Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2019/20
IF17/MI17

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Bayes-Klassifikator / Naive Bayes

Der Naive-Bayes-Algorithmus ist ein wahrscheinlichkeitsbasiertes Verfahren zur Klassifizierung. Vorhergesagt wird die wahrscheinlichste Klasse. Es wird kein Modell trainiert, stattdessen erfolgt die Berechnung der Vorhersage direkt aus den Trainingsdaten. Grundlage bildet die *Bayessche Formel*, wobei angenommen wird, dass alle Attribute voneinander *unabhängig* sind (vgl. Cleve und Lämmel 2014, S. 111):

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{1}$$

Mit der Bayesschen Formel ist die Berechnung von $P(X|Y)$ mittels $P(Y|X)$ und umgekehrt möglich. Dies ist vorteilhaft, da häufig nur eine der beiden bedingten Wahrscheinlichkeiten vorliegt (vgl. ebd., S. 112).

Beispiel: *Vorhersage der Kreditwürdigkeit.*

Gegeben sind die folgenden Datensätze (Trainingsmenge):

Alter	*Einkommen*	*Akademiker*	*Kredit gewähren?*
jung	*hoch*	*nein*	*nein*
jung	*hoch*	*nein*	*nein*
mittel	*hoch*	*nein*	*ja*
alt	*mittel*	*nein*	*ja*
alt	*niedrig*	*ja*	*ja*
alt	*niedrig*	*ja*	*nein*
mittel	*niedrig*	*ja*	*ja*
jung	*mittel*	*nein*	*nein*
jung	*niedrig*	*ja*	*ja*
alt	*mittel*	*ja*	*ja*

Zur Vereinfachung soll die Vorhersage zunächst mit nur einem Attribut bestimmt werden, z.B. für das Alter*. Das Zielattribut hat zwei Ausprägungen:* ja *und* nein *(Binominal). Wir nehmen an, dass uns für einen bisher nicht klassifizierten Datensatz das Alter mit der Ausprägung* jung *bekannt ist.*

Wir kennen aus der Trainingsmenge die Wahrscheinlichkeiten von $P(jung) = 4/10$ *und* $P(kred) = 6/10$ *sowie deren Gegenwahrscheinlichkeiten. Außerdem kennen wir die relative Häufigkeit des Auftretens von* jung *unter der Bedingung der Kreditwürdigkeit* kred = ja *bzw.* kred = nein:

$$P(jung|kred) = \frac{P(jung \wedge kred)}{P(kred)} = \frac{\frac{1}{10}}{\frac{6}{10}} = \frac{1}{6}$$

$$P(jung|\overline{kred}) = \frac{P(jung \wedge \overline{kred})}{P(\overline{kred})} = \frac{\frac{3}{10}}{\frac{4}{10}} = \frac{3}{4}$$

Mit der Bayesschen Formel können wir nun die Wahrscheinlichkeiten dafür berechnen, dass eine Kreditwürdigkeit unter der Bedingung Alter = jung *gegeben oder nicht gegeben ist:*

$$P(kred|jung) = \frac{P(jung|kred) \cdot P(kred)}{P(jung)} = \frac{\frac{1}{6} \cdot \frac{6}{10}}{\frac{4}{10}} = \frac{1}{4}$$

$$P(\overline{kred}|jung) = \frac{P(jung|\overline{kred}) \cdot P(\overline{kred})}{P(jung)} = \frac{\frac{3}{4} \cdot \frac{4}{10}}{\frac{4}{10}} = \frac{3}{4}$$

Wir sagen kreditwürdig *vorher, wenn $P(kred|jung)$ größer ist als $P(\overline{kred}|jung)$. Der neue Datensatz in diesem Beispiel wird somit der Klasse „nicht kreditwürdig“ zugeordnet.*

Es lässt sich beobachten, dass der Term im Nenner in beiden Formeln identisch ist. Da wir nur bezüglich der Größe vergleichen, können wir diesen weglassen. Es handelt sich nun aber nicht mehr um eine Wahrscheinlichkeit P, sodass dies als sogenannter Likelihood L bezeichnet wird (vgl. Cleve und Lämmel 2014, S. 112):

$$L(X|Y) = P(Y|X) \cdot P(X) \tag{2}$$

Anstatt die Wahrscheinlich für $P(Y|X)$ zu berechnen, kann auch die relative Häufigkeit $h[a, k]$ eines Attributs $a \in A$ in der Klasse k bestimmt werden. Der Likelihood wird dann wie folgt berechnet:

