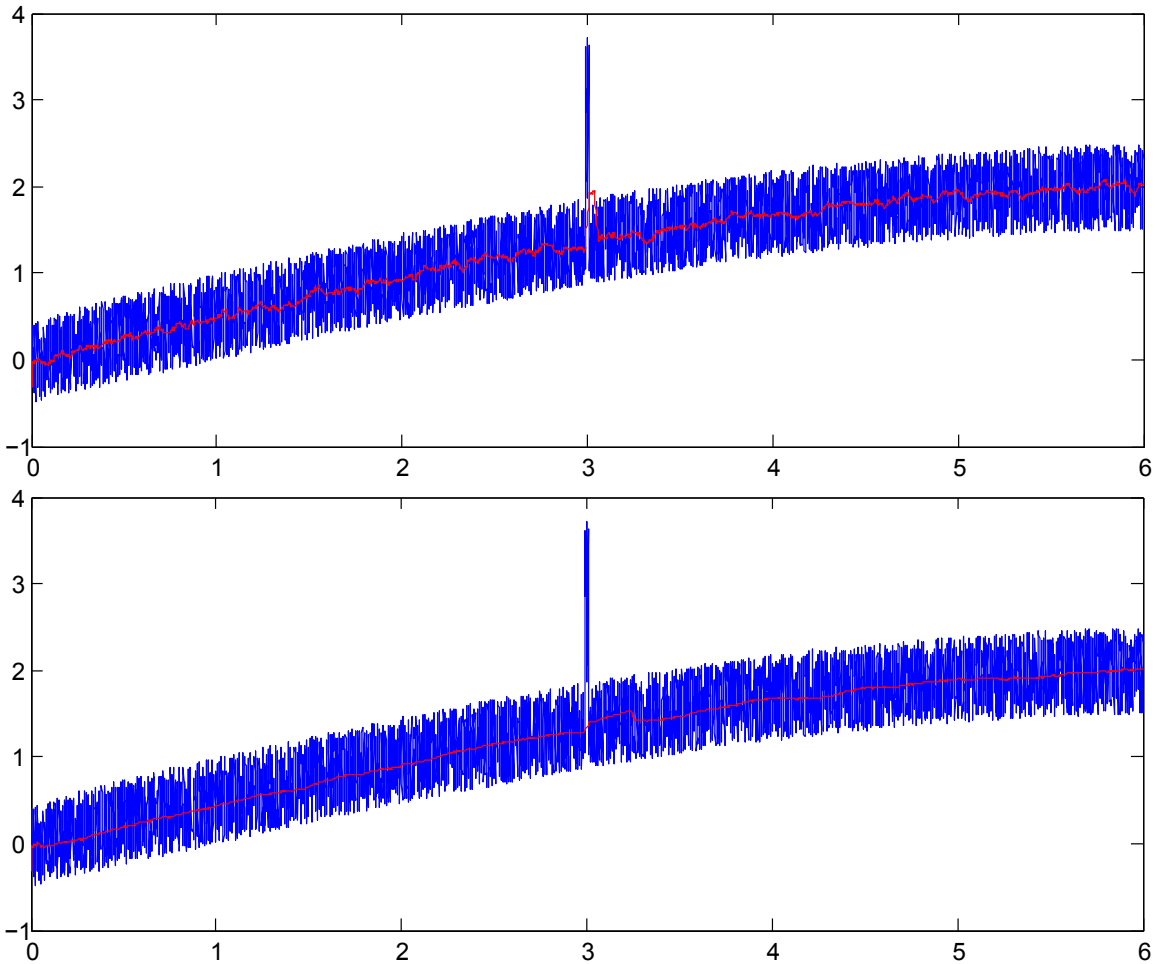
Weitere Techniken:

- Filterung bei Zeitreihen, z. B. FIR-Filter
- modellbasierte Ergänzung durch Zeitreihenmodelle, z. B. ARMA-Modelle
- modellbasierte Ergänzung durch Schätzung von Wahrscheinlichkeitsdichten
- usw.

Intelligent Embedded Systems

Ausreißer – 5

Beispiel: Elimination von Ausreißern durch gleitenden Mittelwert bei einer Zeitreihe



Originaldatensatz mit Ausreißer, Ergebnis der Filterung durch kleines Zeitfenster (oben), durch großes Zeitfenster(unten)

Datenkodierung

Datenkodierung – 1

Manche DM-Algorithmen arbeiten nur auf numerischen Daten.
Nicht-numerische Daten müssen also geeignet kodiert werden.

- **Ordinale Attribute:** rangbasierte Kodierung
- **Nominale Attribute:** orthogonale Kodierung (z. B. 1-aus-k-Kodierung: 00...010...00), wenn k die Zahl der Möglichen Ausprägungen des Attributs ist.

Manchmal bei Kodierung von Klassen: orthogonale Kodierung, wobei die Länge des Vektors die Klassenstärke (Zahl der in den Trainingsdaten verfügbaren Muster) widerspiegelt.

Intelligent Embedded Systems

Datenkodierung – 2

Beispiel für rangbasierte Kodierung:

Ausbildung	Repräsentation
Hauptschulabschluss	1
Realschulabschluss	2
Abitur	3
Diplom	4
Promotion	5

Intelligent Embedded Systems

Datenkodierung – 3

Beispiel für orthogonale Kodierung von Klassen bei quadratischem Fehler als Fehlermaß bei der Modellbildung:

Klasse	Klassenstärke	Repräsentation
$\mathcal{A}$	$\lvert\mathcal{A}\rvert$	$\left(\frac{1}{\sqrt{\lvert\mathcal{A}\rvert}}, 0, 0, 0, 0\right)^{\mathrm{T}}$
$\mathcal{B}$	$\lvert\mathcal{B}\rvert$	$\left(0, \frac{1}{\sqrt{\lvert\mathcal{B}\rvert}}, 0, 0, 0\right)^{\mathrm{T}}$
$\mathcal{C}$	$\lvert\mathcal{C}\rvert$	$\left(0, 0, \frac{1}{\sqrt{\lvert\mathcal{C}\rvert}}, 0, 0\right)^{\mathrm{T}}$
$\mathcal{D}$	$\lvert\mathcal{D}\rvert$	$\left(0, 0, 0, \frac{1}{\sqrt{\lvert\mathcal{D}\rvert}}, 0\right)^{\mathrm{T}}$
$\mathcal{E}$	$\lvert\mathcal{E}\rvert$	$\left(0, 0, 0, 0, \frac{1}{\sqrt{\lvert\mathcal{E}\rvert}}\right)^{\mathrm{T}}$

Intelligent Embedded Systems

Ende

–

Fragen zum Thema "Datenvorverarbeitung"?