Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2020/21
IF18/MI18

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Gütemaße und Fehlerkosten

Sie haben bereits ein paar sogenannte Gütemaße kennengelernt: den MSE und RMSE sowie die *Korrektheitsrate (accuracy)*. Wir beschäftigen uns also mit der Frage, wie man die Qualität einer Vorhersage messen kann. Ein einfacher Ansatz ist das Berechnen des Fehlers einer Vorhersage (Klassifikation oder numerische Vorhersage). Man vergleicht einfach den erwarteten Wert mit dem vorhergesagten Wert (vgl. Cleve und Lämmel 2014, S. 227 f.):

Definition: *Fehlerrate bei Klassifikationen*

Bei Klassifikationsproblemen ist unter der Fehlerrate der relative Anteil der falsch klassifizierten Beispiele einer Instanzenmenge zu verstehen:

$$Fehlerrate = \frac{Falsche\ Klassenzuordnungen}{Alle\ Klassenzuordnungen} \tag{1}$$

Definition: *Fehlerrate bei numerischer Vorhersage*

Bei numerischer Vorhersage wird diese absolute Aussage durch den Abstand der vorhergesagten zu den erwarteten Ergebnissen ergänzt:

$$Fehlerrate = \frac{\sum_i (Realwert_i - Vorhersagewert_i)^2}{\sum_i Realwert_i^2} \tag{2}$$

Um einen subjektiv besseren Eindruck zu vermitteln, wird häufig die Erfolgsrate – statt der Fehlerrate – angegeben:

$$\text{Erfolgsrate} = 1 - \text{Fehlerrate} \tag{3}$$

Weitere Bezeichnungen für diesen Wert sind Korrektheitsrate, Korrektklassifikationsrate, Vertrauenswahrscheinlichkeit, Treffergenauigkeit oder engl. **accuracy**.

Speziell für Klassifikatoren gibt es eine Reihe weiterer Gütemaße. Voraussetzung ist, dass wir den erwarteten Wert, also die „wahre“ Klasse für die vorhergesagten Objekte kennen.

Beurteilung eines binären Klassifikators

In einer Wahrheitsmatrix (Konfusionsmatrix) lassen sich vier Verteilungszustände bestimmen, die meist in englisch angegeben werden: TP (true positive), FN (false negative), FP (false positive) und TN (true negative). Die Konfusionsmatrix ist eine spezielle Form einer Kontingenztafel (vgl. Tabelle 1):

Richtig positiv ein *positives* Objekt wird als solches vorhergesagt.

	tatsächlich positiv	tatsächlich negativ	$\sum$
positiv vorhergesagt	TP richtig positiv	FP falsch positiv	P
negativ vorhergesagt	FN falsch negativ	TN richtig negativ	N
$\sum$	R	I	$\sum$

Tabelle 1: 2×2 Kontingenztafel (Konfusionsmatrix) für binäre Klassifikatoren

Richtig negativ ein *negatives* Objekt wird als solches vorhergesagt.

Falsch positiv ein *negatives* Objekt wird fälschlicherweise als *positiv* vorhergesagt (Fehler 2. Art).

Falsch negativ ein *positives* Objekt wird fälschlicherweise als *negativ* vorhergesagt (Fehler 1. Art).

Darauf aufbauend werden abgeleitete Kenngrößen definiert (vgl. Cleve und Lämmel 2014, S. 228 f.; vgl. Runkler 2015, S. 89 ff.):

Relevanz die Anzahl der tatsächlich *positiven* Objekte.

$$R = TP + FN \tag{4}$$

Irrelevanz die Anzahl der tatsächlich *negativen* Objekte.

$$I = FP + TN \tag{5}$$

Positivität die Anzahl der als *positiv* vorhergesagten/klassifizierten Objekte.

$$P = TP + FP \tag{6}$$

Negativität die Anzahl der als *negativ* vorhergesagten/klassifizierten Objekte.

$$N = TN + FN \tag{7}$$

Korrekte Klassifikationen alle korrekten Vorhersagen.

