Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Anna Beer, Florian Richter

Knowledge Discovery in Databases

WS 2017/18

Übungsblatt 10: Classification

Besprechung: 25. und 26.01.2018

Aufgabe 10-1 Bewertung von Klassifikatoren

Gegeben sei ein Datensatz mit bekannter Klassenzugehörigkeit der Objekte. Um die Qualität eines Klassifikators K zu ermitteln wurden die Objekte mittels K klassifiziert. Die Klassifikationsergebnisse sind in der folgenden Tabelle dargestellt.

ID	Objektklasse	$K(o)$	ID	Objektklasse	$K(o)$
O_1	A	A	O_2	B	A
O_3	A	C	O_4	C	C
O_5	C	B	O_6	B	B
O_7	A	A	O_8	A	A
O_9	A	A	O_{10}	B	C
O_{11}	B	A	O_{12}	C	A
O_{13}	C	C	O_{14}	C	C
O_{15}	B	B			

- Berechnen Sie anhand der tabellierten Ergebnisse Precision und Recall jeder Klasse.

  Konfusionsmatrix:

	A	B	C	C_i	$\|TP\|$	$\|FP\|$	$\|FN\|$
A	4	0	1	5	4	3	1
B	2	2	1	5	2	1	3
C	1	1	3	5	3	2	2
K_i	7	3	5				

  Zur Erinnerung:

  - True Positives für Klasse i $TP_i = \{o \mid C(o) = i \wedge K(o) = i\}$
  - False Positives für Klasse i $FP_i = \{o \mid C(o) \neq i \wedge K(o) = i\}$
  - False Negatives für Klasse i $FN_i = \{o \mid C(o) = i \wedge K(o) \neq i\}$

  Precision:

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

  oder

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

Recall:

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

oder

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

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

- Um ein vollständiges Maß für die Güte der Klassifikation bezüglich einer Klasse zu haben, wird häufig auch das sogenannte F_1-Measure (harmonisches Mittel zwischen Precision und Recall) verwendet. Das F_1-Measure für Klasse i ist wie folgt definiert:

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

  Berechnen Sie das F_1-Measure für alle Klassen.

  $F_1(K, A) = 2/3$
  $F_1(K, B) = 1/2$
  $F_1(K, C) = 3/5$

- Berechnen Sie die durchschnittliche Precision, den durchschnittlichen Recall und daraus das F_1-Measure.

  Durchschnitt über alle Klassen, nicht Objekte – sonst wird alles zur Accuracy!

  $\text{Precision}(K) = 1/3 \cdot (4/7 + 2/3 + 3/5) \approx 0.613$
  $\text{Recall}(K) = 1/3 \cdot (4/5 + 2/5 + 3/5) = 3/5$
  $F_1(K) \approx \frac{2 \cdot 0.6 \cdot 0.613}{0.6 + 0.613} = 0.606$
  Beachte: Durchschnitt $F_1(K, \ldots) = 0.589$.

Aufgabe 10-2 Naive Bayes

Die Ski-Saison ist eröffnet. Um zuverlässig zu entscheiden, wann Sie Skifahren gehen können und wann nicht, können Sie einen Klassifikator (z.B. Naive Bayes) benutzen. Der Klassifikator wird mit Ihren Erfahrungswerten aus dem letzten Jahr trainiert. Berücksichtigt werden dabei folgende Attribute:
Das Wetter: Das Attribut `Wetter` kann die folgenden drei Werte annehmen: `Sonne`, `Regen` und `Schnee`.
Die Schneehöhe: Das Attribut `Schneehöhe` kann die folgenden zwei Werte annehmen: ≥ 50 (Es liegen mindestens 50 cm Schnee) und < 50 (Es liegen weniger als 50 cm Schnee).

