KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Prognosemodelle (Klassifikation und Regression):
02_02Evaluation_Metriken_02

Metriken für eine Klassifikation

Erarbeitet von
Dr. Katja Theune

Lernziele

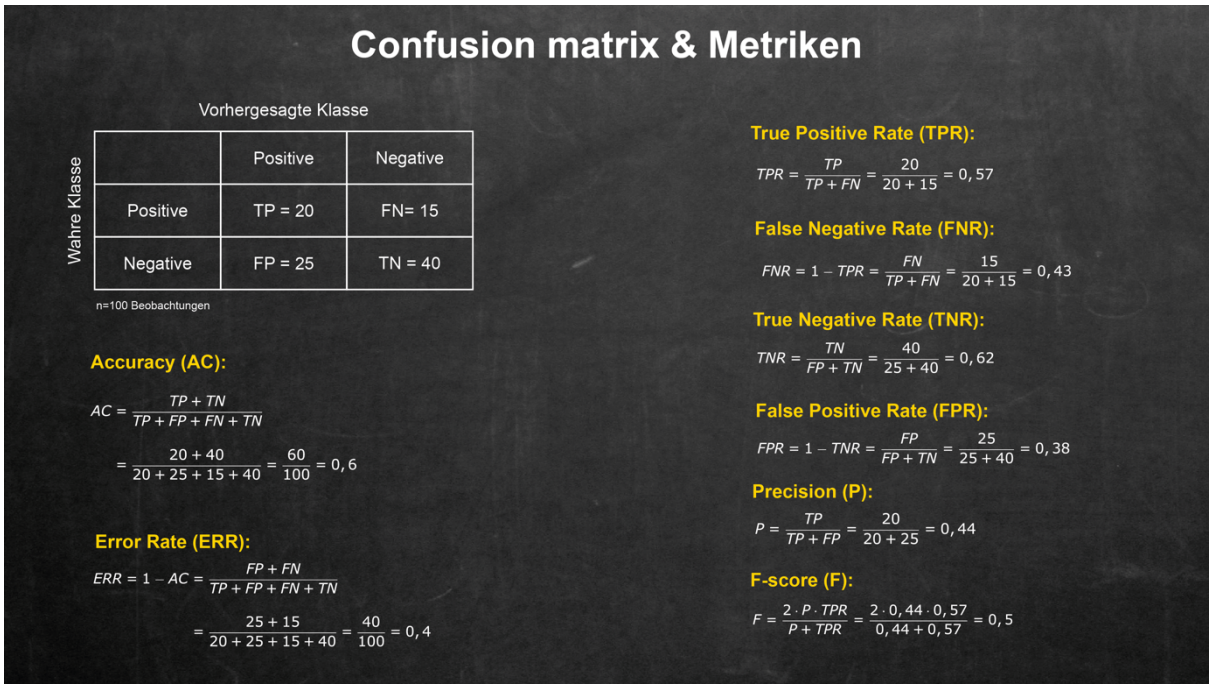
- Du kannst verschiedene Metriken für eine Evaluation erläutern
- Du kannst verschiedene Metriken anhand eines einfachen Beispiels berechnen
- Du kannst erklären, dass die Auswahl von sinnvollen Metriken kontextspezifisch ist
- Du kannst erklären, dass eine Einordnung der Ergebnisse kontextspezifisch ist

Inhalt

Einstieg

Wir kennen bereits die confusion matrix, mit deren Hilfe wir nun sogenannte Evaluationsmetriken für eine Klassifikation ableiten können. Diese Metriken legen ihren Fokus auf unterschiedliche Aspekte und je nach Disziplin haben sich andere Metriken durchgesetzt.

Evaluationsmetriken Klassifikation



Zur Veranschaulichung sehen wir uns direkt mal ein Beispiel an. Wir haben hier eine confusion matrix mit n=100 fiktiven Beobachtungen, die schon in die passenden Felder der Matrix einsortiert sind. Die Metriken mögen jetzt etwas kompliziert klingen, sind es aber eigentlich gar nicht und sie lassen sich auch gut nachvollziehen.

