Modulkonzept zu Multipler Linearer Regression

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:

Die Multiple Lineare Regression ermöglicht die Abbildung eines linearen Einflusses mehrerer Parameter auf einen Zielwert. Damit eignet sie sich hervorragend als Methode für die Erstellung von Energieverbrauchsprognosen. In diesem Modul soll exemplarisch der Heizenergieverbrauch eines Nicht-Wohngebäudes am Umwelt Campus Birkenfeld anhand historischer Daten prognostiziert werden.

Die Features die hierbei betrachtet werden sind die Verbräuche und gemittelten täglichen Außentemperaturen der letzten vierzehn Vortage und eine errechnete Saisonkomponente sowie ein Werktagsindikator für den zu prognostizierenden Folgetag.

Die linearen Einflüsse der Features werden durch die folgende multiple lineare Modellfunktion beschrieben:

$$\begin{aligned} Q(t+1) = x(t) \cdot p = p_0 &+ p_1 \cdot Q(t) + p_2 \cdot Q(t-1) + \ldots + p_{14} \cdot Q(t-13) \\ &+ p_{15} \cdot T(t) + p_{16} \cdot T(t-1) + \ldots + p_{28} \cdot T(t-13) \\ &+ p_{29} \cdot S(t+1) \\ &+ p_{30} \cdot W(t+1), \end{aligned}$$

wobei folgende Notationen gelten:

$$\begin{aligned} p &= [\ p_0, \ldots, p_{30}\ ]^\intercal, \\ x(t) &= [\ 1\ ,\ Q(t), \ldots, Q(t-13)\ ,\ T(t), \ldots, T(t-13)\ ,\ S(t+1)\ ,\ W(t+1)\ ], \\ Q(t) &= \text{Wärmeverbrauch von Tag t}, \\ T(t) &= \text{Außentemperatur von Tag t}, \\ S(t) &= \text{Saisonkomponente von Tag t}, \\ W(t) &= \text{Werktagindikator von Tag t}. \end{aligned}$$

Da Prognosemodelle im Allgemeinen nicht lineare Einflüsse besitzen und Daten Messfehler behaftet sind, ist es nicht möglich, p so zu wählen, dass in der Modellfunktion für alle t Gleichheit gilt. Folglich handelt es sich um ein Optimierungsproblem bei dem p so gewählt wird, dass $\sum_{t=1}^{m}(Q(t+1) - x(t) \cdot p)^2$ minimal ist.

Es sei $X = \begin{bmatrix} x(1) \\ \vdots \\ x(m) \end{bmatrix} \in \mathbb{R}^{m \times 30}$ und $Q = \begin{bmatrix} Q(1+1) \\ \vdots \\ Q(m+1) \end{bmatrix} \in \mathbb{R}^m$.

Dann kann das optimale p^* durch die Moore-Penrose Pseudoinverse $X^{\#}$, die in Matlab mit dem Befehl $pinv(X)$ bestimmt werden kann, wie folgt berechnet werden:

$$p^* = X^{\#} \cdot Q.$$

Die Auswahl der Features wird in diesem Modul nicht behandelt. Stattdessen wird ein vorverarbeiteter Datensatz bereitgestellt der alle Features die verwendet werden sollen enthält.

Aufgaben:

Die Studierenden sollen anhand eines gegebenen Datensatzes und mittels multipler Regression eine Verbrauchsprognose erstellen. Die Teilschritte dabei sind:

- Laden des vorgegebenen Datensatzes
- Einteilung in Trainings- und Testdaten
- Implementierung und Training des Regressionsmodells
- Bewertung der Güte des Modells anhand der Testdaten

Ziele:

In diesem Modul sollen Studierende multiple Regression als Methode zur Modellierung linearer Einflüsse kennenlernen. Darüber hinaus sollen sie durch die praktische Durchführung ein Gefühl für die Möglichkeiten und Beschränkungen der Regression entwickeln.

Literatur:

- Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. (2012). Learning from data: a short course (S. 82-88). [United States]: AMLBook.com