Lösungshinweise zu den Aufgaben

Unterrichtsmodul *Einführung maschinelles Lernen mit Entscheidungsbäumen*

Pascal Schmidt, Stefan Strobel

Aufgabe 1

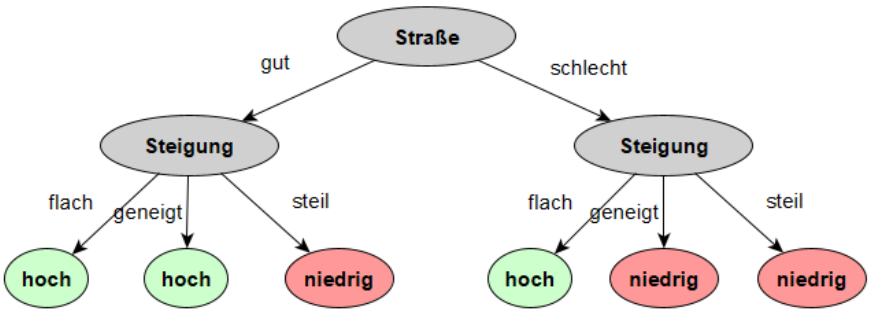
a)

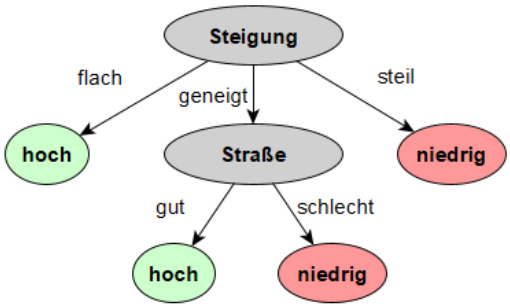
	Wetter	Tag	Entfernung	Ski fahren?
1	Niederschlag	WE	78	Nein
2	sonnig	AT	210	Nein
3	bewölkt	AT	145	Nein
4	bewölkt	WE	145	Ja

b) Ein Entscheidungsbaum ist ein Baum, in dessen inneren Knoten Merkmale der Datenelemente abgefragt werden. Die jeweils ausgehenden Kanten sind mit den möglichen Merkmalsausprägungen bzw. vorgegebenen Alternativen beschriftet. Der jeweilige Verzweigungsgrad ist also von dem abgefragten Merkmal abhängig. Die Blätter geben die Klasse an, mit der ein Datenelement klassifiziert wird, wenn es bei der Baumtraversierung das entsprechende Blatt erreicht.

Aufgabe 2

a) Mögliche Lösungen:





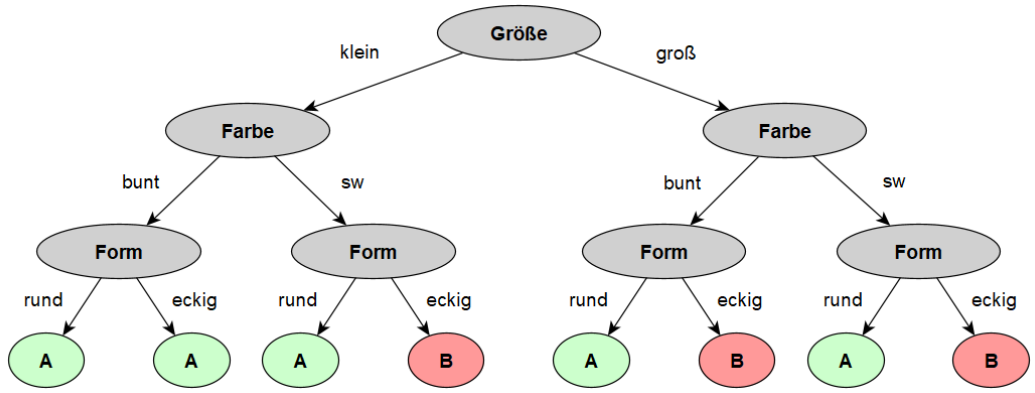
b) In beiden Bäumen landet das Datenelement in einem Blatt mit der Beschriftung „hoch“, wird also falsch klassifiziert.

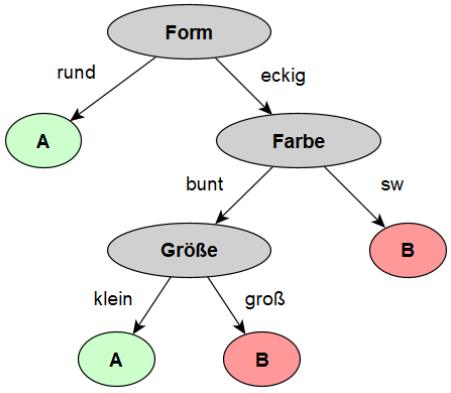
Da es jedoch zwei Datenelemente mit der Merkmalskombination (gut; flach) und dem Label „hoch“ gibt – man bezeichnet sollte Datenelemente als *Duplikate* -, sollte man den Baum auch in Kenntnis des neuen Datenelements nicht ändern, sondern per Mehrheitsentscheid weiterhin in dem betreffenden Blatt das Label „hoch“ ausweisen.