Eine Metrik, die sehr häufig verwendet wird, ist die accuracy. Sie gibt den Anteil aller richtig klassifizierten Beobachtungen an, die sogenannte Korrektklassifikationsrate. Im Zähler des Bruchs stehen hier also alle richtig prognostizierten Beobachtungen, das sind die True Positives und die True Negatives. Zusammen sind es 60. Im Nenner steht die Summe aller klassifizierten Beobachtungen, also 100. Die accuracy beträgt damit 0,6. Wir klassifizieren also 60 % aller Beobachtungen richtig.

Die Falschklassifikationsrate, oder auch error rate, ist dagegen der Anteil an falsch klassifizierten Beobachtungen. Wir teilen also die Summe aus False Positives und False Negatives durch alle 100 Beobachtungen. Sie ergibt sich auch einfach als 1-accuracy und beträgt hier 0,4. Zusammen ergeben die accuracy und die error rate dann 1 bzw. 100 %.

Das Prinzip der Berechnung bleibt nun bei den weiteren folgenden Metriken sehr ähnlich und lässt sich mit den Einträgen der confusion matrix sehr gut nachvollziehen.

Die True Positive Rate gibt den Anteil aller vom Modell richtig als Positive prognostizierten Beobachtungen an allen wahren Positives an. Man nennt sie auch Sensitivität, Recall oder Trefferquote. Sie beträgt hier im Beispiel 0,57. 57 % aller Positives wurden damit auch vom Modell richtig erkannt.

Die False Negative Rate ergibt sich als 1-True Positive Rate. Sie gibt analog den Anteil der fälschlicherweise als Negative prognostizierten Beobachtungen an allen wahren Positives an. Hier hat sie einen Wert von 0,43. 43 % aller wahren Positives wurden also nicht erkannt.

Die True Negative Rate gibt den Anteil aller richtig als Negatives prognostizierten Beobachtungen an allen wahren Negatives an. Man nennt sie auch Spezifität. Sie beträgt hier 0,62.

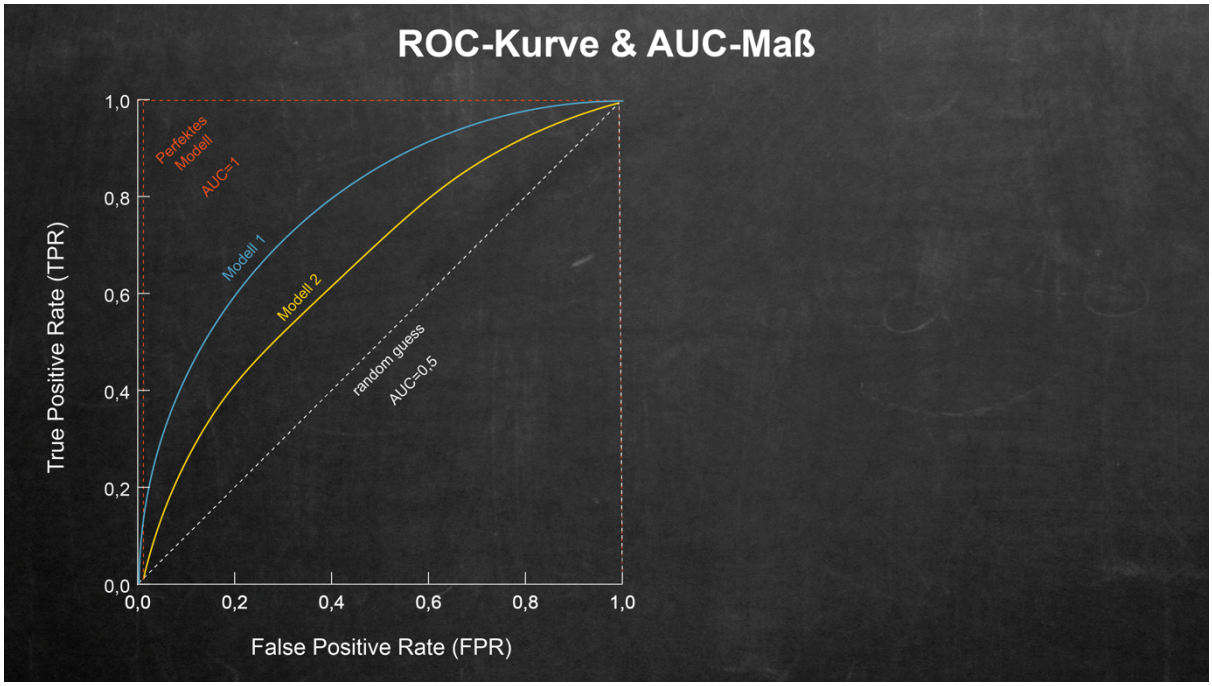
Die False Positive Rate ergibt sich dann als 1-True Negative Rate und entspricht dem Anteil aller falsch als Positives prognostizierten Beobachtungen an allen wahren Negatives. Man nennt sie auch Ausfallrate oder Fehlalarm. Hier beträgt sie 0,38.