$$L[k](A) = \prod_{i=1}^{n} h[a_i, k] \cdot h[k] \tag{3}$$

Die relative Häufigkeit $h[a, k]$ entspricht exakt unserem $P(a|k)$. Wegen der Unabhängigkeit der Attribute des Naive-Bayes-Algorithmus gilt:

$$P(a_1, a_2, \ldots, a_n|k) = \prod_{i=1}^{n} P(a_i|k) \tag{4}$$

Beispiel (fort.)**:** *Vorhersage der Kreditwürdigkeit mit mehreren Attributen.*

Nehmen wir nun an, wir wollen die Klassenzugehörigkeit für einen Datensatz bestimmen, für den Alter = jung, Einkommen = hoch *sowie* Akademiker = nein *bekannt sind.*

Wir berechnen zunächst die Likelihoods für kred = ja *bzw.* kred = nein:

$$\begin{aligned}
L(kred|jung, hoch, \overline{akad}) &= P(jung, hoch, \overline{akad}|kred) \cdot P(kred) \\
&= P(jung|kred) \cdot P(hoch|kred) \cdot P(\overline{akad}|kred) \cdot P(kred) \\
&= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{2}{6} \cdot \frac{6}{10} = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{180} \\
L(\overline{kred}|jung, hoch, \overline{akad}) &= P(jung, hoch, \overline{akad}|\overline{kred}) \cdot P(\overline{kred}) \\
&= P(jung|\overline{kred}) \cdot P(hoch|\overline{kred}) \cdot P(\overline{akad}|\overline{kred}) \cdot P(\overline{kred}) \\
&= \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{3}{4} \cdot \frac{4}{10} = \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{2}{5} = \frac{9}{80}
\end{aligned}$$

Damit wir wieder Wahrscheinlichkeiten erhalten, normalisieren wir die Likelihoods:

$$P[k_j](A) = \frac{L[k_j](A)}{\sum_i L[k_i](A)} \tag{5}$$

Beispiel (fort.)**:** *Berechnen der Wahrscheinlichkeiten:*

$$\begin{aligned}
P(kred|jung, hoch, \overline{akad}) &= \frac{L(kred|jung, hoch, \overline{akad})}{L(kred|jung, hoch, \overline{akad}) + L(\overline{kred}|jung, hoch, \overline{akad})} \\
&= \frac{\frac{1}{180}}{\frac{1}{180} + \frac{9}{80}} = \frac{4}{85} \approx 0,047 \\
P(\overline{kred}|jung, hoch, \overline{akad}) &= \frac{L(\overline{kred}|jung, hoch, \overline{akad})}{L(kred|jung, hoch, \overline{akad}) + L(\overline{kred}|jung, hoch, \overline{akad})} \\
&= \frac{\frac{9}{80}}{\frac{1}{180} + \frac{9}{80}} = \frac{81}{85} \approx 0,953
\end{aligned}$$

Wir sagen die Klasse vorher, die die größte Wahrscheinlichkeit liefert:

$$P(kred|jung, hoch, \overline{akad}) = 0,047 < P(\overline{kred}|jung, hoch, \overline{akad}) = 0,953 \Rightarrow \textit{nicht kreditwürdig}$$

Aufgabe 1

Gegeben sind die folgenden Datensätze:

Kelchblatt-länge	Kelchblatt-breite	Blütenblatt-länge	Blütenblatt-breite	id	label
kurz	mittel	kurz	schmal	id_5	Iris-setosa
mittel	breit	kurz	schmal	id_19	Iris-setosa
kurz	mittel	kurz	schmal	id_23	Iris-setosa
lang	schmal	mittel	mittel	id_77	Iris-versicolor
mittel	mittel	mittel	mittel	id_96	Iris-versicolor
mittel	mittel	mittel	mittel	id_97	Iris-versicolor
lang	mittel	lang	breit	id_103	Iris-virginica
lang	mittel	lang	breit	id_110	Iris-virginica
mittel	mittel	lang	breit	id_125	Iris-virginica
kurz	mittel	kurz	schmal	id_50	(Iris-setosa)
mittel	mittel	mittel	mittel	id_66	(Iris-versicolor)
mittel	mittel	lang	breit	id_111	(Iris-virginica)

a) Berechnen Sie jeweils für die letzten drei Datensätze (Testmenge) die Wahrscheinlichkeiten für jede Klasse und ermitteln Sie die Klassenzugehörigkeit. Vergleichen Sie Ihr Ergebnis mit der Klassenvorgabe aus der Testmenge.

b) Wie können Sie mit numerischen Attributen umgehen?

