Klassifikation

5

Bei der Klassifikation werden Beobachtungen anhand von Eigenschaften in vorher festgelegte Klassen eingeteilt. Wir beschränken uns auf nur zwei Klassen und sprechen dann von „binärer" Klassifikation. Die beiden Klassen werden häufig „positiv" und „negativ" genannt.

Klassifikation wird in der Praxis sehr häufig verwendet. Zum Beispiel: Ist ein Patient mit gewissen diagnostischen Werten krank oder gesund? Oder: Wird ein Kunde mit bekanntem Kaufverhalten ein neues Produkt kaufen oder nicht?

Die logistische Regression kann zur binären Klassifikation verwendet werden: Sie modelliert die Wahrscheinlichkeit zu einer von zwei Klassen (z. B. „positiv") zu gehören. Um klassifizieren zu können, müssen wir zudem noch eine Grenze für die Wahrscheinlichkeit festlegen, z. B. 50 %. Alle Beobachtungen mit einer Wahrscheinlichkeit von 50 % oder mehr werden der einen Klasse („positiv") und alle Beobachtungen mit einer Wahrscheinlichkeit von unter 50 % werden der anderen Klasse („negativ") zugeordnet (je nach Anwendungszweck kann auch eine andere Grenze besser geeignet sein).

Beispiel: Spende (Fortsetzung): Klassifikation

Das angepasste logistische Regressionsmodell modelliert die Wahrscheinlichkeit für Spendebereitschaft „ja" und wir legen willkürlich fest, dass diese Klasse die „positive" Klasse ist. Falls diese Wahrscheinlichkeit 50 % oder mehr ist, wird die Person als „Spender" („positiv") klassifiziert. Ansonsten wird sie als „Kein Spender" („negativ") klassifiziert. Konkret: Sollte gemäß unserem Modell eine 25-jährige Person eher als „Spender", also „positiv", oder als „Kein Spender", also „negativ", klassifiziert werden?

M. Kalisch und L. Meier, *Logistische Regression,* essentials,
https://doi.org/10.1007/978-3-658-34225-8_5

Zunächst berechnen wir die Wahrscheinlichkeit für Spendebereitschaft:

```
datNew <- data.frame(alter = 25)
predict(fit.spende, newdata = datNew, type = "response")
##         1
## 0.1596947
```

Die Wahrscheinlichkeit für Spendebereitschaft ist gemäß unserem Modell etwa 16 %, also kleiner als die Grenze von 50 %. D. h., wir klassifizieren diese Person als „Kein Spender“ bzw. „negativ“. ◄

Die Daten, mit denen das Modell angepasst bzw. „trainiert“ wurde, werden auch **Trainingsdaten** genannt. Entscheidend für die Anwendung ist häufig die Frage, wie gut die Methode funktioniert, um die Klasse bei *neuen* Daten vorherzusagen. Zum Beispiel im Klinikalltag, bei einem neuen Patienten, dessen diagnostische Werte man kennt: Ist er gesund oder krank?

Um das einschätzen zu können, kann man einen zweiten Datensatz verwenden, der zur Modellanpassung bisher *nicht* verwendet wurde, also „neu“ ist. Man spricht von sogenannten **Testdaten.** Alternativ kann Kreuzvalidierung verwendet werden: Es werden dann die vorhandenen Daten (typischerweise mehrmals) in Trainings- und Testdaten aufgeteilt. Wir verfolgen dies hier aber nicht weiter.

Wir klassifizieren nun jede Beobachtung im Testdatensatz mit unserer Klassifikationsmethode. Wenn sie gut funktioniert, sollten praktisch alle Beobachtungen richtig klassifiziert werden. Um das Ergebnis übersichtlich darzustellen, wird häufig auch eine Tabelle mit den wahren Klassen als Spalten und den vorhergesagten Klassen als Zeilen angegeben (die sogenannte **confusion matrix**). Die möglichen Ausgänge sind in Tab. 5.1 dargestellt. Wir verwenden jeweils gerade die entsprechenden englischen Bezeichnungen. Wenn also z. B. bei einer Beobachtung, die in der Tat zur Kategorie „negativ“ gehört, die Vorhersage „positiv“ gemacht wird, dann spricht man von einem „false positive“.

