Hochschule Konstanz
Fakultät Informatik

Data Science

Classification 2 (Mittwoch)

Aufgabe 1 Bestimmung des Geschlechtes von Krabben mit einem Decission Tree

Wir betrachten den *crabs* Datensatz, der im MASS package enthalten ist. Bei dieser Spezies von Krabben (blaue oder orange Leptograpsus variegatus) ist das Geschlecht nicht leicht ersichtlich. Hier wollen wir ein Klassifikationsmodell entwickeln, mit dem das Geschlecht einer Krabbe mithilfe verschiedenen Variablen, die den Körperbau der Krabbe beschreiben, bestimmt werden kann.

a) Erstellen Sie mit der Funktion rpart aus der rpart Library einen Klassifikationsbaum für die Variable Sex. Stellen Sie den Klassifikationsbaum grafisch dar. Quantifizieren Sie die Performance auf dem Trainingsdatensatz durch eine Kofusionsmatrix. R-Hinweise:

```
t1 <- rpart(...)
windows(14,8)  # damit Raender gross genug fuer Beschriftung
plot(t1)
text(t1,use.n=TRUE)
```

b) Im Klassifikationsbaum wurden als Splitvariablen nur die Variablen RW und CL verwendet. Erstellen Sie ein Streudiagramm mit diesen beiden Variablen und färben Sie die Punkte gemäss dem Geschlecht der Krabben ein. Kennzeichnen Sie den ersten Split als Linie im Streudiagramm. Was ist der grosse Nachteil von Klassifikationsbäumen?

c) Führen Sie mit den numerischen Variablen des Crabs Datensatz eine PCA durch und visualisieren Sie die Position der Datenpunkte im Streudiagramm der ersten beiden Hauptkomponenten. Färben Sie die Punkte gemäss dem Geschlecht der Krabben ein.

d) Fügen Sie die Hauptkomponenten als weitere Variablen an den Datensatz Crabs an. Trainieren Sie mit dem erweiterten Datensatz wieder einen Klassifikationsbaum und stellen Sie ihn dar. Welche Variablen wurden als Splitvariaben verwendet? Stellen Sie die Daten im Streudiagramm der beiden häufigsten benutzen Split-Variablen dar. Quantifizieren Sie die Performance auf dem Trainingsdatensatz durch eine Kofusionsmatrix. Was fällt im Vergleich zu den Ergebnissen aus den letzten Teilaufgaben auf? Können Sie sich die Unterschiede erklären?

Aufgabe 2 Churn bei einem Telephonanbieter (Random Forest und Decission Tree)

In dieser Aufgabe wird wieder das Churn-Verhalten (Wechseln zu einem anderen Anbieter) von Telekommunikationskunden untersucht.

a) Laden Sie die Daten (`Churn1.dat`) ein und teilen Sie die Daten in ein zwei Teile. Verwenden Sie die ersten 80 Prozent der Daten als Trainingset und die zweiten 20 Prozent der Daten als Testset. Wie gross ist die Performance eines Decission Tree Klassifikators auf dem Trainings- und Testset? Vergleichen Sie die Ergebnisse mit dem $k = 1$ NN Klassificator. Die Accuracy eines $k = 1$ NN Klassifikators war 1 auf dem Trainingset und etwa 80% auf dem Testset.

b) Wiederholen Sie a) nun mit einem Randomforest und vergleichen Sie die Ergebnisse.

c) Vergleihen Sie nun den out-of-bag error des Trainingsets mit dem Fehler auf dem Testset.

d) Vergleichen Sie die Confusion Matrixen des Decssion Trees und des Random Forest. Sie können diese mit dem Befehl `confusionMatrix` aus dem Package `caret` darstellen. Bei welcher wahren Klasse (churn, no churn) es Ihnen lieber ist, dass der Klassifikator falsch klassifiziert.