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

Dortmund, 15.05.12

Abgabe: bis Mi, 23.05., 12.00 Uhr an
schiffner@statistik.tu-dortmund.de
bzw. in Briefkasten 146

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 6

Aufgabe 6.1 (4 Punkte)

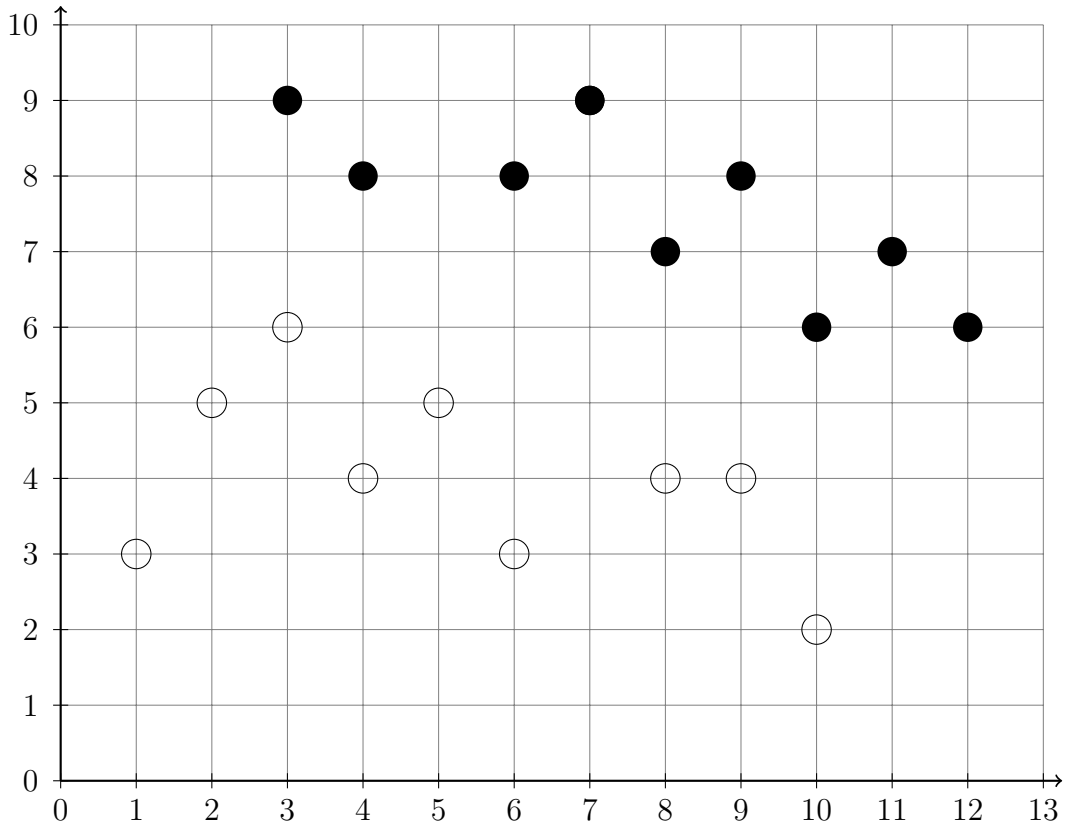


Abbildung 1: Datenpunkte im $\mathbb{R}^2$

a) Im Fall der Linearen Diskriminanzanalyse sind die sich ergebenden Entscheidungsgrenzen Hyperebenen im $\mathbb{R}^p$. Vollziehen Sie zunächst die Umformungen auf S. 186 im Skript nach und bringen Sie die Ebenengleichung anschließend in Hesse Normalform $< \boldsymbol{\beta}, \boldsymbol{x} > +\beta_0 = 0$ mit $\|\boldsymbol{\beta}\| = 1$. Wie groß ist der Abstand β_0 der Hyperebene vom Ursprung?