$$T = TP + TN \tag{8}$$

Falsche Klassifikationen alle falschen Vorhersagen.

$$F = FP + FN \tag{9}$$

Sensitivität und Falsch-negativ-Rate

Richtig-positiv-Rate Wie oft wurde ein *positives* Objekt auch als solches klassifiziert (Sensitivität, Trefferquote, **recall**, **sensitivity**)?

$$TPR = \frac{TP}{R} = \frac{TP}{TP + FN} \tag{10}$$

Falsch-negativ-Rate Wie oft wurde ein *positives* Objekt als *negatives* Objekt klassifiziert?

$$FNR = \frac{FN}{R} = \frac{FN}{TP + FN} \tag{11}$$

Spezifität und Falsch-positiv-Rate

Richtig-negativ-Rate Wie oft wurde ein *negatives* Objekt auch als solches klassifiziert (Spezitivität, **specificity**)?

$$TNR = \frac{TN}{I} = \frac{TN}{FP + TN} \tag{12}$$

Falsch-positiv-Rate Wie oft wurde ein *negatives* Objekt als *positiv* klassifiziert?

$$FPR = \frac{FP}{I} = \frac{FP}{FP + TN} \tag{13}$$

Positiver und negativer Vorhersagewert

Positiver Vorhersagewert Wie oft ist ein als *positiv* vorhergesagtes Objekt ein *positives* Objkekt (Genauigkeit, Präzision, **precision**)?

$$\frac{TP}{P} = \frac{TP}{TP + FP} \tag{14}$$

Negativer Vorhersagewert Wie oft ist ein als *negativ* vorhergesagtes Objekt ein *negatives* Objekt?

$$\frac{TN}{N} = \frac{TN}{TN + FN} \tag{15}$$

Korrekt- und Falschklassifikationsrate

Korrektheitsrate der Anteil der korrekt klassifizierten Objekte (**accuracy**).

$$\frac{T}{n} \tag{16}$$

Inkorrektheitsrate der Anteil der nicht korrekt klassifizierten Objekte.

$$\frac{F}{n} \tag{17}$$

Negative Falschklassifikationsrate Wie oft ist ein als *negativ* vorhergesagtes Objekt ein *positives* Objekt?

$$\frac{FN}{N} = \frac{FN}{TN + FN} \tag{18}$$

Positive Falschklassifikationsrate Wie oft ist ein als *positiv* vorhergesagtes Objekt ein *negatives* Objekt?

$$\frac{FP}{P} = \frac{FP}{TP + FP} \tag{19}$$

Fehlerkosten

Vor allem am Beispiel von medizinischen Tests wird schnell klar: Fehler ist nicht gleich Fehler (vgl. Runkler 2015, S. 89 ff.). Es wiegt schwerer, wenn eine Diagnose nicht festgestellt wird, der Patient aber dennoch krank ist, als wenn ein Patient fälschlicherweise eine Diagnose erhält und gesund ist. Aus diesem Grund werden die Fehlerarten mit einer Gewichtung bzw. Fehlerkosten versehen.

Man kann hierzu eine Kostenmatrix oder eine Bonusmatrix aufstellen, die je nach Vorhersagetreffer oder -fehler Plus- beziehungsweise Minuspunkte vergibt:

	Vorhersage	
	0	1
0	10	−20
1	−30	20

Ziel ist nun nicht mehr, den Prozentsatz der korrekt vorhergesagten Datensätze, sondern den Gewinn zu maximieren (Cleve und Lämmel 2014, S. 230).