Aufgabe 2

Vergleich mit RapidMiner

Importieren Sie die den RapidMiner-Prozess aus der Datei Naive␣Bayes.rmp und vergleichen Sie das Resultat mit den Ergebnissen aus Aufgabe 1.

Aufgabe 3

Iris-Datensatz mit RapidMiner

Erstellen Sie einen neuen RapidMiner-Prozess und laden Sie den Iris-Datensatz aus dem Ordner *Samples* des lokalen Repository.

a) Wenden Sie den Operator *Naive Bayes* an. Teilen Sie dazu den Datensatz in eine Trainings- und in eine Testmenge. Messen Sie die *Korrektheitsrate (accuracy)* und beurteilen Sie das Ergebnis (Abbildung 1).

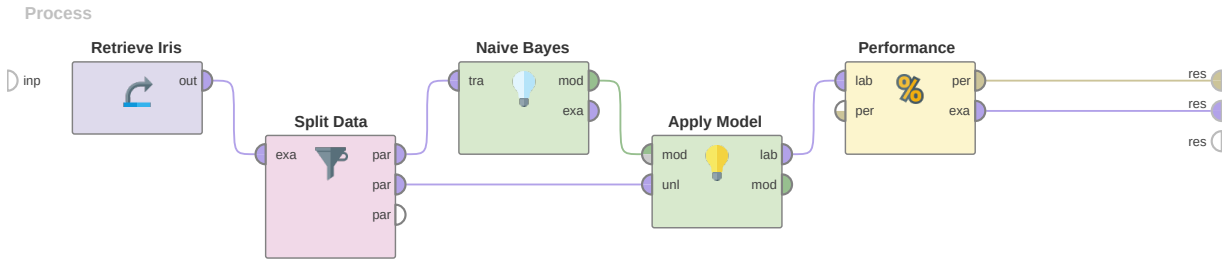

Abbildung 1: RapidMiner-Prozess mit Bayesschem Klassifikator

b) Beschaffen Sie sich Informationen zum sogenannten *Iris-Datensatz* und notieren Sie sich Ihre Quelle. Ordnen Sie den numerischen Attributwerten diskrete Werte durch Intervallbildung zu (vgl. Aufgabe 2). Versuchen Sie die Korrektheitsrate zu verbessern.

Lösungen

1. a) Lösungswege für die drei Datensätze der Testmenge.

Ges.: $P(\text{setosa}|\text{kurz,mittel,kurz,schmal})$; $P(\text{versicolor}|\text{kurz,mittel,kurz,schmal})$;
$P(\text{virginica}|\text{kurz,mittel,kurz,schmal})$

Geg.: $h[\text{setosa}] = 1/3$; $h[\text{versicolor}] = 1/3$; $h[\text{virginica}] = 1/3$;
$h[\text{kurz, setosa}] = 2/3$; $h[\text{kurz, versicolor}] = 0$; $h[\text{kurz, virginica}] = 0$;
$h[\text{mittel, setosa}] = 2/3$; $h[\text{mittel, versicolor}] = 2/3$; $h[\text{mittel, virginica}] = 1$;
$h[\text{kurz, setosa}] = 1$; $h[\text{kurz, versicolor}] = 0$; $h[\text{kurz, virginica}] = 0$;
$h[\text{schmal, setosa}] = 1$; $h[\text{schmal, versicolor}] = 0$; $h[\text{schmal, virginica}] = 0$;

$$\begin{aligned} L(\text{setosa}|\text{kurz,mittel,kurz,schmal}) &= h[\text{kurz, setosa}] \cdot h[\text{mittel, setosa}] \cdot h[\text{kurz, setosa}] \\ &\quad \cdot h[\text{schmal, setosa}] \cdot h[\text{setosa}] \\ &= \frac{2}{3} \cdot \frac{2}{3} \cdot 1 \cdot 1 \cdot \frac{1}{3} = \frac{4}{27} \end{aligned}$$

$$L(\text{versicolor}|\text{kurz,mittel,kurz,schmal}) = 0$$

$$L(\text{virginica}|\text{kurz,mittel,kurz,schmal}) = 0$$

$$P((\text{setosa}|\text{kurz,mittel,kurz,schmal}) = 1$$
$$P(\text{versicolor}|\text{kurz,mittel,kurz,schmal}) = 0$$
$$P(\text{virginica}|\text{kurz,mittel,kurz,schmal}) = 0$$

Antw.: Klassifizierung in Iris-Setosa.

