Modulkonzept zu Statistik

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept Versuch 4 - Umweltdaten

Allgemeines:

In diesem Modul werden die für die übrigen Module notwendigen Grundlagen der Statistik behandelt.

Zur Bewertung der Güte einer Prognose gibt es eine Vielzahl an Metriken, die alle verschiedene Vor- und Nachteile bezüglich ihrer Aussagekraft haben. Im folgenden werden drei solche Methoden vorgestellt, der **RMSE** (Root Mean Square Error), der **Determinationskoeffizient** r^2 und **SMAPE** (Symmetric Mean Absolute Percentage Error).

Der **RMSE** berechnet die Wurzel des MSE (Mean Square Error), welcher das arithmetische Mittel der Fehlerquadrate bestimmt. Es sei n gleich der Anzahl der Samples in der Testmenge, $y(t)$ der wahre Messwert und $\hat{y}(t)$ der vorhergesagte Wert, dann ist der RMSE durch folgende Formel gegeben:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y(t) - \hat{y}(t))^2}{n}}$$

Der **Determinationskoeffizient** ist ein normiertes Maß und berechnet die Gesamtvariation als das Verhältnis der durch die Regression gegebenen Variation und der zu erklärenden Variation. Für die Heizenergieprognose wird das Bestimmtheitsmaß r^2 für multiple Lineare Regression verwendet. Dieses ist durch folgende Formel gegeben:

$$r^2 = \frac{[\sum_{t=1}^{n}(y(t) - \overline{y}(t)) \cdot (\hat{y}(t) - \overline{\hat{y}}(t))]^2}{[\sum_{t=1}^{n}(y(t) - \overline{y}(t))^2] \cdot [\sum_{t=1}^{n}(\hat{y}(t) - \overline{\hat{y}}(t))^2]},$$

wobei $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y(t)$ und $\bar{\hat{y}} = \frac{1}{n}\sum_{t=1}^{n} \hat{y}(t)$ ist.

Der **SMAPE** ist ein normiertes Gütemaß welches auf prozentualen Fehlern basiert und sich wie folgt berechnet:

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|y(t) - \hat{y}(t)|}{((|y(t)| + |\hat{y}(t)|)/2)}$$

Die **Varianz** ist ein stochastisches Maß für die Streuung der Wahrscheinlichkeitsdichte um ihren Schwerpunkt. Sie berechnet sich durch das Quadrat der Standardabweichung s wie folgt:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Die **Normalverteilung** (Gauß-Verteilung) ist eine stetige Wahrscheinlichkeitsverteilung mit der sich die Abweichung von Messwerten vom Erwartungswert in vielen natur-, wirtschafts- und ingenieurswissenschaftlichen Bereichen exakt oder mit einer sehr guten Näherung beschreiben lassen. Die Dichtefunktion der Normalverteilung um das Zentrum μ und der Breite σ ist gegeben durch:

$$f_N(x; \bar{x}, \sigma) = \frac{1}{\theta\sqrt{2\pi}} \cdot exp(-\frac{(x-\bar{x})^2}{2\sigma}^2$$

Im multivarianten Fall ergibt sich die Normalverteilung mit der Wahrscheinlichkeit P wie folgt:

$$p_N(x; \bar{x}, P) = \frac{1}{\sqrt{det(2\pi P)}} exp(-\frac{1}{2}(x-\bar{x})^T p^{-1}(x-\bar{x}))$$

wobei P über das Volumenintegral der Parametervektoren Overfitting berechnet wird:

$$P\{x \in V\} = \int_V p(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

Der **Korrelationskoeffizient** (auch Pearson'scher Korrelationskoeffizient) gibt für mindestens intervallskalierte Daten den linearen Zusammehang zwischen je zwei Merkmalen an. Für zwei Merkmale X und Y berechnet sich der Korrelationskoeffizient wie folgt:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Der Korrelationskoeffizient ist normiert auf das Intervall $[-1, 1]$. Dabei bedeuten Werte nahe 1 oder -1 einen starken linearen Zusammehang und Werte nahe 0 einen schwachen oder keinen linearen Zusammenhang.

Ein **Konfidenzintervall** (auch Erwartungsbereich) beschreibt in der Statistik ein Intervall in dem ein Erwartungswert mit einer bestimmten Wahrscheinlichkeit liegt. Dazu wird aus einer Stichprobe der Mittelwert $\bar{x}$ als Erwartungswert bestimmt. Dann werden anhand einer Standardnormalverteilung symmetrische Intervallgrenzen x_u und x_o bestimmt in denen die wahren Parameterwerte mit einer Wahrscheinlichkeit von $1 - \alpha$ liegen. Die Intervallgrenzen werden also wie folgt berechnet:

$$x_u = \bar{x} + z(\frac{\alpha}{2}) \cdot s_x \quad \text{und} \quad x_o = \bar{x} + z(1 - \frac{\alpha}{2}) \cdot s_x$$

Aufgaben:

- Berechnung der Korrelation zwischen Außentemperatur und Heizenergieverbrauch
- Berechnung des RMSE, r^2 und SMAPE eines Datensatzes und Interpretation der Ergebnisse
- Berechnung der Varianz eines Datensatzes, plotten der Daten und Interpretation der Ergebnisse
- Erzeugung eines Histogrammplots eines Datensatzes mit der zugehörigen Normalverteilung und Konfidenzintervall und Interpretation der Ergebnisse

Ziele:

Die Studierenden verstehen und beherrschen die notwendigen statistischen Grundlagen und Methoden um mit diesen Vorkenntnissen die weiterführenden Module bearbeiten zu können.

Literatur:

- Fahrmeir, L., Künstler, R., Piegot, I., Tutz, G. (1997). Statistik - Der Weg zur Datenanalyse. Springer Verlag, Berlin/Heidelberg/New York.