Einführung in die Vorhersage-Modellierung

Prof. Dr. Christian Soost

2020-03-11

Vorhersage

Neben der erklärenden, rückwärtsgerichteten Modellierung spielt insbesondere in der Praxis die *vorhersageorientierte* Modellierung eine wichtige Rolle: Ziel ist es, bei gegebenen, neuen Beobachtungen die noch unbekannte Zielvariable y *vorherzusagen*, z.B. für neue Kunden auf Basis von soziodemographischen Daten den Kundenwert zu prognostizieren. Dies geschieht auf Basis der vorhandenen Daten der Bestandskunden, d.h. inklusive des für diese Kunden bekannten Kundenwertes (Supervised Learning).

Es werden zwei Teildatenmengen unterschieden: Zum einen gibt es die Trainingsdaten (auch Lerndaten genannt), die aus einer Lern- oder Schätzstichprobe stammen, und zum anderen gibt es Anwendungsdaten, auf die man das Modell anwendet.

1. Bei den Trainingsdaten liegen sowohl die erklärenden Variablen $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ als auch die Zielvariable y vor. Auf diesen Trainingsdaten wird das Modell $y = f(\mathbf{x}) + \epsilon = f(x_1, x_2, \ldots, x_n) + \epsilon$ gebildet und durch $\hat{f}(\cdot)$ geschätzt. Muss keine Regression sein
2. Dieses geschätzte Modell ($\hat{f}(\cdot)$) wird auf die Anwendungsdaten $\mathbf{x_0}$, für die (zunächst) die Zielvariable unbekannt ist, angewendet, d.h., es wird $\hat{y}_0 := \hat{f}(\mathbf{x}_0)$ berechnet. Der unbekannte Wert y_0 der Zielvariable y wird durch $\hat{y}_0$ prognostiziert. Prof prüft dann ob gute Variable

Eventuell liegt zu einem noch späteren Zeitpunkt der eingetroffene Wert y_0 der Zielvariable y vor. Dann kann die eigene Vorhersage $\hat{y}_0$ evaluiert werden, d.h. z.B. kann der Fehler $y_0 - \hat{y}_0$ zwischen prognostiziertem Wert $\hat{y}_0$ und wahrem Wert y_0 analysiert werden. Residuen

In der praktischen Anwendung können zeitlich drei aufeinanderfolgende Abschnitte unterschieden werden (vergleiche oben):

1. die Trainingsphase, d.h., die Phase für die sowohl erklärende ($\mathbf{x}$) als auch die erklärte Variable (y) bekannt sind. Hier wird das Modell geschätzt (gelernt): $\hat{f}(\mathbf{x})$. Schätzung Modell was die Zielvariable erklärt
2. In der folgenden Anwendungsphase sind nur die erklärenden Variablen ($\mathbf{x_0}$) bekannt, nicht y_0. Auf Basis der Ergebnisses aus 1. wird $\hat{y}_0 := \hat{f}(\mathbf{x}_0)$ prognostiziert.
3. Evt. gibt es später noch die Evaluierungsphase, für die dann auch die Zielvariable (y_0) bekannt ist, so dass die Vorhersagegüte des Modells überprüft werden kann.

Im Computer kann man dieses Anwendungsszenario *simulieren*: man teilt die Datenmenge *zufällig* in eine Lern- bzw. Trainingsstichprobe (Trainingsdaten; $(\mathbf{x}, \mathbf{y})$) und eine Teststichprobe (Anwendungsdaten, $(\mathbf{x_0})$) auf: Die Modellierung erfolgt auf den Trainingsdaten. Das Modell wird angewendet auf die Testdaten (Anwendungsdaten). Da man hier aber auch die Zielvariable (y_0) kennt, kann damit das Modell evaluiert werden.

Wettbewerb

Ihre Aufgabe ist: Spielen Sie den Data-Scientist. Konstruieren Sie ein Modell auf Basis der Trainingsdaten $(\mathbf{x}, \mathbf{y})$) und sagen Sie für die Anwendungsdaten ($(\mathbf{x_0})$) die Zielvariable voraus ($\hat{y}_0$).

Ihr Dozent kennt den Wert der Zielvariable (y_0). Zur Bewertung der Vorhersagegüte wird der mittlere absolute Fehler MAE (**m**ean **a**bsolute **e**rror) auf die Anwendungsdaten herangezogen:

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

nicht r^2, nur an den Differenzen zw. geschätzen und echten Werten interessiert

Dabei sind y_i die wahren Werte, $\hat{y}_i$ die prognostizierten Werte des geschätzten Modells $\hat{f}(\cdot)$ und n_{Test} die Anzahl der Beobachtungen des Testdatensatzes (Anwendungsdatensatz). Für eine gute Prognose sollte daher MAE_{Test} möglichst klein sein.

Hinweise

Sie haben relativ freie Methodenwahl bei der Modellierung und Vorverarbeitung: Sie können z.B. eine lineare Regression mit Variablen Ihrer Wahl rechnen; Sie können aber auch Baumverfahren oder Neuronale Netze anwenden.

Eine gute Einführung in verschiedene Methoden gibt es z.B. bei Sebastian Sauer (2019): *Moderne Datenanalyse mit R.* https://link.springer.com/book/10.1007/978-3-658-21587-3 aber auch bei Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning – with Applications in R*, http://www-bcf.usc.edu/~gareth/ISL/. Die Bücher beinhalten jeweils Beispiele und Anwendung mit R.