Es handelt sich hier um ein Beispiel für *widersprüchliche Daten*. Eine solche Situation ist nicht ungewöhnlich: Das Datenelement kann beispielsweise dadurch verursacht werden, dass am betreffenden Tag schlechte Witterungsverhältnisse vorlagen oder ein Tempolimit auf einem Streckenabschnitt besteht. Grundsätzlich können auch Messfehler und -ungenauigkeiten zu widersprüchlichen Daten führen.

Aufgabe 3

Mögliche Lösungen (Auswahl):



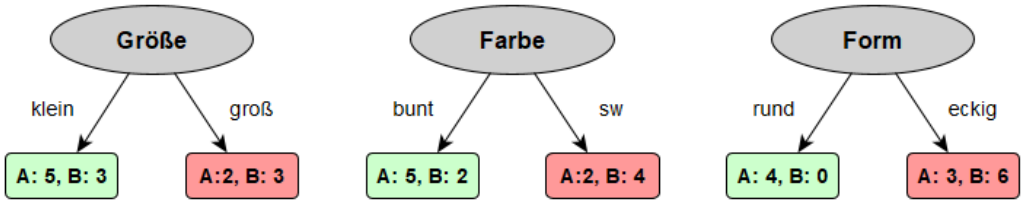


Der obere Baum entsteht, wenn die Merkmale in der Reihenfolge genutzt werden, in der sie in der Tabelle vorkommen (Größe, Farbe, Form). Der linke Teilbaum mit der Wurzel *Form* kann zur Vereinfachung noch durch ein Blatt mit der Beschriftung „A“ ersetzt werden. Eine Analyse der untersten Knotenebene deutet daraufhin, dass eine initiale Aufteilung der Daten nach dem Merkmal *Form* einen kleineren Baum erzeugen könnte, da beispielsweise alle runden Elemente zur Klasse A gehören. Die untere Abbildung bestätigt dies.

Das Datenelement 5 (klein, bunt, eckig: B) steht im Widerspruch zu den Datenelementen (Duplikaten) 8, 9 und 12. Es landet – aufgrund des Mehrheitsentscheides – in einem Blatt mit Label A und wird damit von beiden Bäumen als einziges Element falsch klassifiziert.

Aufgabe 4

a) Für jede mögliche Aufteilung ist die Aufteilung der Datenelemente auf die beiden Blätter angegeben. Die Färbung richtet sich nach der in dem jeweiligen Blatt dominierenden Klasse.

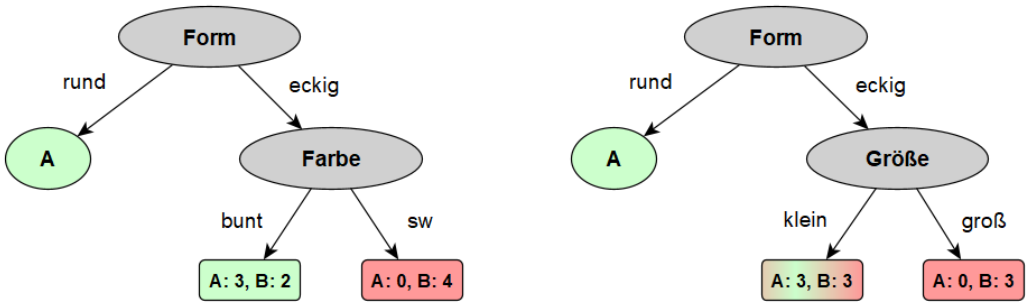


b) Die Zahl der falsch klassifizierten Datenelemente (= Summe der Minderheiten in den beiden Kinderknoten) variiert:

- Größe: 3 + 2 = 5
- Farbe: 2 + 2 = 4
- Form: 0 + 3 = 3

Insofern erzeugt der rechts abgebildete Baum mit der Wurzel *Form* die wenigsten Fehlklassifikationen auf den Trainingsdaten und wird daher bevorzugt.

