Hinweise zur Prüfungsleistung 'Vorhersagemodellierung'

Sebastian Sauer

2021-02-17

Inhaltsverzeichnis

1 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 unbekannten Werte der Zielvariablen y *vorherzusagen*, z.B. für neue Kunden auf Basis von soziodemographischen Daten den *Kundenwert* – möglichst genau – zu prognostizieren. Dies geschieht auf Basis der vorhandenen Daten der Bestandskunden, d.h. inklusive des für diese Kunden bekannten Kundenwertes.

Ihnen werden *zwei Teildatenmengen* zur Verfügung gestellt: Zum einen gibt es die Trainingsdaten (auch *Lerndaten* genannt) und zum anderen gibt es Anwendungsdaten (auch *Testdaten* genannt), 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.

2. Dieses geschätzte Modell ($\hat{f}(\cdot)$) wird auf die Anwendungsdaten $\mathbf{x_0}$, für die (Ihnen) die Werte der Zielvariable unbekannt sind, 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.

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

In der praktischen Anwendung können zeitlich drei aufeinanderfolgende Schritte 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): $\hat{f}(\mathbf{x})$. Dafür wird der Trainingsdatensatz genutzt.
2. In der folgenden *Anwendungsphase* sind nur die erklärenden Variablen ($\mathbf{x_0}$) bekannt, nicht y_0. Auf Basis der Ergebnisses aus dem 1. Schritt 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.

2 Koeffizient zur Güte der Vorhersagegüte

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 möglichst genau voraus ($\hat{y}_0$).

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

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

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.

3 Formale Hinweise

1. Es sind nur Einzelarbeiten zulässig.
2. In der Analyse muss als Ausgangspunkt der vom/von der Dozenten/in bereitgestellten Datensatz genutzt werden. Dazu muss der Datensatz zu Beginn des Skripts von der entsprechenden Internetadresse heruntergeladen werden.
3. Alle Analyseschritte bzw. alle Veränderungen an den Daten müssen im (eingereichten) *Analyseskript* nachvollziehbar (transparent und reproduzierbar) aufgeführt sein. Das Analyseskript ist als R-Skript, Rmd-Datei oder Rmd-Notebook-Datei abzugeben. Sie können die bereitgestellte Vorlage als Analyseskript nutzen (Template-Dokumentation-Vorhersagemodellierung.Rmd).

4. Es dürfen keine weiteren Informationen (Daten) als die vom Dozenten ausgegebenen verwendet werden. Sonstige Hilfe (z.B. von Dritten) ist ebenfalls unzulässig.

5. Nichtbeachtung der für dieses Modul formulierten Regeln kann zu Nichtbestehen oder Punkteabzug führen.

4 Zum Aufbau Ihrer Prognosedatei im CSV-Format

1. Die CSV-Datei muss aus zwei Spalten mit (exakt) folgenden Spaltennamen bestehen:

a) `id`: Den ID-Wert jedes vorhergesagten Wertes
b) `pred`: Der vorhergesagte Wert.

2. Der Name der Datei muss wie folgt lauten: `Nachname_Vorname_Matrikelnummer_Prognose.csv`. Beispiel: `Sauer_Sebastian_0123456_Prognose.csv`.

3. Umlaute sind zu ersetzen (also `Süß` wird `Suess` etc.).

4. Die CSV-Datei muss als *Spaltentrennzeichen* ein *Komma* verwenden und als *Dezimaltrennzeichen* einen *Punkt* (d.h. also die *Standardformatierung* einer CSV-Datei; *nicht* die deutsche Formatierung).

5. Die CSV-Datei muss genau die Anzahl an Zeilen aufweisen, die der Zeilenlänge im Test-Datensatz entspricht.

5 Einzureichende Dateien

1. Folgende Dateien sind einzureichen:

    1. Ihre Prognose-Datei (CSV-Datei)
    2. Ihr Analyseskript (R-, Rmd- oder Rmd-Notebook-Datei)

2. Weitere Dateien sind nicht einzureichen.

3. Komprimieren Sie die Dateien *nicht* (z.B. via *zip*).

4. Prüfen Sie, dass Ihre CSV-Datei sich problemlos lesen lässt. Falls keine (funktionstüchtige) CSV-Datei eingereicht (hochgeladen) wurde, ist die Prüfung nicht bestanden. Tipp: Öffnen Sie die CSV-Datei mit einem Texteditor und schauen Sie sich an, ob alles vernünftig aussieht. Achtung: Öffnen Sie die CSV-Datei besser nicht mit Excel, da Excel einen Bug hat, der CSV-Dateien verfälschen kann auch ohne dass man die Datei speichert.