Tab. 5.1 Schematische Darstellung einer confusion matrix

		Wahrheit	
		negativ	positiv
Vorhersage	negativ	true negative (TN)	false negative (FN)
	positiv	false positive (FP)	true positive (TP)

Beispiel: Spende (Fortsetzung): Confusion Matrix und Fehlerrate

Unsere Klassifikationsmethode wurde mit den Trainingsdaten im data frame `spende` trainiert. Wie gut würde dieser Klassifikator die Spendebereitschaft von *neuen* Personen vorhersagen? Um das herauszufinden, verwenden wir einen Testdatensatz: Im data frame `spende.test` sind 1000 *weitere* Personen zu Alter und Spendebereitschaft befragt worden. Für jede Person machen wir nun basierend auf ihrem Alter eine Vorhersage bezüglich Spendebereitschaft und vergleichen dann mit der wahren Spendebereitschaft, die ja in `spende.test` verfügbar ist.

```
## Berechne Wahrscheinlichkeit
p.pred <- predict(fit.spende, newdata = spende.test,
                  type = "response")
## Leite aus Wahrscheinlichkeit die Klasse ab
vorhersage <- factor(ifelse(p.pred >= 0.5, "ja", "nein"),
                     levels = c("nein", "ja"))
wahrheit <- spende.test$antwort

## Tabelliere Ergebnis: confusion matrix
table(vorhersage, wahrheit)
##           wahrheit
## vorhersage nein  ja
##       nein  443 194
##       ja    140 223
```

In der Tabelle (entspricht der confusion matrix) ist die wahre Spendebereitschaft in den Spalten und die vorhergesagte Spendebereitschaft in den Zeilen zu sehen. In der ersten Spalte sehen wir $443 + 140 = 583$ Personen, die in Wahrheit keine Spendebereitschaft hatten („Spendebereitschaft nein“): 443 Personen wurden in die richtige Klasse „Spendebereitschaft nein“ eingeteilt, während die übrigen 140 Personen fälschlicherweise in die Klasse „Spendebereitschaft ja“ eingeteilt wurden.

Analog sehen wir in der zweiten Spalte $194 + 223 = 417$ Personen, die in Wahrheit zu einer Spende bereit sind. Davon hat unsere Klassifikationsmethode aber nur 223 korrekterweise in die Klasse „Spendebereitschaft ja“ eingeteilt. Die übrigen 194 Personen wurden fälschlicherweise in die Klasse „Spendebereitschaft nein“ eingeteilt.

Zusammenfassend hat unsere Klassifikationsmethode also bei $140 + 194 = 334$ Personen (von insgesamt 1000) einen Fehler gemacht. Die sogenannte **Fehlerrate** (oder: **misclassification error**) auf diesem Testdatensatz ist also $\frac{334}{1000} = 33.4\,\%$. ◀

Übliche Gütezahlen für einen Klassifikator sind die **True Positive Rate** (TPR),

$$\text{TPR} = \frac{\text{Anzahl true positives}}{\text{Anzahl Beob., die in Wahrheit positiv sind}} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}},$$

wobei wir mit dem Symbol „#“ das Wort „Anzahl“ abkürzen. Die TPR gibt uns also an, wieviel Prozent der in der Tat positiven Beobachtungen wir korrekt vorhersagen können.

Umgekehrt ist die **False Positive Rate** (FPR) gegeben durch

$$\text{FPR} = \frac{\text{Anzahl false positives}}{\text{Anzahl Beob., die in Wahrheit negativ sind}} = \frac{\#\text{FP}}{\#\text{FP} + \#\text{TN}}.$$

Sie entspricht dem Anteil „positiv“ klassifizierter Beobachtungen unter allen Beobachtungen, die in Wahrheit „negativ“sind. Wünschenswert ist also eine große TPR und eine kleine FPR. Ein perfekter Klassifikator hat $\text{TPR} = 1$ („wir erwischen alle in der Tat positiven Fälle“) und $\text{FPR} = 0$ („wir machen nie den Fehler, dass wir eine in der Tat negative Beobachtung als positiv vorhersagen“).

