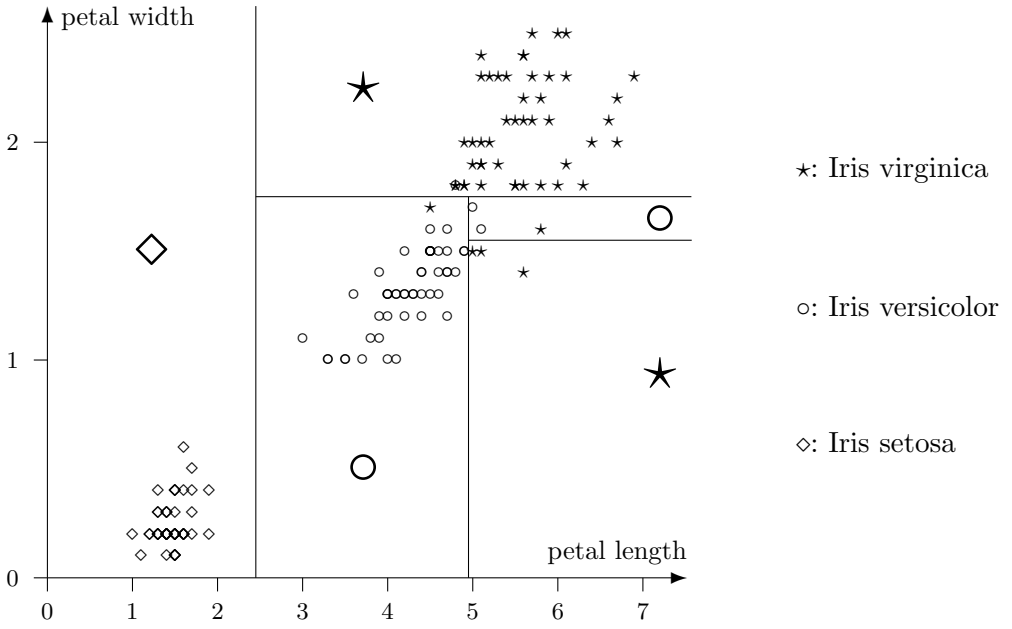
Data Mining / Intelligente Datenanalyse WS 2017/2018
PD Dr.-Ing. habil. Christian Borgelt

Lösung des 5. Übungsblattes

Aufgabe 25 Entscheidungsbäume: Visualisierung

In dem unten gezeigten Diagramm sind die Regionen eingezeichnet, die von dem Entscheidungsbaum der Aufgabenstellung unterschieden werden. Durch große Symbole für die drei Irisarten sind die in diesen Regionen vorhergesagten Klassen angegeben.



Was die Fehlerzahl auf den Trainingsdaten angeht, so ist der Entscheidungsbaum (3 Fehler) dem naiven Bayes-Klassifikator (6 Fehler) überlegen, und das, obwohl er nur zwei Attribute benutzt, während der naive Bayes-Klassifikator auf alle vier Attribute zurückgreift. (Allerdings führt ein naiver Bayes-Klassifikator, der nur die beiden Attribute benutzt, die auch der Entscheidungsbaum verwendet, ebenfalls zu 6 Fehlern.) Der volle Bayes-Klassifikator der Vorlesung (2 Fehler) ist etwas besser, aber auch nur deshalb, weil er alle vier Attribute benutzt. Schränkt man den vollen Bayes-Klassifikator auf die beiden Attribute ein, die im Baum benutzt werden, so erreicht er die gleiche Güte (3 Fehler) wie der Entscheidungsbaum.

Nach der Fehlerzahl sind also Entscheidungsbaum und voller Bayes-Klassifikator gleich gut. Was aber die Struktur und Lage der Entscheidungsregionen angeht, ist der volle Bayes-Klassifikator zu bevorzugen. Er paßt sich sehr gut an die Struktur der Daten an, die in der Tat näherungsweise mehrdimensional normalverteilt sind. (Die Klasse Iris virginica ist allerdings eventuell aus zwei Normalverteilungen zusammengesetzt.) Dagegen ist speziell der schmale horizontale Streifen rechts im Diagramm, in dem der Entscheidungsbaum die Klasse Iris versicolor vorhersagt, nicht sehr plausibel, da in Testdaten in diesem Streifen mit hoher Wahrscheinlichkeit Fälle der Klasse Iris virginica liegen. Ein Entscheidungsbaum hat daher zwar Vorteile dadurch, daß er als Menge von Regeln (eine Regel je Pfad von der Wurzel zu einem Blatt) interpretiert werden kann, während die Interpretation der Kovarianzmatrix eines vollen Bayes-Klassifikators schwierig ist, doch sind die Entscheidungsregionen, die er beschreibt, nicht immer den Daten gut angepaßt.