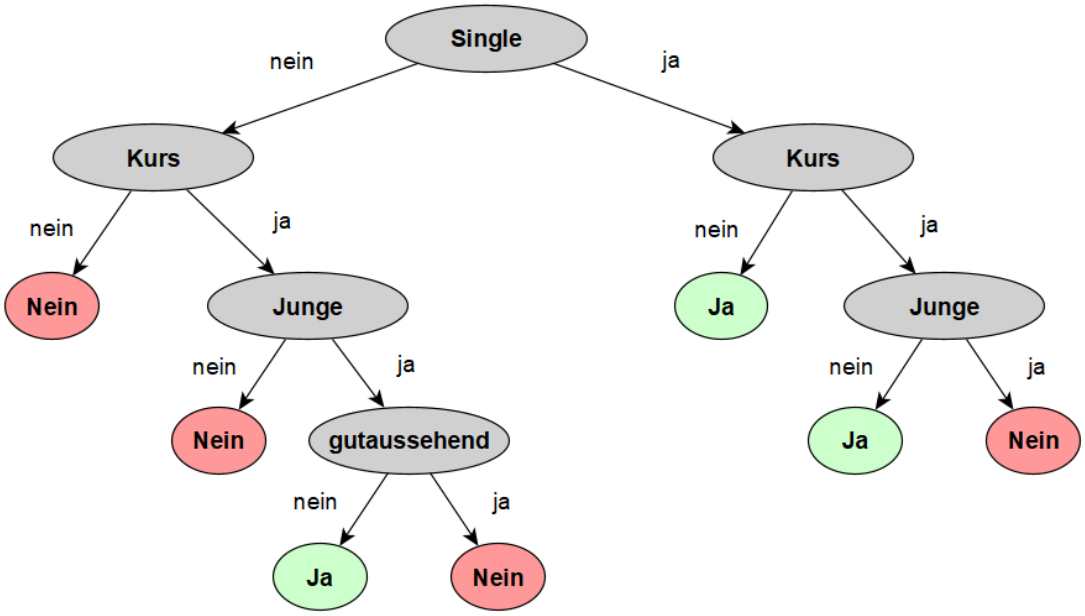
c) Das linke Blatt des Baumes ist bereits sortenrein, hier gibt es also nichts weiter zu tun. Das rechte Blatt enthält allerdings sowohl Elemente der Klasse A (3) als auch der Klasse B (6). Bezüglich dieser Elemente ist zu überprüfen, ob eine Aufspaltung nach *Größe* oder nach *Farbe* weniger Fehlklassifikationen verursacht.



Aus der Abbildung wird deutlich, dass die Aufteilung anhand des Merkmals *Farbe* eine Fehlklassifikation weniger als die Aufteilung anhand der *Größe* verursacht und daher zu bevorzugen ist.

Aufgabe 5

a) Lösung (bei gleicher Anzahl an Fehlklassifikationen wurden die Merkmale in der Reihenfolge ausgewählt, in der sie in der Tabelle stehen: Kurs, Junge, Single, gutaussehend):



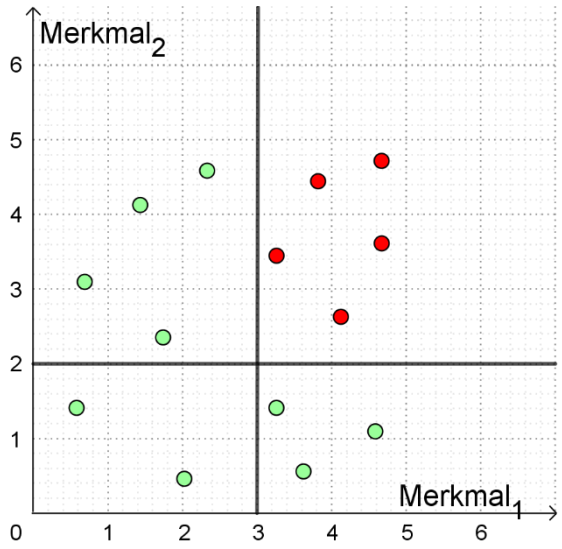
b) Peter lädt offenbar nur Single-Mädchen und nicht gutaussehende Jungs ein. Möglicherweise ist er auf der Suche nach einer Freundin und betrachtet gutaussehende Jungs als Rivalen.

c) Der Lernalgorithmus erzeugt diesen Baum nicht, weil das Merkmal *Junge* für sich genommen keinen hohen Prognosewert hinsichtlich einer Einladung hat: Peter bevorzugt nicht (eindeutig) Jungs bzw. Mädchen. Dass die initiale Aufteilung anhand dieses Merkmals so vorteilhaft ist liegt daran, dass für Jungs und Mädchen unterschiedliche Kriterien eine sehr gute Aufteilung bewirken. Der Lernalgorithmus berücksichtigt jedoch nur Auswirkungen auf die unmittelbar folgende Ebene.