Angenommen, Sie wollten letztes Jahr 8-mal zum Skifahren gehen. Die folgende Tabelle gibt Ihre jeweiligen Entscheidungen wieder:

Wetter	**Schneehöhe**	**Skifahren ?**
Sonne	< 50	nein
Regen	< 50	nein
Regen	≥ 50	nein
Schnee	≥ 50	ja
Schnee	< 50	nein
Sonne	≥ 50	ja
Schnee	≥ 50	ja
Regen	< 50	ja

(a) Berechnen Sie die *a priori* Wahrscheinlichkeiten für die beiden Klassen `Skifahren = ja` und `Skifahren = nein` (auf den Trainingsdaten)!

$$\begin{aligned} P(ski) &= 0.5 \\ P(\neg ski) &= 0.5 \end{aligned}$$

(b) Berechnen Sie für die Klassen die Werteverteilungen aller Attribute.

$$
\begin{aligned}
P(Wetter = Sonne|ski) &= \frac{1}{4} \\
P(Wetter = Schnee|ski) &= \frac{2}{4} \\
P(Wetter = Regen|ski) &= \frac{1}{4} \\
P(Wetter = Sonne|\neg ski) &= \frac{1}{4} \\
P(Wetter = Schnee|\neg ski) &= \frac{1}{4} \\
P(Wetter = Regen|\neg ski) &= \frac{2}{4}
\end{aligned}
$$

$$
\begin{aligned}
P(Schnee \geq 50|ski) &= \frac{3}{4} \\
P(Schnee < 50|ski) &= \frac{1}{4} \\
P(Schnee \geq 50|\neg ski) &= \frac{1}{4} \\
P(Schnee < 50|\neg ski) &= \frac{3}{4}
\end{aligned}
$$

(c) Entscheiden Sie, ob Sie bei den folgenden Wetter- und Schneebedingungen Skifahren gehen oder nicht! Verwenden Sie dazu den naiven Bayes-Klassifikator.

	Wetter	**Schneehöhe**
Tag A	Sonne	≥ 50
Tag B	Regen	< 50
Tag C	Schnee	< 50

$$P(C_i|M) \overset{\text{Bayes}}{=} \frac{P(M|C_i) \cdot P(C_i)}{P(M)} = \frac{P(M|C_i) \cdot P(C_i)}{\sum_{C_j \in C} P(C_j) \cdot P(M|C_j)}$$

A:

$$\begin{aligned} & P(ski|Wetter = Sonne, Schnee \geq 50) \\ = \; & \frac{P(Wetter = Sonne|ski) \cdot P(Schnee \geq 50|ski) \cdot P(ski)}{P(Wetter = Sonne, Schnee \geq 50)} \\ = \; & \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(Wetter = Sonne, Schnee \geq 50)} \\ = \; & \frac{\frac{3}{32}}{P(Wetter = Sonne, Schnee \geq 50)} \end{aligned}$$

$$\begin{aligned} & P(\neg ski|Wetter = Sonne, Schnee \geq 50) \\ = \; & \frac{P(Wetter = Sonne|\neg ski) \cdot P(Schnee \geq 50|\neg ski) \cdot P(\neg ski)}{P(Wetter = Sonne, Schnee \geq 50)} \\ = \; & \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(Wetter = Sonne, Schnee \geq 50)} \\ = \; & \frac{\frac{1}{32}}{P(Wetter = Sonne, Schnee \geq 50)} \end{aligned}$$

$\Rightarrow$ Skifahren

B:

$$
\begin{aligned}
& P(ski|Wetter = Regen, Schnee < 50) \\
= \; & \frac{P(Wetter = Regen|ski) \cdot P(Schnee < 50|ski) \cdot P(ski)}{P(Wetter = Regen, Schnee < 50)} \\
= \; & \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(Wetter = Regen, Schnee < 50)} \\
= \; & \frac{\frac{1}{32}}{P(Wetter = Regen, Schnee < 50)}
\end{aligned}
$$

$$
\begin{aligned}
& P(\neg ski|Wetter = Regen, Schnee < 50) \\
= \; & \frac{P(Wetter = Regen|\neg ski) \cdot P(Schnee < 50|\neg ski) \cdot P(\neg ski)}{P(Wetter = Regen, Schnee < 50)} \\
= \; & \frac{\frac{2}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(Wetter = Regen, Schnee < 50)} \\
= \; & \frac{\frac{6}{32}}{P(Wetter = Regen, Schnee < 50)}
\end{aligned}
$$