Aufgabe 26 Induktion von Entscheidungsbäumen

Wird kein Attribut getestet und einfach die Mehrheitsklasse „Play“ vorhergesagt, so macht man 5 Fehler (36%). Für die einzelnen Attribute erhält man folgende Tabellen:

Outlook	Play	Don't Play	Vorhersage	Fehler
sunny	2	3	Don't Play	2 (40%)
overcast	4	0	Play	0 (0%)
rain	3	2	Play	2 (40%)
gesamt	9	5		4 (29%)

Temp.	Play	Don't Play	Vorhersage	Fehler
< 84	9	4	Play	4 (31%)
> 84	0	1	Don't Play	0 (0%)
gesamt	9	5		4 (29%)

Humidity	Play	Don't Play	Vorhersage	Fehler
< 82	7	2	Play	2 (22%)
> 82	2	3	Don't Play	2 (40%)
gesamt	9	5		4 (29%)

Windy?	Play	Don't Play	Vorhersage	Fehler
false	6	2	Play	2 (25%)
true	3	3	Don't Play	3 (50%)
gesamt	9	5		5 (36%)

Bis auf das Attribut „Windy?“ führen alle Attribute nur zu einer Verbesserung auf 4 Fehler (29%). Diese Verbesserung wird bei der Feuchtigkeit (Humidity) sogar nur dadurch erreicht, daß ein einzelnes Beispiel abgetrennt wird, was sicherlich kein sehr sinnvolles Vorgehen ist. Da bei gleicher Fehlerzahl das in der Tabelle weiter links stehende Attribut gewählt werden sollte, wählen wir „Outlook“ als erstes Testattribut und teilen den Datensatz nach den Werten dieses Attributes auf. Für den Wert „overcast“ erhalten wir eine klassenreine Teilmenge (alle Beispiele haben die Klasse „Play“), so daß dieser Zweig mit einem Blatt abgeschlossen werden kann. Für die Werte „sunny“ und „rain“ müssen wir dagegen noch einmal die anderen Attribute betrachten. Für den Attributwert „sunny“ erhält man folgende Tabellen:

Temp.	Play	Don't Play	Vorhersage	Fehler
< 77	2	1	Play	1 (33%)
> 77	0	2	Don't Play	0 (0%)
gesamt	2	3		1 (20%)

Humidity	Play	Don't Play	Vorhersage	Fehler
< 75	2	0	Play	0 (0%)
> 75	0	3	Don't Play	0 (0%)
gesamt	2	3		0 (0%)

Windy?	Play	Don't Play	Vorhersage	Fehler
false	1	2	Don't Play	1 (25%)
true	1	1	Play	1 (50%)
gesamt	2	3		2 (40%)

Mit einem Schwellenwert von 75% führt das Attribut „Humidity“ als einziges zu 0 Fehlern. Wir wählen es daher als Testattribut im Zweig „Outlook = sunny“. Da die Aufteilung der Fälle zwei klassenreine Teilmengen ergibt, können beide Zweige mit Blättern abgeschlossen werden. Für den Attributwert „rain“ des Attributes „Outlook“ erhält man folgende Tabellen:

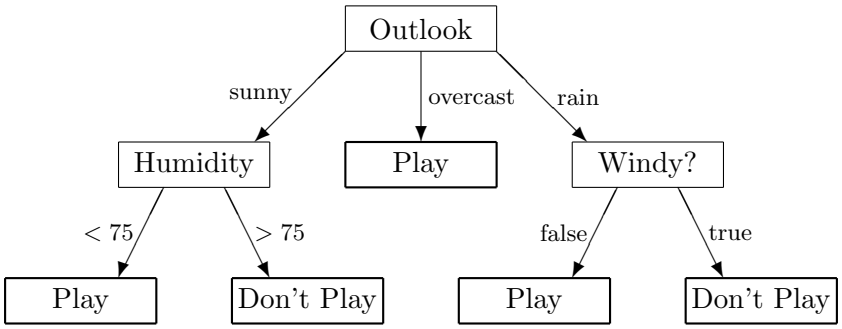
Temp.	Play	Don't Play	Vorhersage	Fehler
< 66	3	1	Play	1 (25%)
> 66	0	1	Don't Play	0 (0%)
gesamt	3	2		1 (20%)

