Hinweise zur Hausarbeit - Vorhersagemodellierung

FOM-Data-Literacy-Team

2020-09-01

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 – möglichst genau – 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.
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.

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.

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): $\hat{f}(\mathbf{x})$.
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.

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 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.

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 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. 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.

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

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.

Bewertung

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*: u.a. inhaltliche Korrektheit in Beschreibung und Interpretation sowie **Vorhersagegüte**.
- Vorhersagegüte: Die Vorhersagegüte des Nullmodelles entspricht einer $4,0$, die eines (unbekannten) einfachen Referenzmodells Ihr(e)r 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. 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.
- Die Gesamtnote muss sich nicht als arithmetischer Mittelwert der Teilnoten ergeben. Einzelne besonders gute oder schwache Aspekte in den Teilnoten können die Gesamtnote ggf. substanziell nach oben oder unten beeinflussen.

Formalia

- Es sind sowohl Gruppenarbeiten möglich (max. 3 Personen) als auch Einzelarbeiten. Bei Gruppenarbeiten ist von allen Mitgliedern das exakt gleiche Thema sowie einen (gleichen) Gruppennamen anzumelden. Jede Person erhält eine individuelle Note.
- 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`).
- Ein ausgedrucktes Exemplar muss nicht abgegeben werden.
- Einzureichen (hochzuladen) sind a) Ihre Vorhersagen für die Zielvariable im Anwendungs-Datensatz (als Standard-CSV-Datei) sowie b) Ihre Auswertung (auf Basis des R-Markdown-Templates `Template-Vorhersagemodellierung.Rmd`). Ggf. können Sie Ihre Dateien als ZIP-Archiv hochladen. Prüfen Sie, dass Ihre CSV-Datei sich problemlos lesen lässt.
- Falls keine (funktionstüchtige) CSV-Datei eingereicht (hochgeladen) wurde, ist die Prüfung nicht bestanden.
- Bei Gruppenleistungen ist anzugeben, welche Person welchen Abschnitt bearbeitet hat.
- Nichtbeachtung dieser Regeln kann zu Nichtbestehen oder Punkteabzug führen.

Datenbeschreibung

Im Datensatz werden Prädiktoren des Gehalts von Angestellten und insbesondere Prädiktoren zum Gender-Pay-Gap untersucht, d.h. zum Unterschied der Bezahlung zwischen Frauen und Männern bei gleichwertiger Tätigkeit und Qualifikaton.

Es handelt sich nicht um eine Studie mit Erklärungs- bzw. Kausalanspruch, sondern die Güte (Genauigkeit) der Prognose des Gehalts steht im Mittelpunkt.

Zielvariable: `pay`

Prädiktorvariablen

- `jobTitle`: Berufsbezeichnung
- `gender`: Geschlecht
- `age`: Alter
- `perEval`: Wert in der letzten Leistungsbeurteilung durch die Führungskraft (höhere Werte sind besser)
- `dept`: Abteilung
- `seniority`: Dauer der Zugehörigkeit (Seniorität)

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.

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.

Checkliste

- Haben Sie eine Vorhersage für die 300 Anwendungsdaten erzeugt und als csv Datei exportiert: `Prognose_IhrName.csv` (Ihr Name entsprechend angepasst)?
- Entspricht diese in der Struktur dem Beispiel `Vorhersage_Zufall.csv`?
- Haben Sie die Vorhersage im OC eingereicht?
- Bei Gruppenarbeiten: Sind die individuellen Kapitelzuordnungen erkennbar?
- Läuft die Rmd Datei beim knitten durch?
- Haben Sie das pdf Ihrer Auswertung hochgeladen?