6 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*[1] aber auch bei Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning – with Applications in R*[2]. 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

[1] https://link.springer.com/book/10.1007/978-3-658-21587-3
[2] http://www-bcf.usc.edu/~gareth/ISL/

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-Science.

Alles, was Sie tun, Datenvorverarbeitung, Modellierung und Anwenden, muss transparent und reproduzierbar sein. Im Übrigen lautet die Aufgabe: Finden Sie ein Modell, von dem Sie glauben, das es die Testdaten 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.

7 Tipps

7.1 Tipps für eine gute Prognose

- **Schauen Sie in die Literatur.**
- Evtl. kann eine Datenvorverarbeitung (Variablentransformation, z.B. log(), oder die Elimination von Ausreißern) helfen.
- Überlegen Sie sich Kriterien zur Modell- und/ oder Variablenauswahl. Auch hierfür gibt es Algorithmen und R-Funktionen.
- Vermeiden Sie Über-Anpassung (Overfitting).

7.2 Tipps zur Datenverarbeitung

- Ein "deutsches" Excel kann Standard-CSV-Dateien nicht ohne Weiteres lesen. Online-Dienste wie Google Sheets können dies allerdings.

8 Bewertung

- Es gibt drei Bewertungskriterien:
  - *Formalia*: u.a. Reproduzierbarkeit der Analyse, Lesbarkeit der Syntax, Übersichtlichkeit der Analyse.
  - *Methode*: u.a. methodischer Anspruch und Korrektheit in der Explorativen Datenanalyse, Datenvorverarbeitung, Variablenauswahl und Modellierungsmethode.
  - *Inhalt*: **Vorhersagegüte**.
- Das zentrale Bewertungskriterium ist *Inhalt*; die übrigen beiden Kriterien fließn nur bei besonders guter oder schlechter Leistung in die Gesamtnote ein.
- Zur Vorhersagegüte: Die Vorhersagegüte des Nullmodells entspricht einer $4,0$, die eines (unbekannten) einfachen Referenzmodells Ihres/Ihrer Dozent*in 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. Zufälliges identisches Vorgehen, z.B. im R Code, ist sehr unwahrscheinlich und kann als **Plagiat** bewertet werden.
- Die Gesamtnote muss sich nicht als arithmetischer Mittelwert der Teilnoten ergeben.
- Es werden keine Teilnoten vergeben, sondern nur eine Gesamtnote.

9 Formalia

- Der Schwerpunkt dieser Hausarbeit liegt auf der quantitativen Modellierung, der formale Anspruch liegt daher unter dem von anderen Hausarbeiten.
- Es muss keine Literatur zitiert werden.
- Ein ausgedrucktes Exemplar muss nicht abgegeben werden.
- Der Datensatz `Trainigsdaten.csv` enthält die Zielvariable (y, `pay`), anhand dieser Daten können Sie Ihr Modell entwickeln ("trainieren"), getestet wird es im Test-Datensatz `Anwendungsdaten.csv`. Dieser enthält die Zielvariable *nicht*. Die Aufteilung erfolgte zufällig. Erstellen Sie auf Basis der Beobachtungen ein Modell zur Vorhersage der Zielvariablen. Wenden Sie Ihr Modell auf die Beobachtungen aus an und erstellen Sie so für diese Beobachtungen eine Prognose für das Gehalt.

10 Wo finde ich Beispiele?

Eine Beispiel-Modellierung finden Sie in der Datei `Beispielanalyse-Prognose-Wettbewerb.Rmd`. Eine beispielhafte Vorlage (Template), die Sie als Richtschnur nutzen können, ist mit der Datei `Template-Vorhersagemodellierung.Rmd` bereitgestellt.