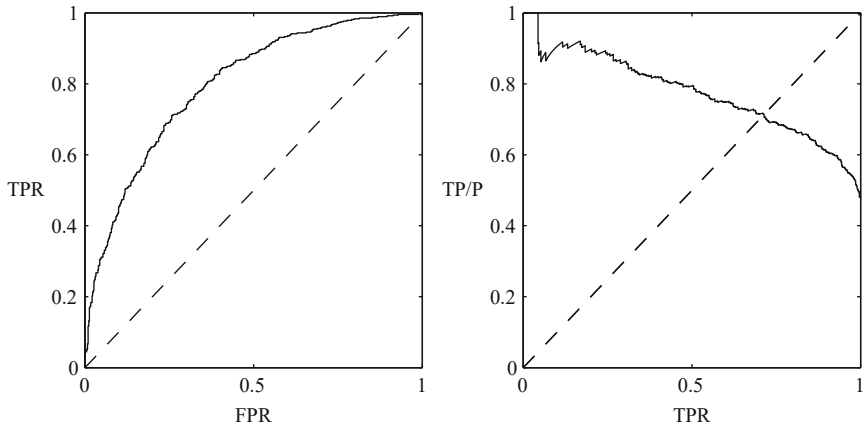

Abbildung 1: ROC- und PR-Diagramm (Runkler 2015, S. 92)

ROC- und PR-Diagramm

Für die Bewertung eines Klassifikators ist es nicht ausreichend, nur eines der o.g. Kriterien zu betrachten. Es kann vorkommen, dass sowohl die Richtig-Positiv-Rate (TPR) als auch die Falsch-Positiv-Rate (FPR) 100 Prozent betragen. Daher werden zumeist zwei oder mehr Kriterien einbezogen, wie z.B. die Grenzwertoptimierungskurve (engl. *Receiver Operating Characteristic*, kurz ROC-Diagramm), ein Streudiagramm von TPR und FPR. Damit ist der Vergleich eines Klassifikators mit unterschiedlichen Parameterwerten möglich, wie z.B. unterschiedliche Parameter zur Erzeugung der Trainings- und Validierungsmenge. Es können ebenso verschiedene Klassifikatoren bewertet oder optimale Werte für Parameter eines Klassifikators ermittelt werden. Es werden vier Bereiche in einem ROC unterschieden (vgl. Runkler 2015, S. 92 f.):

Idealer Klassifikator: 100% TPR, 0% FPR (obere linke Ecke)

immer Positiv: 100% TPR, 100% FPR, (obere rechte Ecke)

immer Negativ: 0% TPR, 0% FPR (untere linke Ecke)

immer Falsch: 0% TPR, 100 FPR (untere rechte Ecke)

Durch Invertierung lässt sich jeder Klassifikator in einen idealen Klassifikator überführen, d.h. eine Spiegelung am Mittelpunkt (50% TPR, 50% FPR) Die Darstellung in ROC erfolgt daher meist nur oberhalb der Hauptdiagonalen (vgl. Abbildung 1; vgl. ebd., S. 92 f.).

Das Genauigkeit-Trefferquote-Diagramm (engl. *Precision Recall* (PR) Diagramm) ist ein Streudiagramm von Genauigkeit (positiver Vorhersagewert) TP/P und Trefferquote (Richtig-Positiv-Rate) FPR. Eine hohe Genauigkeit zu erreichen ist bei niedriger Trefferquote leicht, sodass ein trivialer Klassifikator leicht wenige positive Objekte richtig klassifiziert. Die Genauigkeit sinkt mit steigender Trefferquote. Für einen guten Klassifikator gilt, die Genauigkeit auch bei höherer Trefferquote zu behalten. Gemessen wird dies am Schnittpunkt der PR-Kurve mit der Hauptdiagonalen, dem sogenannten Genauigkeit-Trefferquote-Grenzwert (engl. Precision Recall Breakeven Point). Dieser Grenzwert sollte möglichst hoch sein (vgl. Abbildung 1; vgl. ebd., S. 92 f.).

Aufgabe 1

Beschreiben Sie alle Kenngrößen anhand eines Beispiels.

Aufgabe 2

Eine Klassifikation hat zu folgendem Ergebnis geführt:

- Von den 1000 gegebenen Beispielen gehören 300 Kunden zur Klasse schlecht, 700 zur Klasse gut.
- Unser Verfahren hat von den 300 schlechten Kunden 290 korrekt klassifiziert; von den 700 guten Kunden wurden 650 korrekt erkannt.

a) Erstellen Sie eine entsprechende Konfusionsmatrix.

b) Berechnen Sie Korrekte und Falsche Klassifikation, Relevanz, Recall, Positivität und Precision.

