

Musterlösung zu Aufgabenblatt: Entscheidungsbäume in Orange3

Aufgabe 1b:

Die Merkmale der Äffchen sind wie folgt modelliert:

- Mund (mit Ausprägungen wie Zahn, Grinsen oder grimmig)
- Auge links (mit den Ausprägungen offen, zu und x)
- Auge rechts (mit den Ausprägungen offen, zu und x)
- Accessoire (mit Ausprägungen wie keins oder Augenbinde)
- beißt (mit den Ausprägungen ja und nein)

Als Zusatzinformation ist jeweils außerdem die Nummer des Äffchens angegeben.

Aufgabe 2a:

Trainingsphase: *Tree-Widget*; Testphase: *Predictions-Widget* und *Confusion Matrix-Widget*

Aufgabe 2b:

Individuell. Unterschiede könnten sich bspw. hinsichtlich der Anzahl der richtig und falsch klassifizierten Äffchen ergeben. Auch bei der Tiefe des Baums, dem Verzweigungsgrad oder den Merkmalen der jeweils falsch klassifizierten Äffchen können sich Unterschiede ergeben.

Aufgabe 2c:

Individuell. Metriken helfen schnell dabei einen Überblick über die Modelle zu bekommen und das bessere Modell für die Aufgabe auszuwählen. Auch das computergenerierte Modell wird nicht alle Äffchen richtig klassifizieren.

Aufgabe 3a:

Der Baum ohne Einschränkungen hat einen höheren Verzweigungsgrad als der binäre Baum und verfügt über mehr Entscheidungsregeln, die die Trainingsdaten noch spezifischer aufteilen.

Aufgabe 3b:

Der nicht-binäre Baum klassifiziert deutlich mehr Testdaten falsch bzw. hat eine deutlich geringere Genauigkeit. Das ist zunächst überraschend, da man aufgrund der zusätzlichen Möglichkeiten, die Zuordnung zwischen Ein- und Ausgabe zu lernen, davon ausgehen könnte, dass dadurch ein genaueres Modell möglich ist.

Aufgabe 3c:

Wie in Aufgabe 3b gesehen kann ein Modell, das über mehr Möglichkeiten verfügt, die Beziehung zwischen Ein- und Ausgabe zu modellieren, zu schlechteren Ergebnissen führen. Dies liegt daran, dass während des Trainingsprozesses eine Überanpassung an die Spezifika der Trainingsdaten erfolgt – der Algorithmus lernt die Trainingsdaten gewissermaßen „auswendig“. Das resultierende Modell kann nicht mehr ausreichend auf die Testdaten verallgemeinern. Obwohl es erstmal nicht intuitiv erscheint, sind daher einfache Modelle in vielen Fällen erfolgreicher in der Vorhersage als deutlich komplexer erscheinende Modelle.

Aufgabe 4:

Wenn der Anteil der Trainings- und Testdaten variiert wird, verändern sich der Entscheidungsbaum und die Konfusionsmatrix.