Humidity	Play	Don't Play	Vorhersage	Fehler
< 90	2	2	Don't Play	2 (50%)
> 90	1	0	Play	0 (0%)
gesamt	3	2		2 (40%)

Windy?	Play	Don't Play	Vorhersage	Fehler
false	3	0	Play	0 (0%)
true	0	2	Don't Play	0 (0%)
gesamt	3	2		0 (0%)

Hier führt das Attribut „Windy“ als einziges zu 0 Fehlern. Wir wählen es daher als Testattribut im Zweig „Outlook = rain“. Da die Aufteilung der Fälle zwei klassenreine Teilmengen ergibt, können beide Zweige mit Blättern abgeschlossen werden.

Damit haben wir insgesamt den Entscheidungsbaum:



Dieser Entscheidungsbaum liefert auf den Trainingsdaten eine perfekte Vorhersage, macht also keine Fehler.

Aufgabe 27 Induktion von Entscheidungsbäumen

Wir bauen den Entscheidungsbaum auf die gleiche Weise auf wie in Aufgabe 26. Da die beiden Klassen gleich häufig sind, macht man bei einer Vorhersage ohne Testattribut 4 Fehler (50%). Die Tabellen für die drei Attribute sind:

A_1	f	w	V.	Fehler
f	2	2	f	2 (50%)
w	2	2	w	2 (50%)
ges.	4	4		4 (50%)

A_2	f	w	V.	Fehler
f	2	2	f	2 (50%)
w	2	2	w	2 (50%)
ges.	4	4		4 (50%)

A_3	f	w	V.	Fehler
f	3	1	f	1 (25%)
w	1	3	w	1 (25%)
ges.	4	4		2 (25%)

Da das Attribut A_3 die beste Vorhersagegüte liefert, wird es als Testattribut für die Wurzel ausgewählt. In der Teilmenge $A_3 = f$ erhalten wir die Tabellen:

A_1	f	w	V.	Fehler
f	1	0	f	0 (0%)
w	2	1	f	1 (33%)
ges.	3	1		1 (25%)

A_2	f	w	V.	Fehler
f	1	1	w	0 (50%)
w	2	0	f	1 (0%)
ges.	3	1		1 (25%)

Keines der beiden Attribute führt also zu einer Verbesserung. Wir könnten nun entscheiden, den Baumaufbau abzubrechen und ein Blatt zu erzeugen, oder wir können willkürlich eines der beiden Attribute als Testattribut auswählen, in der Hoffnung, daß in den Zweigen doch noch eine bessere Klassifikation erzielt werden kann. Wir wählen hier das Attribut A_2, da es eine gleichmäßigere Aufteilung der Beispiele bewirkt, außerdem zu einer größeren klassenreinen Teilmenge führt. Für $A_2 = w$ erzeugen wir ein Blatt, für $A_2 = f$ erhalten wir die Tabelle:

A_1	f	w	V.	Fehler
f	1	0	f	0 (0%)
w	0	1	w	0 (0%)
ges.	1	1		0 (0%)

Mit diesem Test erzielen wir 0 Fehler und können damit den Baumaufbau in diesem Zweig abbrechen. Wenden wir uns nun der Teilmenge $A_3 = w$ zu. Wir erhalten die Tabellen:

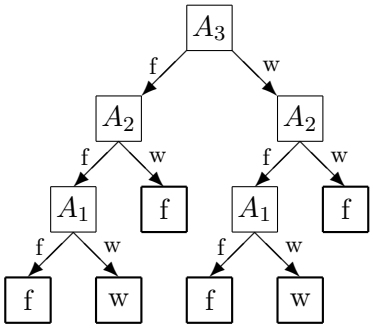
A_1	f	w	V.	Fehler
f	1	2	f	0 (0%)
w	0	1	f	1 (33%)
gesamt	1	3		1 (25%)

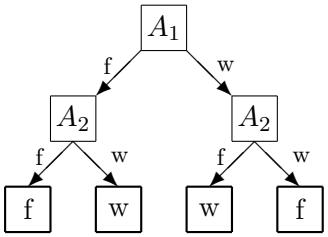
A_2	f	w	V.	Fehler
f	1	1	w	0 (50%)
w	0	2	w	1 (0%)
gesamt	1	3		1 (25%)