$\Rightarrow$ nicht Skifahren

C:

$$
\begin{aligned}
& P(ski|Wetter = Schnee, Schnee < 50) \\
= \; & \frac{P(Wetter = Schnee|ski) \cdot P(Schnee < 50|ski) \cdot P(ski)}{P(Wetter = Schnee, Schnee < 50)} \\
= \; & \frac{\frac{2}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(Wetter = Schnee, Schnee < 50)} \\
= \; & \frac{\frac{2}{32}}{P(Wetter = Schnee, Schnee < 50)}
\end{aligned}
$$

$$
\begin{aligned}
& P(\neg ski|Wetter = Schnee, Schnee < 50) \\
= \; & \frac{P(Wetter = Schnee|\neg ski) \cdot P(Schnee < 50|\neg ski) \cdot P(\neg ski)}{P(Wetter = Schnee, Schnee < 50)} \\
= \; & \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(Wetter = Schnee, Schnee < 50)} \\
= \; & \frac{\frac{3}{32}}{P(Wetter = Schnee, Schnee < 50)}
\end{aligned}
$$

$\Rightarrow$ nicht Skifahren

Aufgabe 10-3 Entscheidungsbäume

Sie wollen die Risikoklasse eines Autofahrers anhand der folgenden Merkmale vorhersagen:

- Zeit seit Bestehen der Fahrprüfung(1-2 Jahre, 2-7 Jahre, >7 Jahre)
- Geschlecht (männlich, weiblich)
- Wohnort(Stadt, Land)

Für Ihre Analyse stehen Ihnen folgende manuell eingeteilte Testbeispiele zu Verfügung:

Person	Zeit seit der Fahrprüfung	Geschlecht	Wohnort	Risikoklasse
1	1-2	m	Stadt	niedrig
2	2-7	m	Land	hoch
3	>7	w	Land	niedrig
4	1-2	w	Land	hoch
5	>7	m	Land	hoch
6	1-2	m	Land	hoch
7	2-7	w	Stadt	niedrig
8	2-7	m	Stadt	niedrig

(a) Konstruieren Sie anhand dieser Trainingsdaten einen Entscheidungsbaum. Benutzen Sie beim Split den Informationsgewinn als Maß für die Unreinheit. Erzeugen Sie dabei für jeden Attributwert einen eigenen Ast. Der Entscheidungsbaum soll terminieren, wenn alle Instanzen im Blatt die gleiche Klasse haben. Die Anwendung eines Pruning-Algorithmus ist nicht erforderlich!

Zur Erinnerung: bei split von T durch Wahl von Attribut A in Partitionen $T_1 \dots T_m$:

$$\begin{aligned} \mathit{entropie}(T) &= -\sum_{i=1}^{k} p_i \cdot \log p_i \\ \mathit{informationsgewinn}(T, A) &= \mathit{entropie}(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} \mathit{entropie}(T_i) \end{aligned}$$

$\mathit{entropie}(T) = 1$, da $p(R = \texttt{niedrig}) = \frac{1}{2} = p(R = \texttt{hoch})$

- IG Zeit
  (i) 1-2 Jahre: $T_1 =$ Person 1,4,6

$$\begin{aligned} p(R = \texttt{niedrig}) &= \frac{1}{3} \\ p(R = \texttt{hoch}) &= \frac{2}{3} \\ \mathit{entropie}(T_1) &= -\sum_{i=1,2} p_i \log p_i \\ &= -\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) \\ &\approx 0,918 \end{aligned}$$

  (ii) 2-7 Jahre: $T_2 =$ Person 2,7,8

$$\begin{aligned} p(R = \texttt{niedrig}) &= \frac{2}{3} \\ p(R = \texttt{hoch}) &= \frac{1}{3} \\ \mathit{entropie}(T_2) &= \mathit{entropie}(T_1) \\ &\approx 0,918 \end{aligned}$$

  (iii) $>$ 7 Jahre: $T_3 =$ Person 3,5

$$
\begin{aligned}
p(R = \texttt{niedrig}) &= \frac{1}{2} \\
p(R = \texttt{hoch}) &= \frac{1}{2} \\
entropie(T_3) &= -\left(\frac{1}{2}\log\frac{1}{2}\right)\cdot 2 \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
&informationsgewinn(T, Zeit) \\
&= entropie(T) - \sum_{i=1,2,3}\frac{|T_i|}{|T|}entropie(T_i) \\
&= 1 - \left(\frac{3}{8}\cdot 0,918 + \frac{3}{8}\cdot 0,918 + \frac{1}{4}\cdot 1\right) \\
&\approx 0,06
\end{aligned}
$$

