Foto: Battlecreek Coffee Roasters auf Unsplash

Evaluierung von KI-Modellen

Wie bewertet man ein Modell?

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

INNOVET
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Binäre Klassifikation

Ein (fiktives) Modell zur Erkennung von Hautkrebs

Die Daten

- 1000 Fotos von Hautunregelmäßigkeiten
- Davon 100 mit Hautkrebs („positiv“), 900 gesund („negativ“)

Die Vorhersagen

- Modell klassifiziert 980 Fälle korrekt, 20 falsch

Wie viele Vorhersagen sind korrekt?

$$\frac{980}{1000} = 0.98 = 98\,\%$$

Klingt super?

Accuracy („Korrektklassifikationsrate“, „Genauigkeit“)

Die Accuracy (auch: Korrektklassifikationsrate) ist ein Evaluationsmaß zum Vergleich von (KI-)Modellen.

Sie gibt den Anteil der Fälle an, für die das Modell die korrekte Klasse vorhersagt.

Welches Modell ist besser?

- Beide Modelle haben eine Accuracy von 98%
- Modell 1 liegt bei tatsächlichen Krebsfällen immer richtig
- Modell 2 erkennt ein Fünftel der Fälle von Krebs nicht

Korrektheit der Vorhersagen von der Erkennung der tatsächlichen Fälle trennen

Krebs?	Modell: ja	Modell: nein
Diagnose: ja	100	0
Diagnose: nein	20	880

Krebs?	Modell: ja	Modell: nein
Diagnose: ja	80	20
Diagnose: nein	0	900

Modell 1

Krebs?	Modell: ja	Modell: nein
Diagnose: ja	Richtig positiv	Falsch negativ
Diagnose: nein	Falsch positiv	Richtig negativ

Precision: 100/120 = 83.3%
Recall: 100/100 = 100%

- Wie viele der positiven Vorhersagen sind korrekt?

  Precision

  (auch: positiver Vorhersagewert)

$$\frac{RP}{RP + FP}$$

- Wie viele der tatsächlichen Fälle werden als positiv erkannt?

  Recall

  (auch: Sensitivität)

$$\frac{RP}{RP + FN}$$

Modell 2

Krebs?	Modell: ja	Modell: nein
Diagnose: ja	80	20
Diagnose: nein	0	900

Precision: 80/80 = 100%
Recall: 80/100 = 80%

- Wie viele der positiven Vorhersagen sind korrekt?

  Precision

  (auch: positiver Vorhersagewert)

$$\frac{RP}{RP + FP}$$

- Wie viele der tatsächlichen Fälle werden als positiv erkannt?

  Recall

  (auch: Sensitivität)

$$\frac{RP}{RP + FN}$$

Precision und Recall sind weitere Evaluationsmaße.

Ein alternativer Begriff für Precision ist „positiver Vorhersagewert“.

Ein alternativer Begriff für Recall ist „Sensitivität“.

Die Precision gibt an, welcher Anteil der positiven Vorhersagen eines binären Klassifikators korrekt ist.

Man berechnet sie also als die Anzahl der richtig positiven Vorhersagen, geteilt durch die Anzahl der richtig positiven Vorhersagen plus die Anzahl der falsch positiven Vorhersagen.

Der Recall gibt an, welcher Anteil der positiven Fälle in den Daten von einem binären Klassifikator korrekt erkannt wird.

Man berechnet ihn also als die Anzahl der richtig positiven Vorhersagen, geteilt durch die Anzahl der richtig positiven Vorhersagen plus die Anzahl der falsch negativen Vorhersagen.

Klassifikation mit mehreren Klassen

Übertragung von Precision und Recall auf Klassifikation mit mehreren Klassen

- Precision und Recall: für 2 Klassen:
  - Positiv vs. Negativ
- Oft aber mehrere Klassen, z.B. Katze vs. Hund vs. Goldfisch

Idee:

- Precision und Recall für jede Klasse einzeln berechnen
  - Katze: Katze vs. nicht Katze
  - Hund: Hund vs. nicht Hund
  - Goldfisch: Goldfisch vs. nicht Goldfisch
- Hinterher zu einem Wert mitteln

Ein Beispiel mit mehreren Klassen

- Übersicht: welche echten Klassen wurden wie klassifiziert?
- **Konfusionsmatrix**
- Diagonale: korrekte Klassifikationen

		Vorhersage		
		Katze	**Hund**	**Goldfisch**
Tatsächliche Klasse	**Katze**	93	7	0
	Hund	14	186	0
	Goldfisch	9	11	20

Die Konfusionsmatrix zeigt, welche tatsächlichen Klassen vom Modell wie klassifiziert wurden.