Im medizinischen Bereich werden alternativ auch die Begriffe **Sensitivität** (= TPR) und **Spezifität** ($= 1 - \text{FPR}$) verwendet.

Beispiel: Spende (Fortsetzung): TPR und FPR

Insgesamt gibt es 417 „positive“ Beobachtungen (also Personen mit Spendebereitschaft). Davon wurden 223 Beobachtungen richtigerweise in die „positive“ Klasse („Spender“) eingeteilt. Für die True Positive Rate gilt also:

$$\text{TPR} = \frac{223}{417} \approx 0.53$$

Umgekehrt gab es 583 in der Tat „negative“ Beobachtungen (Personen ohne Spendebereitschaft). Davon wurden 140 Beobachtungen fälschlicherweise in die „positive“ Klasse („Spender“) eingeteilt. Für die False Positive Rate gilt also:

$$\text{FPR} = \frac{140}{583} \approx 0.24$$

Die Sensitivität ist also 0.53 und die Spezifität $1 - 0.24 = 0.76$. ◄

Bei unserer Klassifikationsmethode haben wir die Grenze für die Wahrscheinlichkeit, den sogenannten „cutoff“, bei 50 % angesetzt: Alle Beobachtungen mit einer

Wahrscheinlichkeit von 50 % oder mehr werden der „positiven“ Klasse und alle Beobachtungen mit einer Wahrscheinlichkeit von unter 50 % werden der „negativen“ Klasse zugeordnet. Daraus hat sich eine gewisse TPR und FPR ergeben.

Wenn wir diese Grenze verschieben, ändern sich die Vorhersagen und somit auch die TPR bzw. FPR. Wenn die Grenze z. B. 0 % ist, werden alle Beobachtungen in die Klasse „positiv“ eingeteilt. D. h., alle Beobachtungen, die in Wahrheit „positiv“ sind, werden korrekterweise als „positiv“ klassifiziert. Somit gilt TPR $= 1$. Allerdings werden auch alle in Wahrheit „negativen“ Beobachtungen (fälschlicherweise) als „positiv“ klassifiziert. Daher gilt FPR $= 1$.

Wenn wir diese Grenze für die Wahrscheinlichkeit erhöhen, ändert sich die Einteilung bei mehr und mehr Personen von „positiv“ zu „negativ“. Dadurch nehmen sowohl TPR als auch FPR ab. Wenn die Grenze schliesslich 100 % ist, wird jede Person in die Klasse „negativ“ eingeteilt. Damit gilt sowohl TPR $= 0$ als auch FPR $= 0$.

Je nach „cutoff“ ergibt sich also ein anderer Kompromiss zwischen (möglichst großer) TPR und (möglichst kleiner) FPR. Die **ROC-Kurve** (ROC steht für „Receiver Operating Characteristic“) visualisiert alle möglichen Kombinationen von TPR und FPR, die durch eine Einstellung des „cutoffs“ erzielt werden können: Auf der horizontalen Achse wird die FPR und auf der vertikalen Achse die TPR aufgetragen. Nun wird für jeden denkbaren Wert des „cutoffs“ ein Punkt bei der entsprechenden TPR und FPR eingezeichnet. Daraus ergibt sich eine Kurve, die links unten bei TPR $= 0$ und FPR $= 0$ (entspricht einem „cutoff“ von 100 %) beginnt und bis rechts oben bei TPR $= 1$ und FPR $= 1$ (entspricht einem „cutoff“ von 0 %) monoton ansteigt. D.h., wenn man den „cutoff“ von 0 % schrittweise auf 100 % erhöht, dann wird die Kurve von rechts oben nach links unten durchlaufen.