Auch ist es Ihnen überlassen, welche Variablen Sie zur Modellierung heranziehen – und ob Sie diese eventuell vorverarbeiten, d.h., transformieren, zusammenfassen, Ausreißer bereinigen o.ä.. Denken Sie nur daran, die Datentransformation, die Sie auf den Trainingsdaten durchführen, auch auf den Testdaten (Anwendungsdaten) durchzuführen.

Hinweise zur Modellwahl usw. gibt es auch in erwähnter Literatur, aber auch in vielen Büchern zum Thema Data-Mining/Data-Science.

Alles was Sie tun, Datenvorverarbeitung, Modellierung und Anwenden, muss transparent und reproduzierbar sein. Ansonsten lautet die Aufgabe: Finden Sie ein Modell, von dem Sie glauben, das es gut vorhersagt. $\hat{y} = 42$ tut es leider oft nicht. Eine gute Modellierung auf den Trainingsdaten (z.B. hohes R^2) bedeutet nicht zwangsläufig eine gute Vorhersage.

Tipps für eine gute Prognose

- Vermeiden Sie Über-Anpassung.
- Evtl. kann eine Datenvorverarbeitung (Variablentransformation, z.B. log() oder die Elimination von Ausreißern) helfen.
- Überlegen Sie sich Kriterien zur Modell- und/ oder Variablenauswahl.
- Schauen Sie in die Literatur.

Bewertung

Gruppenarbeiten mit bis zu 3 Personen sind möglich. In die Bewertung fließen u.a. ein:

- Methode: methodischer Anspruch und Korrektheit in der Explorativen Datenanalyse, Datenvorverarbeitung, Variablenauswahl und Modellierungsmethode.
- Inhalt: inhaltliche Korrektheit in Beschreibung und Interpretation.
- Vorhersagegüte: Die Vorhersagegüte des Nullmodelles entspricht einer $4,0$, die eines (unbekannten) einfachen Referenzmodells Ihres Dozenten einer $2,0$. Ihre Bewertung erfolgt entsprechend Ihrer Vorhersagegüte, d.h., sind Sie besser als das Referenzmodell erhalten Sie hier in diesem Teilaspekt eine bessere Note als $2,0$!

Die quantitative Datenanalyse in Durchführung und Interpretation ist der Schwerpunkt dieser Arbeit. Identisches Vorgehen, z.B. im R Code, ist zufällig sehr unwahrscheinlich und kann als **Plagiat** bewertet werden.

Falls Sie hypothesengesteuert vorgehen: Achten Sie auf die korrekte Formulierung der Null- und Alternativhypothesen, sowie auf die richtige Interpretation des Testergebnisses.

Nicht reproduzierbare Auswertungen sowie fehlende Abgaben der Prognosedatei führen zur Abwertung.

Organisatorisches

Der Schwerpunkt dieser Hausarbeit liegt auf der quantitativen Modellierung, der formale Anspruch, aber auch der Anspruch in Bezug auf Literatur etc., liegen daher unter dem von anderen Hausarbeiten. Um eine komplett transparente und reproduzierbare Analyse zu ermöglichen, muss das beigefügte R Markdown Template verwendet werden (`Template-Vorhersagemodellierung.Rmd`). Dies kann dann in eine Word Datei überführt werden (`knit`). Das pdf dieser Datei kann dann im OC hochgeladen werden. Ein ausgedrucktes Exemplar muss nicht abgegeben werden.

Senden Sie die csv Datei Ihrer Prognose (`Prognose_IhrName.csv`) als Anhang einer Email bis zum 05.07.2020, 23:59 Uhr an christian.soost@fom.de. Betreff der Email: "WMQD Do: Vorhersagewettbewerb".

Im (vorläufigen!) Zeitplan ist vorgesehen, dass Sie innerhalb eines Präsenztermins an Ihrer Hausarbeit arbeiten können.

Checkliste

- Haben Sie eine Vorhersage für die 100 Anwendungsdaten erzeugt und als csv Datei exportiert: `Prognose_IhrName.csv` (Ihr Name entsprechend angepasst)?
- Haben Sie die Vorhersage per Email versendet?
- Läuft die Rmd Datei beim knitten durch?
- Haben Sie das pdf Ihrer Auswertung hochgeladen?

Datenbeschreibung

Es liegen Daten der Social Media Abteilung eines Unternehmens vor, Zielvariable y ist die Anzahl der Klicks (`klicks`) auf einem Post in einem Sozialen Netzwerk des Unternehmens.

Als (potentiell) erklärende Variablen liegen folgende Daten des Posts vor:

- `wochentag`: Wochentag des Posts, beginnend mit Montag.
- `stunde`: Uhrzeit (volle Stunde des Posts)
- `likes`: Anzahl der Likes der Seite des Unternehmens.
- `typ`: Inhalt des Posts.
- `kategorie`: Charakterisierung des Inhalts des Posts.
- `bezahlt`: Bezahlte Werbung auf der Plattform.

Der Datensatz `Trainingsdaten.csv` enthält die Zielvariable (`klicks`), anhand dieser Daten können Sie Ihr Modell entwickeln, angewendet wird es auf den Datensatz `Anwendungsdaten.csv`. Dieser enthält die Zielvariable nicht. Die Aufteilung erfolgte zufällig. Erstellen Sie auf Basis der Beobachtungen `Trainingsdaten.csv` ein Modell für die Anzahl Klicks, `klicks`. Wenden Sie Ihr Modell auf die Beobachtungen aus `Anwendungsdaten.csv` an und erstellen Sie so für diese Beobachtungen eine Prognose für die Anzahl Klicks.

Exportieren Sie Ihre Prognose für `klicks` ebenfalls als `csv` Datei (`Vorhersage_IhrName.csv`, siehe Template `Template-Vorhersagemodellierung.Rmd`).