In den Zeilen findet man normalerweise die tatsächlichen Klassen; in den Spalten die Vorhersagen des Modells.

In der Diagonalen stehen die korrekt klassifizierten Fälle.

Fill in the Blanks: Bestimmen Sie die Accuracy eines Klassifikators für Hunde und Katzen

Fill in the Blanks: Bestimmen Sie die Precision für Hunde

Fill in the Blanks: Bestimmen Sie den Recall für Hunde

Fill in the Blanks: Bestimmen Sie den Recall für Katzen

Fill in the Blanks: Bestimmen Sie die Precision für Katzen

Ein Blick auf die Konfusionsmatrix

- Korrekte Klassen:
  - 100 Katzen
  - 200 Hunde
  - 40 Goldfische
  - Also insgesamt 340 Datenpunkte
- Accuracy:
  - (93 + 186 + 20) / 340 = 87,9%

		Vorhersage			
		Katze	**Hund**	**Goldfisch**	
Tatsächliche Klasse	**Katze**	93	7	0	**100**
	Hund	14	186	0	**200**
	Goldfisch	9	11	20	**40**

Ein Blick auf die Konfusionsmatrix

- Korrekte Klassen:
  - 100 Katzen
  - 200 Hunde
  - 40 Goldfische
  - Also insgesamt 340 Datenpunkte
- Accuracy:
  - (93 + 186 + 20) / 340 = 87,9%

		Vorhersage			
		Katze	**Hund**	**Goldfisch**	Recall
Tatsächliche Klasse	**Katze**	93	7	0	93/100 = 93%
	Hund	14	186	0	186/200 = 93%
	Goldfisch	9	11	20	20/40 = 50 %
	Summe	**116**	**204**	**20**	
	Precision	93/116 = 80,2%	186/204 = 91,2%	20/20 = 100%	

Precision und Recall für alle Klassen

	Katze	Hund	Goldfisch	Mittel	gewichtet
Precision	80,2 %	91,2 %	100 %	90,5 %	89,1 %
Recall	93 %	93 %	50 %	78,7 %	87,8 %
Klassen-wahrscheinlichkeit	$\frac{100}{340} \approx 0{,}29$	$\frac{200}{340} \approx 0{,}59$	$\frac{40}{340} \approx 0{,}12$	**Sinnvoll, wenn alle Klassen gleich wichtig**	**Sinnvoll, wenn Klassen verschieden wichtig**

- Gewichtetes Mittel:
  mit Klassenwahrscheinlichkeiten als Gewichten
  - **Precision** 0,29 * 80,2 % + 0,59 * 91,2 % + 0,12 * 100 % ≈ 89,1 %
  - **Recall** 0,29 * 93 % + 0,59 * 93 % + 0,12 * 50 % ≈ 87,8 %

Mittel für Precision und Recall über alle Klassen

Mittel

- „Macro“ Precision bzw. „Macro“ Recall
- Berücksichtigt alle Klassen gleich
- Sinnvoll, wenn die Klassenverteilung nicht genau bekannt ist
  - D.h. sich für neue Datenpunkte ändern könnte
  - Oder wenn auch seltene Klassen wichtig sind

Gewichtetes Mittel

- „Weighted“ Precision bzw. „weighted“ Recall
- Berücksichtigt die Klassen gemäß ihrer Verteilung im Datensatz
- Sinnvoll, wenn die Klassenverteilung repräsentativ ist
  - D.h. sich auch für neue Datenpunkte eher nicht ändert
  - Und wenn keine der seltenen Klassen besonders wichtig ist

Bei der Berechnung von Durchschnittswerten für Precision und Recall kann man entweder das Mittel über alle Klassen berechnen (Macro Precision bzw. Macro Recall) oder die Werte für die Klassen mit der Klassenwahrscheinlichkeit gewichten.

Beide Werte können sinnvoll sein.

Der F-Score

- Precision und Recall: immer noch 2 Maße
- Für Modellvergleich erwünscht: 1 Maß
- Lösung: Mittelwert aus Precision und Recall
- Aber: diesmal das für Verhältniszahlen mathematisch sinnvollere harmonische Mittel…

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F_1-Score
(auch: F-Score)
F_1-Maß

Der F-Score am Beispiel

	Katze	Hund	Goldfisch	Mittel	Mittel gewichtet
Precision	80,2 %	91,2 %	100 %	90,5 %	89,1 %
Recall	93 %	93 %	50 %	78,7 %	87,8 %

F-Score	F-Score gewichtet
84,1 %	88,4 %

- Aus Macro-Precision und Macro-Recall:

$$F = \frac{2 * 90{,}5\,\% * 78{,}7\,\%}{90{,}5\,\% + 78{,}7\,\%} = 84{,}1\,\%$$

- Mit Gewichtung:

$$F = \frac{2 * 89{,}1\,\% * 87{,}8\,\%}{89{,}1\,\% + 87{,}8\,\%} = 88{,}4\,\%$$

Der F-Score ist das harmonische Mittel aus Precision und Recall.