Ges.: $P(\text{setosa}|\text{mittel,mittel,mittel,mittel})$; $P(\text{versicolor}|\text{mittel,mittel,mittel,mittel})$;
$P(\text{virginica}|\text{mittel,mittel,mittel,mittel})$

Geg.: $h[\text{setosa}] = 1/3$; $h[\text{versicolor}] = 1/3$; $h[\text{virginica}] = 1/3$;
$h[\text{mittel, setosa}] = 1/3$; $h[\text{mittel, versicolor}] = 2/3$; $h[\text{mittel, virginica}] = 1/3$;
$h[\text{mittel, setosa}] = 2/3$; $h[\text{mittel, versicolor}] = 2/3$; $h[\text{mittel, virginica}] = 1$;
$h[\text{mittel, setosa}] = 0$; $h[\text{mittel, versicolor}] = 1$; $h[\text{mittel, virginica}] = 0$;
$h[\text{mittel, setosa}] = 0$; $h[\text{mittel, versicolor}] = 1$; $h[\text{mittel, virginica}] = 0$;

$$L(\text{setosa}|\text{mittel,mittel,mittel,mittel}) = 0$$

$$\begin{aligned} L(\text{versicolor}|\text{mittel,mittel,mittel,mittel}) &= h[\text{mittel, versicolor}] \cdot h[\text{mittel, versicolor}] \\ &\cdot\ h[\text{mittel, versicolor}] \cdot h[\text{mittel, versicolor}] \\ &\cdot\ h[\text{versicolor}] \\ &= \frac{2}{3} \cdot \frac{2}{3} \cdot 1 \cdot 1 \cdot \frac{1}{3} = \frac{4}{27} \end{aligned}$$

$$L(\text{virginica}|\text{mittel,mittel,mittel,mittel}) = 0$$

$$\begin{aligned} P((\text{setosa}|\text{mittel,mittel,mittel,mittel}) &= 0 \\ P(\text{versicolor}|\text{mittel,mittel,mittel,mittel}) &= 1 \\ P(\text{virginica}|\text{mittel,mittel,mittel,mittel}) &= 0 \end{aligned}$$

Antw.: Klassifizierung in Iris-Versicolor.

Ges.: $P(\text{setosa}|\text{mittel,mittel,lang,breit})$; $P(\text{versicolor}|\text{mittel,mittel,lang,breit})$;
$P(\text{virginica}|\text{mittel,mittel,lang,breit})$

Geg.:

$h[\text{setosa}] = 1/3$;	$h[\text{versicolor}] = 1/3$;	$h[\text{virginica}] = 1/3$;
$h[\text{mittel, setosa}] = 1/3$;	$h[\text{mittel, versicolor}] = 2/3$;	$h[\text{mittel, virginica}] = 1/3$;
$h[\text{mittel, setosa}] = 2/3$;	$h[\text{mittel, versicolor}] = 2/3$;	$h[\text{mittel, virginica}] = 1$;
$h[\text{lang, setosa}] = 0$;	$h[\text{lang, versicolor}] = 0$;	$h[\text{lang, virginica}] = 1$;
$h[\text{breit, setosa}] = 0$;	$h[\text{breit, versicolor}] = 0$;	$h[\text{breit, virginica}] = 1$;

$$L(\text{setosa}|\text{mittel,mittel,lang,breit}) = 0$$

$$L(\text{versicolor}|\text{mittel,mittel,lang,breit}) = 0$$

$$\begin{aligned} L(\text{virginica}|\text{mittel,mittel,lang,breit}) &= h[\text{mittel, versicolor}] \cdot h[\text{mittel, versicolor}] \\ &\cdot\ h[\text{lang, versicolor}] \cdot h[\text{breit, versicolor}] \\ &\cdot\ h[\text{versicolor}] \\ &= \frac{2}{3} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{3} = \frac{2}{9} \end{aligned}$$

$$\begin{aligned} P((\text{setosa}|\text{mittel,mittel,lang,breit}) &= 0 \\ P(\text{versicolor}|\text{mittel,mittel,lang,breit}) &= 0 \\ P(\text{virginica}|\text{mittel,mittel,lang,breit}) &= 1 \end{aligned}$$

Antw.: Klassifizierung in Iris-Virginica.

1. b) Intervalle bilden.

3. a) Naive␣Bayes␣Iris.rmp

3. b) Naive␣Bayes␣Iris␣diskret.rmp

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.