Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Prof. Dr. Claus Weihs
Dr. Thomas Liebig
Sebastian Buschjäger
Malte Jastrow
Jennifer Neuhaus-Stern

Übung zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2018

Übungsblatt Nr. 7

Der Abgabetermin ist Dienstag der 19.06.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Logistische Regression) (5 Punkte)

Dieser Aufgabe liegen Daten aus einer zu Beginn der 80er Jahre von der *American Cancer Society* durchgeführten Beobachtungsstudie zum Thema "Rauchen und Lungenkrebs" zugrunde. Insgesamt stehen Informationen für 640 739 Studienteilnehmer zur Verfügung. Im Datensatz `lungcancer.RData` geben die Spalten *FreqLunge* und *FreqSonst* jeweils Aufschluss darüber, wieviele Personen mit den entsprechenden Ausprägungskombinationen der Merkmale

- Geschlecht (0: männlich, 1: weiblich),
- Alter zu Studienbeginn (01.01.1982) in Jahren,
- Ausbildung (0: keine College-Ausbildung, 1: College-Ausbildung) und
- Raucherstatus (0: nie geraucht, 1: Raucher)

innerhalb der Beobachtungszeit von sechs Jahren an Lungenkrebs gestorben sind und wieviele entweder noch leben oder deren Tod eine andere Ursache als Lungenkrebs hatte.

a) Passen Sie ein logistisches Regressions-Modell an die Daten an. Lesen Sie sich dazu die Hilfeseite der Funktion `glm` gründlich durch, um die Funktion richtig einzusetzen! Interpretieren Sie die Modellausgabe. Sind alle Einflussgrößen signifikant?

b) Geben Sie die Odds Ratios für das Modell an. Wie sind diese zu interpretieren?

c) Vergleichen Sie mit Hilfe des aufgestellten Modells einen 63-jährigen Raucher ohne College-Ausbildung mit einer 63-jährigen Nichtraucherin mit College-Ausbildung bezüglich ihres Risikos an Lungenkrebs zu sterben.

Aufgabe 2 (Logistische Regression - Mehrklassenfall) (5 Punkte)

Ein wesentlicher Nachteil der in der Vorlesung vorgestellten logistischen Regression ist, dass lediglich binäre Klassifikationsprobleme behandelt werden können. In der Literatur wurden verschiedene Verfahren vorgeschlagen, um diesen Nachteil zu beheben. Informieren Sie sich im Internet über die folgenden zwei Prinzipien der Verallgemeinerung binärer Klassifikatoren und beschreiben Sie sie jeweils in ca. 5 Sätzen:

a) (2.5 Punkte) One-vs-All

b) (2.5 Punkte) One-vs-One

Aufgabe 3 (Entscheidungsbäume) **(5 Punkte)**

Betrachten Sie den folgenden kleinen Datensatz:

X_1	2	1	4	5	7	3	8	6	10	9
X_2	1	4	3	7	5	2	6	9	8	10
Class	A	A	A	A	A	B	B	B	B	B

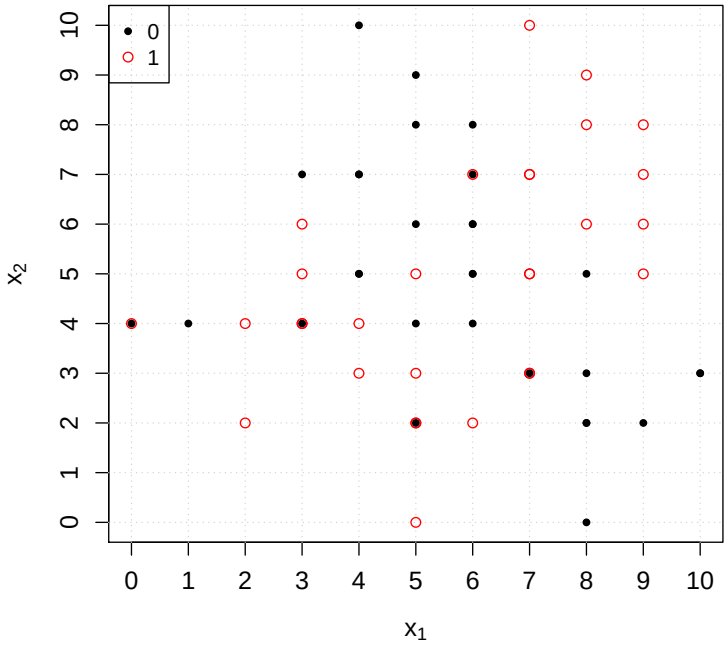
a) (3 Punkte) Wenden Sie den CART-Algorithmus zur Erstellung eines Entscheidungsbaums auf diesen Datensatz an ohne eine entsprechende R-Funktion zu verwenden. Falls an einer Stelle 2 verschiedene Trennungen die gleiche Güte haben, dürfen Sie eine beliebige wählen. Stoppen Sie, wenn ein Knoten rein ist oder er maximal eine *fremde* Beobachtung enthält.

b) (1 Punkt) Zeichnen Sie den Entscheidungsbaum.

c) (1 Punkt) Welchen Klassen werden die Beobachtungen $(7, 7)$, $(1, 8)$ und $(7.5, 2)$ zugeordnet?

Aufgabe 4 (Entscheidungsbäume in R) **(5 Punkte)**

Betrachten Sie die folgende Datensituation (die Werte sind in der Bearbeitungsvorlage angegeben):



a) (2 Punkte) Benutzen Sie die an den CART-Algorithmus angelehnte Implementierung aus dem Paket rpart, um einen Entscheidungsbaum auf den Daten zu lernen. Erzeugen Sie außerdem eine grafische Darstellung des Baumes.

b) (2 Punkte) Zeichnen Sie, z.B. mit der Funktion lines, die resultierenden Entscheidungsgrenzen in das vorgegebene Streudiagramm ein.

c) (1 Punkt) Welchen Klassen werden hier die Beobachtungen $(7, 7)$, $(1, 8)$ und $(7.5, 2)$ zugeordnet? Verwenden Sie die predict-Funktion zur Vorhersage der Klassen.