Häufig wird auch die sogenannte Precision verwendet. Man nennt sie auch positiven Vorhersagewert und ist ein Maß der Exaktheit des Modells. Sie gibt an, welcher Anteil aller als Positive prognostizierten Beobachtungen wirklich Positives sind. Sie beträgt hier 0,44.

In vielen Anwendungsfällen wollen wir den Fokus auf unterschiedliche Metriken legen. Dann ist es schwierig, das beste Modell zu finden. Daher sind kombinierte Metriken von Vorteil. Z. B. ist der sogenannte F-Score eine Kombination aus Precision und True Positive Rate. Er beträgt hier 0,5.

ROC-Kurve und AUC-Maß

Noch besser ist es, sich ein Gesamtbild anzusehen, welches die Stärken und Schwächen eines Modells herausstellt. Im Falle einer Klassifikation ist eine übliche Methode die receiver operating characteristics Kurve oder auch kurz ROC-Kurve genannt. Sie veranschaulicht den Trade-off zwischen True Positive Rate, also der Trefferquote und der False Positive Rate, also dem Fehlalarm. Die Interpretation ist nicht ganz so intuitiv und ich werde hier nicht ins Detail gehen. Aber es ist sehr sinnvoll, auch dieses Maß zur Evaluation einmal gesehen zu haben.



Die True Positive Rate wird auf der vertikalen und die False Positive Rate auf der horizontalen Achse abgetragen. Beobachtungen werden nun absteigend nach ihrer vom Modell prognostizierten Wahrscheinlichkeit, ein Positive zu sein, sortiert und geschaut, ob es wirklich ein Positive ist. Falls ja, steigt natürlich die True Positive Rate, falls nein, steigt die False Positive Rate. Dies wird für die ganze sortierte Liste fortgeführt und es entsteht eine Kurve.
Die Diagonale repräsentiert einen random guess. Hier ist es also gleich wahrscheinlich, ob eine Beobachtung ein True Positive oder False Positive ist. Unser Prognosemodell ist also nicht besser als Raten. Ein perfektes Modell ist durch die orangene gestrichelte Linie gekennzeichnet. Je weiter die Kurve über der Diagonalen liegt, desto besser ist das Modell. Hier wäre z. B. Modell 1 besser als Modell 2.

Die Fläche unter der Kurve wird als Area Under the Curve-Maß, kurz AUC-Maß bezeichnet. Ein Wert nahe bei 0,5 weist auf ein weniger genaues Modell hin. Ein Wert von 1 repräsentiert ein perfektes Modell.

