

Maschinelles Lernen

Datenvorverarbeitung

Prof. Dr. Rainer Stollhoff

Univariate Vorverarbeitung - Datentypkonvertierung

• Nominelle Variablen

– Character \leftrightarrow Factor

- Für viele Verfahren nicht notwendig, da automatisch angewandt

– Dummy-Kodierung (One-Off-Kodierung)

- Übersetzt nominalen Faktor mit n Werten in n separate 0/1-Variablen (Variante $n-1$ Variablen)
- Bei manchen Verfahren notwendige Vorverarbeitung, bei anderen Verfahren integriert

Farbe: rot grün blau

rot	grün	blau
0,1	0,1	0,1
:	:	:
:	:	:

$k=1$ $k=2$

Farbe
grün
rot
blau
rot
:
:

$i=1$ $i=2$

rot	grün	blau
0	1	0
1	0	0
0	0	1

• Numerische Variablen

– Zahlenwert als Faktor

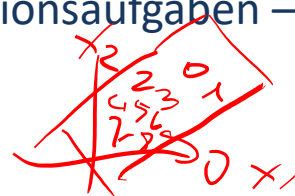
- Notwendig: falls Klassifikationsaufgabe und Klassenkodierung als Zahlenwert – sonst automatisch Regressionsverfahren
- Ermöglicht Einsatz von Klassifikationsverfahren in Regressionsproblemen
- Diskretisierung z.B. zur Visualisierung als Histogramm

– Faktor als Zahlenwert

- Ermöglicht Einsatz von Regressionsverfahren für Klassifikationsaufgaben – aber Vorsicht: nur sinnvoll für ordinale Merkmale!

0. ~~1~~ Bayer 36 Bb Be \rightarrow Faktor \rightarrow Dummy

1 2 3 4



Univariate Vorverarbeitung – Imputation fehlender Werte

Mittelwert / Median

- Ersetzt für eine Beobachtung fehlende Werte in einer Variable durch den Mittelwert bzw. Median dieser Variable in anderen Beobachtungen
- Vorteile
 - Einfach und Robust
 - Alle Beobachtungen können verwendet werden
- Nachteile
 - Ignoriert Zusammenhänge zwischen Variablen

$$\bar{x}_1 = \frac{7,2 + 10,2}{2} = 8,7$$

i	x ₁	x ₂
1	7,2	14,2
2	10,2	NA,4
3	NA,7	5,0

(lokal)-lineare Modell

- Schätzt fehlende Werte in einer Variablen anhand eines Regressionsmodells, das auf allen anderen Variablen geschätzt wird $x_1 = f(x_2)$
 $x_2 = g(x_1)$
- Vorteile
 - Berücksichtigt Zusammenhänge zwischen Variablen
- Nachteile
 - Zusätzlicher Rechenaufwand
 - weniger robust

Univariate Vorverarbeitung - Transformation

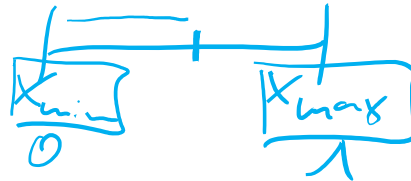
Standardisierung der Variablen

- Abziehen des Mittelwerts und teilen durch die Standardabweichung

$$z = \frac{x - \bar{x}}{\sigma(x)}$$


- Normalisierung des Wertebereichs auf [0,1]

$$\check{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$



• Vorteile

- Eingangsvariablen in Modelle haben vergleichbare Skala / Auflösung
- Parameter in einem Regressionsmodell können direkt verglichen werden

$$|\theta_1| \cdot x_1 + |\theta_2| \cdot x_2$$

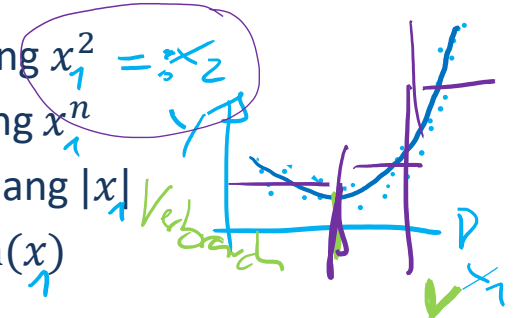
• Nachteile

- Einheiten gehen verloren
- Zusätzlicher Rechenaufwand

Nicht-lineare Zusammenhänge aufnehmen

- Beispiele:

- Quadratischer Zusammenhang x^2
- Polynomieller Zusammenhang x^n
- Betragsmäßiger Zusammenhang $|x|$
- Logarithmische Skalierung $\ln(x)$



• Vorteile

- Abbildung von Vorkenntnissen z.B. physikalischer Gesetzmäßigkeiten $E = \frac{1}{2}mv^2$
- Erweiterung linearer Verfahren z.B. lineare Regression

• Nachteile

- Zusätzlicher Rechenaufwand
- Kein Vorteil bei Verfahren mit eingebauter Transformation bzw. Unabhängigkeit ggB. Transformationen z.B. Bäume / rekursive Partitionierung

Multivariate Vorverarbeitung

Anzahl Variable

Dimensionsreduktion

(vgl. Unsupervised Learning)

– Korrelationen herausnehmen

- Hauptkomponentenanalyse durchführen
- Statt Variablen Hauptkomponenten verwenden

– Embeddings

- Vielzahl von Dummy-Variablen in metrischen Raum einbetten (z.B. Worräume)

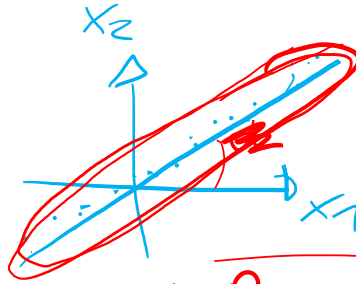
und da war es gut
0,1 0,1 0,1 0,1 0,1

• Vorteile

- Geringere Speicherbelegung
- Notwendig in linearer Regression, falls Ausgangsvariablen linear abhängig

• Nachteile

- U.U. Verlust von Informationen
- Verfahren sind datengetrieben



$$y = \theta_1 \cdot x_1 + \theta_2 \cdot x_2$$
$$\sim \theta_1 \cdot x_1 \cdot x_2 + \theta_2 \cdot x_2$$

Interaktionen berechnen

– Beispiele:

- Zweifaches Produkt $x_1 \cdot x_2$
- mehrfaches Produkt $x_1 \cdot x_2 \cdot x_3 \dots$
- Exponent $x_1^{x_2}$
- Quadratischer Abstand $(x_1 - x_2)^2$



• Vorteile

- Abbildung von Vorkenntnissen z.B. physikalischer Gesetzmäßigkeiten $E = \frac{1}{2}mv^2$
- Erweiterung linearer Verfahren z.B. lineare Regression

• Nachteile

- Zusätzlicher Rechenaufwand
- Kein Vorteil bei Verfahren mit eingebauter Interaktionsmöglichkeit z.B. Bäume/rekursive Partitionierung

da es gut
tal 2 super