Entscheidend für die Güte des Klassifikators ist die *Art* des Anstiegs. Bei einem Klassifikator, der auf bloßem Raten basiert, entspricht die erwartete ROC-Kurve gerade der Winkelhalbierenden. Im Gegensatz dazu würde ein perfekter Klassifikator zunächst vertikal bis TPR $= 1$ ansteigen und dann horizontal bis FPR $= 1$ verlaufen. In der Praxis wird die ROC-Kurve meist irgendwo dazwischen liegen. Grundsätzlich ist ein Klassifikator mit einer größeren Fläche unter der ROC-Kurve („area under the curve“ oder kurz **AUC**) besser. Bei bloßem Raten erwartet man AUC $= 0.5$ und bei einem perfekten Klassifikator ist AUC $= 1$.

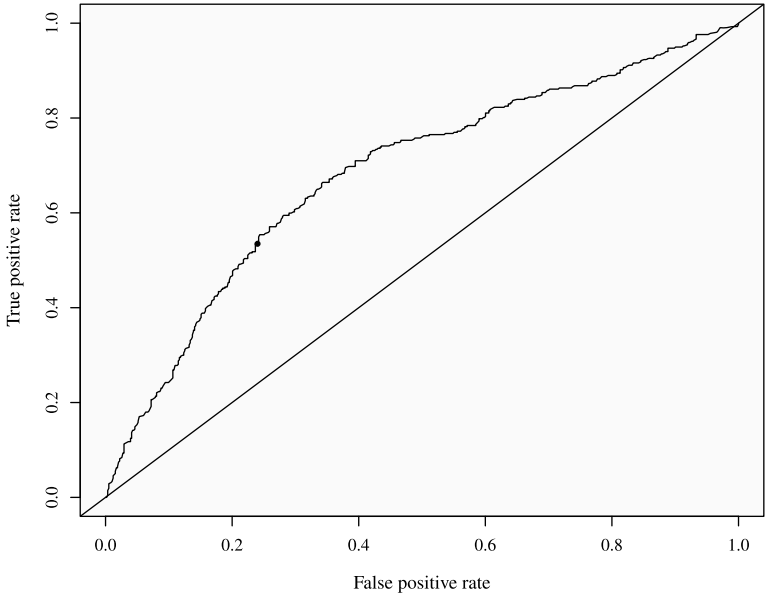
Die ROC-Kurve kann helfen, einen guten „cutoff“ zu finden. Hier gibt es keine eindeutige Regel, allerdings sollte die TPR möglichst groß und die FPR möglichst klein sein. D. h., wir suchen auf der ROC-Kurve einen Punkt, der möglichst weit „links oben“ liegt. Weitere Informationen zur Analyse einer ROC-Kurve findet man z. B. in Fawcett (2006).

In R kann die ROC-Kurve z. B. mit dem Paket ROCR (Sing et al. 2005) oder pROC (Robin et al. 2011) erzeugt werden.

Beispiel: Spende (Fortsetzung): ROC-Kurve

Wir verwenden das Paket ROCR um die ROC-Kurve des angepassten logistischen Regressionsmodells zu berechnen.

```
library(ROCR)
## Wahrscheinlichkeiten gemäß logistischem Regressionsmodell
pred.test <- predict(fit.spende, newdata = spende.test,
                     type = "response")
## Erstelle prediction-Objekt für ROCR
pred <- prediction(pred.test, spende.test$antwort,
                   label.ordering = c("nein", "ja"))
perf <- performance(pred, "tpr", "fpr")
plot(perf)
points(x = 140 / 583, y = 223 / 417, pch = 20) ## cutoff 0.5
abline(a = 0, b = 1) ## Winkelhalbierende
```



Der gewählte cutoff von 0.5 (schwarzer Punkt) scheint ein vernünftiger Kompromiss zwischen großer TPR und kleiner FPR zu sein.

Die AUC ist etwa 0.68, also größer als der Wert von 0.5, den wir mit bloßem Raten erwarten würden.

```
auc <- performance(pred, "auc")
auc@y.values
## [[1]]
## [1] 0.6834471
```

◀

Um mehrere Klassifikationsmethoden miteinander zu vergleichen, werden häufig die entsprechenden ROC-Kurven in einem Bild gezeigt. Die zugehörigen AUC-Werte können zudem mit statistischen Tests miteinander verglichen werden, zum Beispiel mit der Funktion `roc.test` im Paket `pROC`.