Diskussion: Eignung Metriken

Die Beurteilung und Einordnung solcher Metriken ist sehr kontextspezifisch und in unterschiedlichen Anwendungsszenarien können unterschiedliche Metriken sinnvoll sein. Die accuracy ist z. B. nicht immer sinnvoll, wenn eine Klasse deutlich größer ist als die andere. Stellen wir uns vor, die Anzahl an Negatives ist 990 und die Anzahl an Positives 10. Klassifiziert das Modell einfach alle Beobachtungen als Negatives, beträgt die accuracy 990/1000 = 99 %, obwohl alle Positives, also die besonders relevante Klasse, falsch klassifiziert wurden. Dagegen ist das AUC-Maß unabhängig von der Klassenverteilung. Es kann also auch bei sehr ungleichen Klassengrößen verwendet werden.

Wichtig können für uns auch die Irrtümer, also die falsch prognostizierten Beobachtungen sein. Dabei sollten wir auch beachten, dass Fehler unterschiedlich schwerwiegende monetäre und nicht-monetäre Folgen nach sich ziehen. Es ist dann also nicht egal, in welche Richtung man sich irrt. Z. B. kann im medizinischen Bereich ein Nichterkennen von Positives, z. B. nicht erkannte Krankheiten, mit schwerwiegenden Konsequenzen verbunden sein. Bei Spamfiltern dagegen wäre vielleicht eher ein möglicher Fehlalarm kritischer.

Aber auch die Beurteilung der Werte solcher Metriken ist sehr kontextspezifisch. Nicht in jedem Fall ist gut auch wirklich gut. Stellen wir uns vor, die Fehlalarmrate für eine Gesichtserkennung am Bahnhof beträgt 0,1 %. Das kommt uns sehr wenig vor. Aber betrachtet man mal die absoluten Zahlen und auch den Kontext, so können selbst 0,1 % Fehlalarme drastische Konsequenzen haben. Bei so vielen Personen am Bahnhof gäbe es sehr viele falsch Verdächtigte, was dann einen erheblichen Aufwand und Kosten für Personenkontrollen durch die Polizei nach sich zieht. Zudem ist in sensibleren Bereichen eine sehr hohe Genauigkeit vermutlich wesentlich wichtiger als z. B. bei Serien- oder Musikvorschlägen.

Quelle [1]

Schauen wir uns nochmal unser Beispiel der Frühwarnsysteme für Studienabbrüche an. Hier wäre es vermutlich sinnvoll, sich nicht nur die korrekten Prognosen anzusehen, sondern auch die Irrtümer. Eine hohe False Positive Rate würde z. B. bedeuten, dass wir viele Studierende, die vermutlich ihr Studium beenden werden, mit Maßnahmen unterstützen. Das würde unnötigerweise den Studierenden Zeit und den Hochschulen Geld kosten. Andererseits würde eine hohe False Negative Rate dazu führen, dass wir viele gefährdete Studierende nicht erkennen und damit auch nicht unterstützen würden. Das kann ebenfalls hohe monetäre und nicht-monetäre Kosten verursachen.

Abschluss

Wir kennen nun einige Metriken zur Evaluation von Klassifikationsmodellen und wissen, dass alle Metriken einen anderen Fokus haben, der bei der Interpretation und Einordnung der Ergebnisse auch berücksichtigt werden sollte. Zudem ist es wichtig, sich individuell nach Anwendungsfall und unseren eigenen Zielen mehrere Metriken anzusehen und zu diskutieren, um ein umfassendes Bild unseres Prognosemodells zu bekommen.

Quellen

Quelle [1] https://www.rwi-essen.de/presse/wissenschaftskommunikation/unstatistik/archiv/2018/detail/erfolgreiche-gesichtserkennung-mit-hunderttausenden-fehlalarmen

Weiterführendes Material

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3. Auflage). Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R., & Tylor, J. (2023). *An Introduction to Statistical Learning - with Applications in Python*. Springer.

Lantz, B. (2015). *Machine learning with R* (2. Auflage). Packt Publishing Ltd, Birmingham.

Disclaimer