Die Situation ist also die gleiche wie im Fall $A_3 = f$. Wieder wählen wir das Attribut A_2, da es eine gleichmäßigere Aufteilung der Beispiele bewirkt, außerdem zu einer größeren klassenreinen Teilmenge führt. Für $A_2 = w$ erzeugen wir ein Blatt, für $A_2 = f$ erhalten wir die Tabelle:

A_1	f	w	V.	Fehler
f	1	0	f	0 (0%)
w	0	1	w	0 (0%)
gesamt	1	1		0 (0%)

Mit diesem Test erzielen wir 0 Fehler und können damit den Baumaufbau in diesem Zweig abbrechen. Insgesamt erhalten wir den unten links gezeigten Entscheidungsbaum.





Dieser Entscheidungsbaum ist jedoch deutlich komplexer als nötig. Schon die Tatsache, daß der rechte und linke Unterbaum nach Test des Attributes A_3 gleich aussehen, läßt vermuten,

daß es auch einfacher geht. In der Tat prüft man leicht nach, daß sich die Klasse als Exklusiv-Oder-Verknüpfung der Attribute A_1 und A_2 berechnen läßt, was z.B. durch den oben rechts gezeigten Entscheidungsbaum dargestellt werden kann. Dieser Entscheidungsbaum wird jedoch von dem Standardverfahren wegen der gierigen Auswahl des Testattributes nicht gefunden, da die Attribute A_1 und A_2 einzeln keine Information über die Klasse liefern, das Attribut A_3 dagegen wenigstens eine gewisse Bestimmung der Klasse erlaubt.

Dieses Problem läßt sich durch Entscheidungsbaum-Lernverfahren beheben, die ein Vorausschauen enthalten, also statt nur die Einzelattribute zu bewerten, alle möglichen Entscheidungsbäume bis zu einer bestimmten Tiefe aufbauen und diese bewerten. Ein solches Vorgehen erhöht jedoch erheblich die Rechenzeit für die Entscheidungsbaum-Induktion, die dadurch erzielte Verbesserung rechtfertigt den zusätzlichen Aufwand meist nicht.

Aufgabe 28 Entscheidungsbäume: Attributauswahlmaße

Um die Maße zu berechnen formen wir zunächst deren Formeln (siehe Folien der Vorlesung) so um, daß sie direkt aus den Einträgen der beiden Kontingenztafeln berechnet werden können. Dazu benutzen wir die Beziehungen (siehe ebenfalls Folien der Vorlesung):

$$p_{ij} = \frac{N_{ij}}{N_{\cdot\cdot}}, \qquad p_{i\cdot} = \sum_{j=1}^{n_A} p_{ij} = \frac{N_{i\cdot}}{N_{\cdot\cdot}}, \qquad p_{\cdot j} = \sum_{i=1}^{n_C} p_{ij} = \frac{N_{\cdot j}}{N_{\cdot\cdot}},$$

wobei die N_{ij} die Einträge der Kontingenztafel, die $N_{i\cdot}$ die Zeilensummen, die $N_{\cdot j}$ die Spaltensummen und $N_{\cdot\cdot}$ die Gesamtzahl der Beispielfälle, also die Summe aller Einträge der Kontingenztafel sind. Wir erhalten so

$$\begin{aligned} I_{\text{gain}}(C,A) &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} p_{ij} \log_2 \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} \\ &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} \frac{N_{ij}}{N_{\cdot\cdot}} \log_2 \frac{\frac{N_{ij}}{N_{\cdot\cdot}}}{\frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}}} = \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} \frac{N_{ij}}{N_{\cdot\cdot}} \log_2 \frac{N_{ij}N_{\cdot\cdot}}{N_{i\cdot}N_{\cdot j}} \end{aligned}$$

für den Informationsgewinn und

$$\begin{aligned} \chi^2(C,A) &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} N_{\cdot\cdot} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \\ &= \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} N_{\cdot\cdot} \frac{\left(\frac{N_{ij}}{N_{\cdot\cdot}} - \frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}}\right)^2}{\frac{N_{i\cdot}}{N_{\cdot\cdot}}\frac{N_{\cdot j}}{N_{\cdot\cdot}}} = \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} \frac{(N_{ij}N_{\cdot\cdot} - N_{i\cdot}N_{\cdot j})^2}{N_{i\cdot}N_{\cdot j}N_{\cdot\cdot}} \end{aligned}$$

