Data Literacy Education

FOM

Special 6: Vorhersagemodellierung

Inhalt

Themen von Special 6: Vorhersagemodellierung

- Lineare Regression

Aufgabenstellung:

Modelliere die abbhängige Variable `total.bill`, y, auf Basis der unabhängigen Variablen `time` und `size`, x_1, x_2:

$$y = f(x_1, x_2) + \epsilon$$

Trainingsdaten

Der Trainingsdatensatz (`tips_train`) enthält alle Variablen, d.h. x_1, x_2, y:

Unabhängige Variablen		Abhängige Variable
time	size	total_bill
Dinner	4	21.50
Dinner	2	19.82
Dinner	3	45.35
Dinner	2	11.02

Modellierung

Schätze $f(\cdot)$ z.B. über lineare Regression auf den **Trainingsdaten**:

```
erg_lm <- lm(total_bill ~ time + size, data = tips_train)
erg_lm
```

```
##
## Call:
## lm(formula = total_bill ~ time + size, data = tips_train)
##
## Coefficients:
## (Intercept)    timeLunch         size
##       5.165       -2.131        5.905
```

$$\widehat{\text{total\_bill}}_i = 5.17 - 2.13 \cdot \begin{cases} 1, & \text{i ist Lunch} \\ 0, & \text{i ist nicht Lunch} \end{cases} + 5.9 \cdot \text{size}_i$$

Anwendungsdaten

Die **Anwendungsdaten** (`tips_anwendung`) enthalten nur die unabhänigen Variablen x_1, x_2, die abhängige Variable y nicht:

Unabhängige Variablen	
time	**size**
Lunch	6
Dinner	3
Lunch	2
Lunch	2

Vorhersage

Zuvor gelerntes Modell (`erg_lm`) auf Basis der Trainingsdaten (`tips_train`) zur Prognose der Zielvariable `total_bill` auf den Anwendungsdaten verwenden:

```
lm_predict <- predict(erg_lm, newdata = tips_anwendung)
```

Für die Beobachtungen des Anwendungsdatensatzes gibt es jetzt geschätze Werte für die Rechnungshöhe, $\widehat{\text{total\_bill}}$. Z.B. für $i = 1$:

Unabhängige Variablen	
time	**size**
Lunch	6

$$\widehat{\text{total\_bill}}_1 = 5.17 - 2.1307 \cdot 0 + 5.9 \cdot 6 = 38.46$$

Abhängige Variable
hat_total_bill
38.46

Evaluierung

Sind die wahren Werte der Zielvariable bekannt, kann die Vorhersage evaluiert werden:

Unabhängige Variablen		Abhängige Variable	
time	size	total_bill	hat_total_bill
Lunch	6	34.30	38.46
Dinner	3	35.83	22.88
Lunch	2	16.66	14.84
Lunch	2	10.33	14.84

Z. B.:

$$MAE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |y_i - \hat{y}_i|$$

Tipps

- Je nachdem, welche Variablen zur Modellierung verwendet werden, ergeben sich i.d. R. verschiedene Prognosen: `lm(y ~ 1); lm(y ~ x1); lm(y ~ x1 + x2); lm (y ~ x1 * x2)`

- Werden im Trainingsdatensatz Ausreißer eliminiert ändert sich das geschätzte Modell und damit die Prognose.

- Werden Variablen transformiert (z.B. `mutate(x1l = log(x1))`) ändert sich das geschätzte Modell und damit die Prognose.

- Werden unterschiedliche Modellierungsmethoden (`lm(); rpart()`, ...) verwendet, ändert sich die Prognose.

- Über Kreuzvalidierung o. ä. kann die Vorhersagegüte rein auf Basis der Trainingsdaten evaluiert werden. Dabei wird der Trainingsdatensatz wiederholt selber in einen Teildatensatz zum Modell schätzen und einen zum Modell evaluieren aufgeteilt.