Data Mining Tutorial

Klassifikation I

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2014-05-27 — KDD Übung

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A				
B				
C				
K_i				

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	
B	2	2	1	
C	1	1	3	
K_i				

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$
4		
2		
3		

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$
4	3	1
2		
3		

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$
4	3	1
2	1	3
3		

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$
4	3	1
2	1	3
3	2	2

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\lvert TP\rvert$	$\lvert FP\rvert$	$\lvert FN\rvert$
4	3	1
2	1	3
3	2	2

$$\text{Precision}(K, A) = 4/7$$
$$\text{Precision}(K, B) = 2/3$$
$$\text{Precision}(K, C) = 3/5$$

$$\text{Precision}(K, i) = \frac{|\{o \in K_i \,|\, K(o) = C(o)\}|}{|K_i|} = \frac{|TP_i|}{|TP_i| + |FP_i|}$$

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\lvert TP\rvert$	$\lvert FP\rvert$	$\lvert FN\rvert$
4	3	1
2	1	3
3	2	2

$\text{Precision}(K, A) = 4/7$
$\text{Precision}(K, B) = 2/3$
$\text{Precision}(K, C) = 3/5$

$\text{Recall}(K, A) = 4/5$
$\text{Recall}(K, B) = 2/5$
$\text{Recall}(K, C) = 3/5$

$$\text{Recall}(K, i) = \frac{|\{o \in C_i \,|\, K(o) = C(o)\}|}{|C_i|} = \frac{|TP_i|}{|TP_i| + |FN_i|}$$

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\lvert TP\rvert$	$\lvert FP\rvert$	$\lvert FN\rvert$
4	3	1
2	1	3
3	2	2

$$\text{Precision}(K,A) = 4/7 \quad \text{Recall}(K,A) = 4/5 \quad F_1(K,A) = 2/3$$
$$\text{Precision}(K,B) = 2/3 \quad \text{Recall}(K,B) = 2/5 \quad F_1(K,B) = 1/2$$
$$\text{Precision}(K,C) = 3/5 \quad \text{Recall}(K,C) = 3/5 \quad F_1(K,C) = 3/5$$

$$F_1(K,i) = \frac{2 \cdot \text{Recall}(K,i) \cdot \text{Precision}(K,i)}{\text{Recall}(K,i) + \text{Precision}(K,i)}$$

$$\left(\text{nicht allgemeingültig: } = \frac{2|TP_i|}{2|TP_i| + |FP_i| + |FN_i|}\right)$$

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\|TP\|$	$\|FP\|$	$\|FN\|$
4	3	1
2	1	3
3	2	2

Mittlere Precision, Recall und F_1:

Mittelwert Precision: $\frac{1}{3}(4/7 + 2/3 + 3/5) \approx 0.613$

Mittelwert Recall: $\frac{1}{3}(4/5 + 2/5 + 3/5) = 0.6$

$F_1 \approx \frac{2 \cdot 0.6 \cdot 0.613}{0.6 + 0.613} \approx 0.606.$

Evaluierung von Klassifikatoren

Konfusionsmatrix aufbauen:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$\lvert TP\rvert$	$\lvert FP\rvert$	$\lvert FN\rvert$
4	3	1
2	1	3
3	2	2

Mittelwert der $F_1(K, \ldots) \approx 0.589$. Es ist aber sinnvoller,

$$F_1(\text{Mittlere Precision}(K), \text{Mittlerer Recall}(K))$$

zu verwenden. Precision und Recall sind wichtige Kennzahlen, und F_1 ist "nur" eine Reduktion dieser zwei Kennzahlen auf eine einzige.

Bewertung von Klassifikatoren

Optimaler Klassifikator (nur für zufällige Klassenlabel!):

Immer als die häufigste (Mehrheits-) Klasse klassifizieren.

Erwartete Fehlerrate?

Bewertung von Klassifikatoren

Optimaler Klassifikator (nur für zufällige Klassenlabel!):

Immer als die häufigste (Mehrheits-) Klasse klassifizieren.

Erwartete Fehlerrate?

Da $|A| = |B| = |D|/2$, ist die Fehlerrate 50%.

Leave-one-out Validierung:
Erwartete Fehlerrate?

Bewertung von Klassifikatoren

Leave-one-out Validierung

Leave-one-out Validierung:
Erwartete Fehlerrate?

Die “falsche” Klasse wird jetzt zur Mehrheitsklasse, da wir ja nur das Testobjekt weglassen.

Der erwarteter Fehler wird 100%!
Das ist natürlich zu pessimistisch.

Bewertung von Klassifikatoren

Bootstrap Methode

Bootstrap durch "Ziehen mit Zurücklegen":
Jedes Objekt wird mit einer Wahrscheinlichkeit von ca. $\left(1 - \frac{1}{n}\right)^n \approx 0.368$ *nie* gezogen, also nur ca. 63.2% der Objekte werden zum Training verwendet.
(Bei 10-facher Kreuzvalidierung werden 90% verwendet!)
Die normale Fehlerschätzung wäre pessimistisch.

Bewertung von Klassifikatoren

Bootstrap Methode

Bootstrap durch "Ziehen mit Zurücklegen":
Jedes Objekt wird mit einer Wahrscheinlichkeit von ca. $\left(1 - \frac{1}{n}\right)^n \approx 0.368$ *nie* gezogen, also nur ca. 63.2% der Objekte werden zum Training verwendet.
(Bei 10-facher Kreuzvalidierung werden 90% verwendet!)
Die normale Fehlerschätzung wäre pessimistisch.

Üblicher Ansatz: man integriert auch den beobachteten Klassifikationsfehler (auf den Trainingsdaten!):

$$\begin{aligned}\text{Fehlerrate} = {} & 0.632 \cdot \text{Fehler auf Testdaten} \\ & + 0.368 \cdot \text{Fehler auf Trainingsdaten}\end{aligned}$$

Das wird mehrmals wiederholt (mit unterschiedlichen Stichproben) und dann darüber gemittelt.

Evaluierung von Klassifikatoren

Bootstrap-Methode

Die Fehlerrate des konstanten Klassifikators ist $\approx 50\%$.

Evaluierung von Klassifikatoren

Bootstrap-Methode

Die Fehlerrate des konstanten Klassifikators ist $\approx 50\%$.

Neuer "bester" Klassifikator auf den Trainingsdaten:
"auswendig lernen"!
Auf den Trainingsdaten kann der "auswendig lernen"
Ansatz eine Präzision von bis zu 100% erreichen!

Evaluierung von Klassifikatoren

Bootstrap-Methode

Die Fehlerrate des konstanten Klassifikators ist $\approx 50\%$.

Neuer "bester" Klassifikator auf den Trainingsdaten:
"auswendig lernen"!
Auf den Trainingsdaten kann der "auswendig lernen"
Ansatz eine Präzision von bis zu 100% erreichen!

Dann ergibt sich:

$$\text{Fehlerrate} = 0.632 \cdot 50\% + 0.368 \cdot 0\% = 31.6\%$$

was eine zu optimistische Schätzung ist.