- IG Geschlecht
  (i) m: $T_1 =$ Person 1,2,5,6,8

$$
\begin{aligned}
p(R = \texttt{niedrig}) &= \frac{2}{5} \\
p(R = \texttt{hoch}) &= \frac{3}{5} \\
entropie(T_1) &\approx 0,971
\end{aligned}
$$

  (ii) w: $T_2 =$ Person 3,4,7

$$
\begin{aligned}
p(R = \texttt{niedrig}) &= \frac{2}{3} \\
p(R = \texttt{hoch}) &= \frac{1}{3} \\
entropie(T_1) &\approx 0,918
\end{aligned}
$$

$$
\begin{aligned}
&informationsgewinn(T, Geschlecht) \\
&= entropie(T) - \sum_{i=1,2}\frac{|T_i|}{|T|}entropie(T_i) \\
&= 1 - \left(\frac{5}{8}\cdot 0,971 + \frac{3}{8}\cdot 0,918\right) \\
&\approx 0,05
\end{aligned}
$$

- IG Wohnort
  (i) Stadt: $T_1 =$ Person 1,7,8

$$
\begin{aligned}
p(R = \texttt{niedrig}) &= 1 \\
p(R = \texttt{hoch}) &= 0 \\
entropie(T_1) &= 0
\end{aligned}
$$

(ii) Land: $T_2 =$ Person 2,3,4,5,6

$$\begin{aligned} p(R = \texttt{niedrig}) &= \frac{1}{5} \\ p(R = \texttt{hoch}) &= \frac{4}{5} \\ entropie(T_2) &\approx 0,722 \end{aligned}$$

$$\begin{aligned} &informationsgewinn(T, Wohnort) \\ &= 1 - \left(0 + \frac{5}{8} \cdot 0,722\right) \\ &\approx 0,55 \end{aligned}$$

Wohnort hat höchsten $informationsgewinn$.

Split 2, rechter Ast: $T = \{2, 3, 4, 5, 6\}$
$entropie(T) = -\left(\frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}\right) \approx 0,722$

- IG Zeit
  (i) 1-2 Jahre: $T_1 =$ Person 4,6

$$\begin{aligned} p(R = \texttt{hoch}) &= 1 \\ entropie(T_1) &= 0 \end{aligned}$$

  (ii) 2-7 Jahre: $T_2 =$ Person 2

$$\begin{aligned} p(R = \texttt{hoch}) &= 1 \\ entropie(T_2) &= 0 \end{aligned}$$

  (iii) $>$ 7 Jahre: $T_3 =$ Person 3,5

$$\begin{aligned} p(R = \texttt{niedrig}) &= \frac{1}{2} \\ p(R = \texttt{hoch}) &= \frac{1}{2} \\ entropie(T_3) &= 1 \end{aligned}$$

$$\begin{aligned} &informationsgewinn(T, Zeit) \\ &= entropie(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} entropie(T_i) \\ &= 0,722 - \left(0 + 0 + \frac{2}{5} \cdot 1\right) \\ &= 0,322 \end{aligned}$$

- IG Geschlecht
  (i) m: $T_1 =$ Person 2,5,6

$$\begin{aligned} p(R = \texttt{hoch}) &= 1 \\ entropie(T_1) &= 0 \end{aligned}$$

  (ii) w: $T_2 =$ Person 3,4

$$\begin{aligned} p(R = \texttt{niedrig}) &= \frac{1}{2} \\ p(R = \texttt{hoch}) &= \frac{1}{2} \\ entropie(T_1) &= 1 \end{aligned}$$

$$\begin{aligned} & informationsgewinn(T, Geschlecht) \\ &= 0,722 - \left(0 + \frac{2}{5} \cdot 1\right) \\ &= 0,322 \\ &= informationsgewinn(T, Zeit) \end{aligned}$$

  Wähle einen beliebig:

(b) Wenden Sie Ihren Entscheidungsbaum auf folgende Autofahrer an:
Person A: 1-2, w, Land
Person B: 2-7, m , Stadt
Person C: 1-2, w, Stadt

Person A: 1-2, w, Land: hoch
Person B: 2-7, m , Stadt: niedrig
Person C: 1-2, w, Stadt: niedrig