Aufgabe 3

Iris-Daten mit RapidMiner

Erstellen Sie einen neuen RapidMiner-Prozess und laden Sie die Iris-Daten aus den Beispieldaten im lokalen Repository. Vergeleichen Sie die ROCs für die Verfahren *Naive Bayes* und *k-NN* (vgl. Abbildung 2).

Aufgabe 4

Vergleichen Sie die Definitionen von Cleve und Lämmel (2014, S. 228 f.) oder Runkler (2015, S. 91) mit denen von Fawcett (2006, S. 862) und anderen englischsprachigen Quellen. Notieren Sie Ihre Quellen.

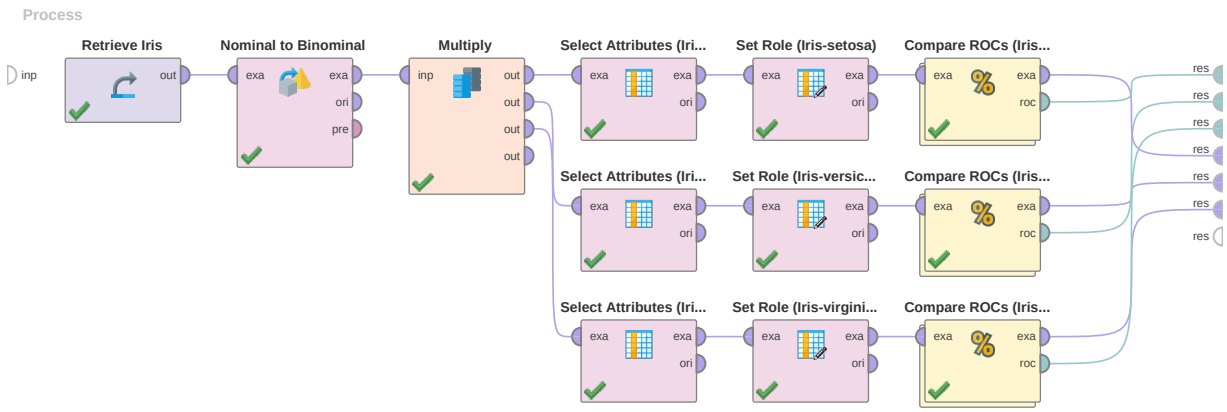

Abbildung 2: RapidMiner-Prozess für ROC-Analyse des Iris-Datensatzes

2. a)

	positiv vorhergesagt	negativ vorhergesagt	
tatsächlich positiv	650	50	700
tatsächlich negativ	10	290	300
	660	340	1000

2. b)

- Korrekte Klassifikationen: $T = TP + TN = 940$
- Falsche Klassifikationen: $F = FP + FN = 60$
- Relevanz: $R = TP + FN = 700$, Irrelevanz: $I = FP + TN = 300$
- Recall: TPR $= \frac{TP}{R} = \frac{650}{700}$
- Positivität: $P = TP + FP = 650 + 10 = 660$
- Precision: $\frac{TP}{P} = \frac{650}{660}$

Literatur

Cleve, Jürgen und Uwe Lämmel (2014). *Data mining*. Studium. München: De Gruyter Oldenbourg. 306 S. ISBN: 978-3-486-72034-1 978-3-486-71391-6.

Fawcett, Tom (Juni 2006). „An introduction to ROC analysis“. In: *Pattern Recognition Letters* 27.8, S. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

Runkler, Thomas A. (2015). *Data mining: Modelle und Algorithmen intelligenter Datenanalyse*. 2., aktualisierte Auflage. Computational intelligence. Wiesbaden: Springer Vieweg. 145 S. ISBN: 978-3-8348-2171-3 978-3-8348-1694-8.