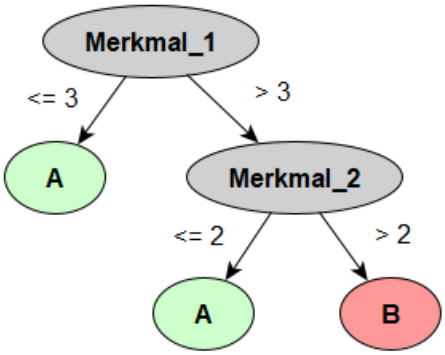
Aufgabe 6

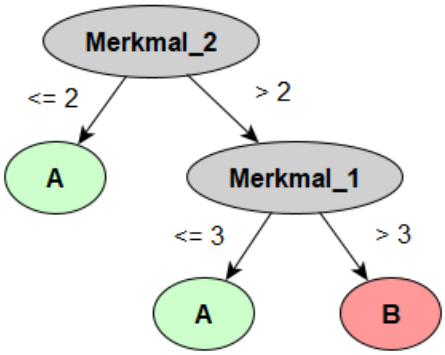
a) Da es keine Duplikate gibt – bei kontinuierlichen Merkmalen wie in der vorliegenden Aufgabe sind solche deutlich seltener als bei diskreten Merkmalen mit nur wenigen Ausprägungen – erzeugt jedes Wertepaar ein eigenes Blatt.
Das Datenelement (1,31; 0,94) kann nicht klassifiziert werden, da diese Merkmalskombination noch nicht vorliegt.

b) Es sind unterschiedliche Lösungen denkbar, die Geraden *Merkmal_1 = 3* bzw. *Merkmal_2 = 2* springen als Trennlinien aber relativ offensichtlich ins Auge:



Je nachdem, ob zuerst anhand von Merkmal_1 oder Merkmal_2 gesplittet wird, entstehen unterschiedliche Bäume. Im Sinne der Minimierung der Fehlklassifikationen sollte zunächst anhand von Merkmal_1 gesplittet werden.

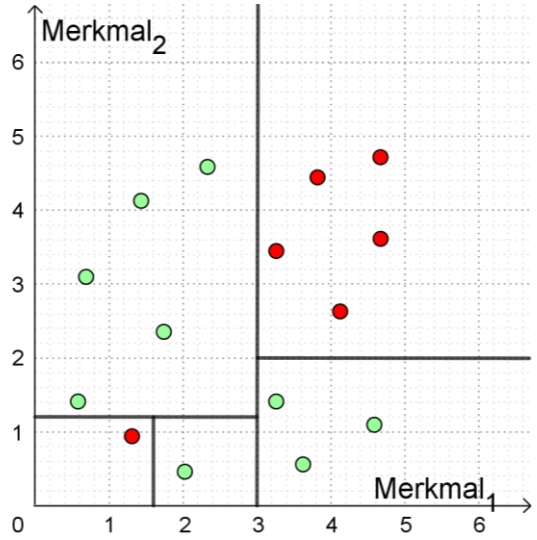




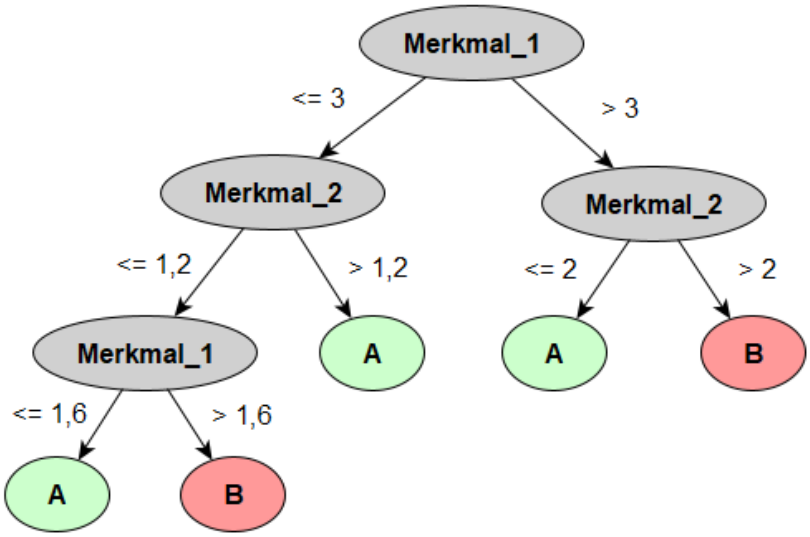
c) In beiden Fällen wird dem Datenelement (1,31; 0,94) die Klasse A zugeordnet.

Aufgabe 7

a) Erneut sind mehrere Lösungen möglich, die folgende Abbildung zeigt ein Beispiel:



Der entsprechende Entscheidungsbaum sieht wie folgt aus:



b) Bei dem Datenelement (1,31; 0,94; B) handelt es sich mit hoher Wahrscheinlichkeit um einen Messfehler bzw. ein verrauschtes Datenelement. Der größere, an diese speziellen Daten angepasste Entscheidungsbaum klassifiziert nun alle Datenelemente, die im unteren linken Rechteck liegen, möglicherweise fälschlich mit B.
Diese Überanpassung an die Trainingsdaten, aus denen der Entscheidungsbaum erzeugt wird, bezeichnet man als *Overfitting*.