Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 12.06.12

Abgabe: bis Mi, 20.06., 12.00 Uhr an
schiffner@statistik.tu-dortmund.de
und/oder Briefkasten 146

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 10

Aufgabe 10.1 (4 Punkte)
Auf der Homepage liegen die Datensätze `orange.train.txt` und `orange.test.txt`. Es handelt sich um ein künstliches Klassifikationsproblem mit 2 Klassen (Variable `class`) und zehn erklärenden Variablen `F1` bis `F10`.

a) Versuchen Sie herauszufinden, welche Variablen nützlich sind, um die Klassen zu trennen. Wie Sie dabei vorgehen, ist Ihnen überlassen. Sie können z. B. verschiedene Kennzahlen wie Klassenmittel, Klassenkovarianzen oder die Korrelationen der erklärenden und der Zielvariable etc. berechnen oder Grafiken betrachten wie z. B. eine Scatterplotmatrix oder Histogramme der einzelnen Variablen für die beiden Klassen. Oder Sie passen einen Entscheidungsbaum an die Daten an (R-Funktion `rpart` aus dem Paket `rpart`) und schauen, welche Variablen für die einzelnen Knoten ausgewählt werden. Probieren Sie mindestens eine Methode aus und interpretieren Sie die Ergebnisse.

b) Trainieren Sie einen Random Forest (R-Funktion `randomForest` aus dem Paket `randomForest`) auf dem Trainingsdatensatz. Sagen Sie die Klassenzugehörigkeiten der Beobachtungen im Testdatensatz vorher (R-Funktion `predict`) und berechnen Sie die Fehlklassifikationsrate.

c) Untersuchen Sie nun die Wichtigkeit der erklärenden Variablen mithilfe eines Random Forest, d.h. erstellen Sie Variablen-Wichtigkeits-Plots (Funktion `varImpPlot`), sowohl basierend auf dem Gini-Index als auch auf der OOB-Fehlerrate, und interpretieren Sie diese. Wählen Sie auf Basis dieser plots Variablen aus. Trainieren Sie anschließend einen Random Forest auf den Trainingsdaten (wobei Sie aber nur die ausgewählten Variablen benutzen) und berechnen Sie die Fehlerrate auf den Testdaten.

Aufgabe 10.2 (6 Punkte)
Am 28.1.1986 ist das Raumschiff Challenger kurz nach dem Start explodiert. Als Ursache wurden Probleme mit Dichtungsringen des Treibstofftanks ermittelt. Bereits vor dem Start wurde die Sorge geäußert, dass niedrige Außentemperaturen beim Start Probleme mit den Dichtungsringen begünstigen könnten. Der bisher kälteste Start wurde bei 53° Fahrenheit durchgeführt. Für den 28.1.1986 waren 31° Fahrenheit vorhergesagt.

Der Datensatz `challenger.txt` enthält 2 Variablen – die Temperatur beim Start (temperature) und eine Indikatorvariable (failure), die angibt, ob es Probleme mit den Dichtungsringen gab (1 bedeutet Probleme).

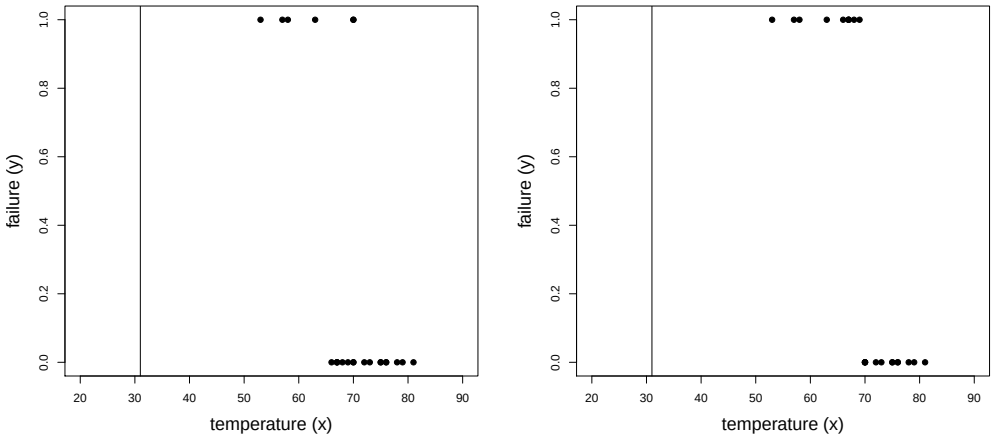

Abbildung 1: links: `challenger`, rechts: `challenger2`

a) Passen Sie ein logistisches Regressionsmodell, das die Wahrscheinlichkeit für Probleme mit den Dichtungsringen in Abhängigkeit von der Temeperatur modelliert, an die `challenger`-Daten an (R-Funktion `glm` mit dem Argument `family = "binomial"`).

b) Sagen Sie die Wahrscheinlichkeit für ein Problem mit den Dichtungsringen bei 31° Fahrenheit vorher. Hätten Sie einen Start empfohlen? (R-Funktion `predict` mit Argument `type = "response"`)

Betrachten Sie nun den leicht veränderten Datensatz `challenger2.txt`.

c) Versuchen Sie ein logistisches Regressionsmodell an die Daten anzupassen. Sie sollten Warnungen erhalten, dass der Algorithmus nicht konvergiert ist und dass angepasste Wahrscheinlichkeiten 0 oder 1 aufgetreten sind.

d) Nun geht es darum, das Phänomen aus c) genauer zu verstehen. Das Problem liegt darin, dass die beiden Klassen im Datensatz `challenger2` vollständig trennbar sind, d.h. es gibt ein $x_0 \in \mathbb{R}$ (hier $x_0 = 70$) mit $y_i = 0$ für $x_i \geq x_0$ und $y_i = 1$ für $x_i < x_0$.

Bei der Anpassung des logistischen Regressionsmodells an die Daten wird die log-likelihood

$$\ln L(\beta) = \sum_{i=1}^{n} (y_i \ln(\pi_{i1}) + (1 - y_i) \ln(1 - \pi_{i1}))$$

maximiert. Im Fall einer einzigen Einflussvariable X sind $\beta' = (\beta_0, \beta_1)$ und $\pi_{i1} = \exp(\beta_0 + x_i\beta_1)/(1+\exp(\beta_0+x_i\beta_1))$. Nehmen Sie der Einfachheit halber an, dass $x_1 \leq x_2 \leq \ldots \leq x_n$ gilt. Im Falle vollständiger Trennbarkeit sind dann $y_1 = \ldots = y_m = 1$ und $y_{m+1} = \ldots = y_n = 0$ für ein $m \in \{1, \ldots, n-1\}$ und $x_m < x_{m+1}$.

Zeigen Sie: Im Fall vollständiger Trennbarkeit existiert der ML-Schätzer für β nicht.

Hier ein paar Tipps zur Vorgehensweise:

- Leiten Sie $\ln L(\beta)$ nach β_0 und β_1 ab. Stellen Sie die Ableitungen in Abhängigkeit der π_{i1} dar und setzen Sie die y_i-Werte ein.
- Setzen Sie zunächst $\frac{\partial L(\beta)}{\partial \beta_0}$ gleich 0.
- Zeigen Sie, dass dann $\frac{\partial L(\beta)}{\partial \beta_1} < 0$ ist. Nutzen Sie dabei aus, dass $x_1 \leq x_2 \leq \ldots \leq x_n$ und $x_m < x_{m+1}$ ist.