für das χ^2-Maß. Mit diesen Formeln berechnen wir nun für den Informationsgewinn

$$\begin{aligned} I_{\text{gain}}(C,A) &= \frac{9}{48}\log_2\frac{9\cdot 48}{16\cdot 16} + \frac{4}{48}\log_2\frac{4\cdot 48}{16\cdot 16} + \frac{3}{48}\log_2\frac{3\cdot 48}{16\cdot 16} \\ &+ \frac{3}{48}\log_2\frac{3\cdot 48}{16\cdot 16} + \frac{9}{48}\log_2\frac{9\cdot 48}{16\cdot 16} + \frac{4}{48}\log_2\frac{4\cdot 48}{16\cdot 16} \\ &+ \frac{4}{48}\log_2\frac{4\cdot 48}{16\cdot 16} + \frac{3}{48}\log_2\frac{3\cdot 48}{16\cdot 16} + \frac{9}{48}\log_2\frac{9\cdot 48}{16\cdot 16} \approx 0.1652 \end{aligned}$$

und

$$I_{\text{gain}}(C,B) = \frac{9}{48}\log_2\frac{9\cdot 48}{16\cdot 16} + \frac{4}{48}\log_2\frac{4\cdot 48}{16\cdot 16} + \frac{3}{48}\log_2\frac{3\cdot 48}{16\cdot 16}$$

$$
\begin{aligned}
&+ \quad \frac{6}{48}\log_2\frac{6\cdot 48}{16\cdot 16} + \frac{6}{48}\log_2\frac{6\cdot 48}{16\cdot 16} + \frac{4}{48}\log_2\frac{1\cdot 48}{16\cdot 16} \\
&+ \quad \frac{1}{48}\log_2\frac{1\cdot 48}{16\cdot 16} + \frac{6}{48}\log_2\frac{6\cdot 48}{16\cdot 16} + \frac{9}{48}\log_2\frac{9\cdot 48}{16\cdot 16} \quad \approx \quad 0.1754.
\end{aligned}
$$

Für das χ^2-Maß erhalten wir

$$
\begin{aligned}
\chi^2(C,A) \quad = \quad & \frac{(9\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(4\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(3\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \\
+ \quad & \frac{(3\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(9\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(4\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \\
+ \quad & \frac{(4\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(3\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(9\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \quad = \quad 11.625
\end{aligned}
$$

und

$$
\begin{aligned}
\chi^2(C,B) \quad = \quad & \frac{(9\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(4\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(3\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \\
+ \quad & \frac{(6\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(6\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(4\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \\
+ \quad & \frac{(1\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(6\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} + \frac{(9\cdot 48-16\cdot 16)^2}{16\cdot 16\cdot 48} \quad = \quad 10.5.
\end{aligned}
$$

Der Informationsgewinn bewertet also das Attribut B besser als das Attribut A, während das χ^2-Maß das Attribut A bevorzugt. Folglich unterscheiden i.a. sich die Entscheidungsbäume, die mit diesen beiden Maßen aufgebaut werden.

Wenn man die Kontingenztafeln etwas genauer betrachtet, kann man anschauliche Prinzipien vermuten, nach denen die beiden Maße das Testattribut auswählen. Da sich die beiden Tafeln nur in den vier Einträgen unterscheiden, müssen die Unterschiede in diese vier Werten die Entscheidung hervorrufen. Offenbar bewertet der Informationsgewinn eine Situation, in der für einen Attributwert eine Klasse (fast) ausgeschlossen werden kann (hier: Attributwert b_1) sehr gut, auch wenn für einen anderen Attributwert (hier: b_2) dann so gut wie Unterscheidung möglich ist. Er „hofft“ damit gewissermaßen auf tiefere Ebenen des Baumes und andere Attribute, um die Klassifikation zu verbessern. Das χ^2-Maß dagegen scheint Situationen zu bevorzugen, in denen klare Evidenz für eine Klasse vorliegt (hier: Attributwerte a_1 und a_2). Es wählt daher mehr nach der Devise „Lieber den Spatz in der Hand als die Taube auf dem Dach.“, indem es eine möglichst gute Klassifikation mit dem einzelnen Attribut anstrebt, und „hofft weniger“ auf tiefere Ebenen des Baumes. In diesem Sinne kann man Attributauswahlmaße als Strategieparameter des Entscheidungsbaumlernens sehen.