Auf Deutsch kann man den F-Score auch als F-Maß bezeichnen.

Wenn man betonen möchte, dass bei der Berechnung Precision und Recall gleich gewichtet wurden, kann man auch explizit von F1-Score sprechen.

Single Choice: Macro-Precision

Multiple Choice: Vergleich von Evaluationsmaßen

Single Choice: Berechne den F1-Score

Regression

Bewertung von vorhergesagten Größen

- Kein „richtig“ oder „falsch“
- Eher ein Kontinuum von „sehr nah dran“ bis “weit daneben“

	Korrekter Wert	Vorhersage
Datenpunkt 1	101,24	101,25
Datenpunkt 2	87,13	287,13
Datenpunkt 3	287,13	87,13
Datenpunkt 4	13,4	16
Datenpunkt 5	16	13,4
Summe		

Bewertung von vorhergesagten Größen

- Kein „richtig“ oder „falsch“
- Eher ein Kontinuum von „sehr nah dran“ bis “weit daneben“

	Korrekter Wert	Vorhersage	Fehler
Datenpunkt 1	101,24	101,25	0,01
Datenpunkt 2	87,13	287,13	200
Datenpunkt 3	287,13	87,13	-200
Datenpunkt 4	13,4	16	2,6
Datenpunkt 5	16	13,4	-2,6
Summe			**0,01**

Bewertung von vorhergesagten Größen

- Kein „richtig“ oder „falsch“
- Eher ein Kontinuum von „sehr nah dran“ bis “weit daneben“

	Korrekter Wert	Vorhersage	Fehler	Fehlerquadrat
Datenpunkt 1	101,24	101,25	0,01	0,0001
Datenpunkt 2	87,13	287,13	200	40.000
Datenpunkt 3	287,13	87,13	-200	40.000
Datenpunkt 4	13,4	16	2,6	6,76
Datenpunkt 5	16	13,4	-2,6	6,76
Summe			**0,01**	**80.013,52**

Bewertung von vorhergesagten Größen

- Kein „richtig“ oder „falsch“
- Eher ein Kontinuum von „sehr nah dran“ bis “weit daneben“

MSE

RMSE

	Korrekter Wert	Vorhersage	Fehler	Fehlerquadrat	Mittleres Fehlerquadrat
Datenpunkt 1	101,24	101,25	0,01	0,0001	16.002,7
Datenpunkt 2	87,13	287,13	200	40.000	
Datenpunkt 3	287,13	87,13	-200	40.000	Wurzel aus mittlerem Fehlerquadrat
Datenpunkt 4	13,4	16	2,6	6,76	
Datenpunkt 5	16	13,4	-2,6	6,76	
Summe			0,01	80.013,52	126,50

Das Mittlere Fehlerquadrat, englisch Mean Squared Error (MSE) ist ein Evaluationsmaß für Regressionsmodelle.

Man berechnet es, indem man die Fehler („error“) bei der Vorhersage quadriert (“squared“) und den Durchschnitt berechnet („mean“).

Die Wurzel aus dem mittleren Fehlerquadrat, englisch Root Mean Squared Error (RMSE) ist ein Evaluationsmaß für Regressionsmodelle.

Man berechnet es, indem man den Mittleren Fehler berechnet (MSE) und dann die Wurzel zieht (“root“).

Der Vorteil beim RMSE liegt darin, dass der Wert etwa in der Größenordnung der Fehler liegt.

Der Determinationskoeffizient

- Siehe lineare Regression
- R^2
- Verwandt mit dem Korrelationskoeffizienten
- Erfasst, wie viel Varianz vom Modell erklärt wird
- Evaluiert also Modelle
- Ist vor allem in der Statistik verbreitet

INNOVET

KI B³

Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM

INNOVET

Bundesministerium für Bildung und Forschung

bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

“Evaluationsmaße" von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.de)
(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: unbenannt, von Battlecreek Coffee Roasters (https://unsplash.com/@battlecreekcoffeeroasters) auf Unsplash (https://unsplash.com/photos/JUPPTsfCuiQ), ist lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.