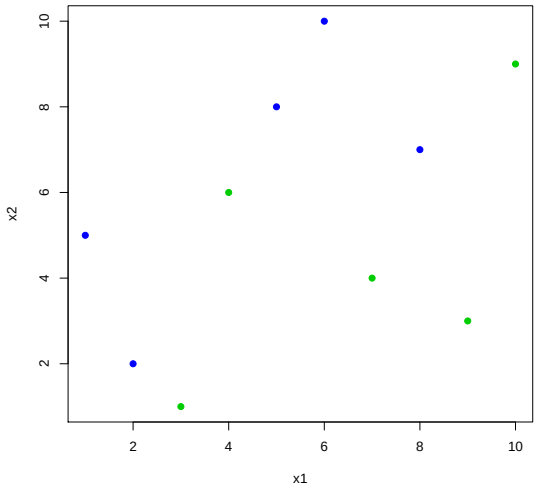
Betrachten Sie den Datensatz in der Datei `daten_a61.txt` auf der Homepage (vgl. Abbildung 1). Er enthält 20 Beobachtungen aus 2 Klassen (-1 und +1).

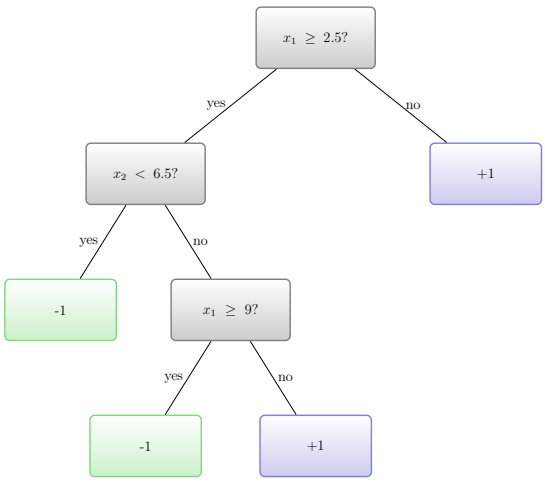
b) Berechnen Sie die trennende Hyperebene für den Beispieldatensatz und zeichnen Sie die Gerade in den Scatterplot ein. (Sie können entweder den Plot selbst per Hand oder mit der Software Ihrer Wahl erstellen. Oder Sie zeichnen die Trenngerade einfach in Abb. 1 mit ein und werfen das Ergebnis in den Briefkasten.)

c) Zu welchen Klassen werden die folgenden Punkte zugeordnet?

$$(4, 6), (7, 6), (12, 4), (-1, 8), (-4, 11)$$

Aufgabe 6.2 (3 Punkte)
Für das links abgebildete 2-Klassen Problem mit den Klassen -1 (grün) und +1 (blau) ergibt sich der rechts abgebildete Entscheidungsbaum.





a) Zeichnen Sie die zum Baum gehörige Entscheidungsgrenze in den Scatterplot mit ein. (Es gilt das gleiche wie in Aufgabe 6.1 c). Die Daten finden Sie in der Datei `daten_a62.txt`.)

b) Berechnen Sie für jeden split im Entscheidungsbaum die Verringerung der Gini-Unreinheit.

Aufgabe 6.3 (4 Punkte)
Zeit für die praktische Anwendung von Klassifikationsverfahren in *RapidMiner*!

Ein Datensatz namens *Sonar* ist bereits im Samples Repository von RapidMiner vorhanden. Es handelt sich um ein Klassifikationsproblem mit 2 Klassen. Infos zum Datensatz finden Sie hier: `http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+\%28Sonar,+Mines+vs.+Rocks\%29`.

Probieren Sie Lineare Diskriminanzanalyse und 3 weitere Klassifikationsverfahren Ihrer Wahl auf den Sonar Daten aus. Splitten Sie den Datensatz im Verhältnis 7:3 zufällig in einen Trainings- und Testdatensatz auf. Nutzen Sie den Trainingsdatensatz, um den jeweiligen Klassifikator an die Daten anzupassen, und den Testdatensatz zur Beurteilung der Vorhersagegüte.

Die Aufteilung in Trainings- und Testdatensatz erreichen Sie mithilfe des *Validation*-Operators. Weitere nützliche Operatoren sind *Multiply*, *ApplyModel* und *Performance*.

Erstellen Sie ein RapidMiner-Experiment, das für die 4 Verfahren die Vorhersagegüte berechnet, und beantworten Sie die folgenden Fragen:

- Wie groß ist die accuracy für die einzelnen Verfahren und welches Verfahren hat die größte/kleinste Vorhersagegüte?
- Welches Verfahren hat die größte/kleinste Vorhersagegüte für Klasse Rock?
- Welches Verfahren hat die größte/kleinste Vorhersagegüte für Klasse Mine?
- Sind die beiden Klassen gut linear trennbar?