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

Dortmund, 05.06.12
Abgabe: bis Mi, 13.06., 12.00 Uhr an
friedrichs@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 9

Aufgabe 9.1 (4 Punkte)
In den bisherigen Vorlesungen wurde die SVM zur Lösung binärer Klassifikationsprobleme verwendet, d.h. das Label y war üblicherweise aus der Menge {-1, +1}. Häufig findet man in Anwendungen jedoch das Problem vor, dass die Menge Y der Klassen größer ist, z.B. bei der Sortierung von Texten in mehrere Kategorien.

a) Überlegen Sie sich, welche der bisher behandelten Lernverfahren direkt für Probleme mit mehreren Klassen angewandt werden können und begründen Sie kurz anhand von 2 Beispielen ihre Antwort.

b) Überlegen Sie sich ein Verfahren um auch die SVM für Mehrklassenprobleme zu nutzen und beschreiben Sie ihren Ansatz.

c) Erstellen Sie ein RapidMiner-Experiment, welches die SVM zur Klassifizierung der Iris-Daten benutzt.
Hinweis: Suchen Sie nach einem geeigneten Operator!

Aufgabe 9.2 (2 Punkte)
In der Vorlesung haben Sie für die Textkategorisierung das TCat-Modell kennengelernt.

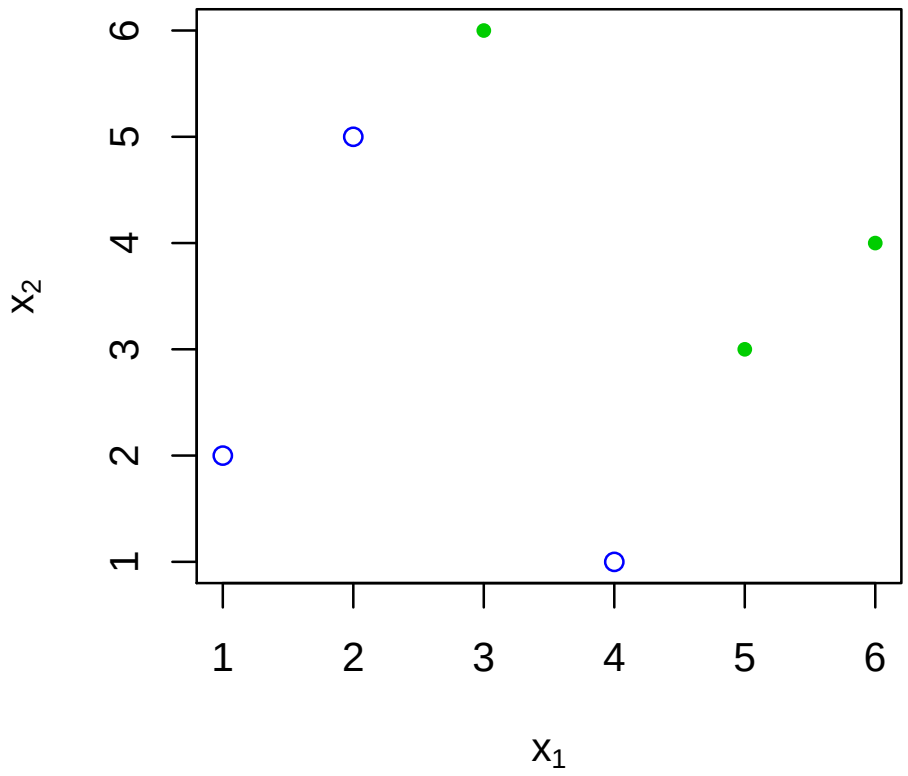
a) Erläutern Sie kurz mit eigenen Worten die Eigenschaften von Textkollektionen bzgl. der Klassifikation.

b) In der Vorlesung wurden auch andere Anwendungen für das TCat-Modell vorgestellt. Überlegen Sie sich eine weitere Anwendung, für die man das Konzept verwenden kann. Begründen Sie ihren Vorschlag!

Aufgabe 9.3 (4 Punkte)
Führen Sie den AdaBoost-Algorithmus aus der Vorlesung „per Hand“ (d.h. entweder mit Bleistift und Papier oder mithilfe eines selbst geschriebenen Programms) durch. Auf der Homepage liegt der Datensatz `daten_a93.txt`. Es handelt sich um ein zweidimensionales Klassifikationsproblem mit 2 Klassen -1 und $+1$. Pro Klasse sind 3 Trainingsbeobachtungen vorhanden. Als Basisklassifikatoren sollen sogenannte decision stumps verwendet werden. Dies sind Entscheidungsbäume, die nur genau einen Split machen (in unserem zweidimensionalen Problem also entweder waagerecht oder senkrecht) und somit Entscheidungsregeln der Form „wähle Klasse -1, wenn $x_2 < 3$“ produzieren. Der Basisklassifikator wählt unter allen möglichen Splits denjenigen aus, bei dem die Summe der Gewichte aller falsch klassifizierten Beobachtungen minimal wird. Gibt es mehrere mögliche Splits, wählen sie den zuerst gefundenen aus. Beginnen Sie mit vertikalen Splits und arbeiten Sie mit aufsteigenden Schwellwerten.

a) Berechnen Sie die ersten drei AdaBoost-Iterationen.

b) Berechnen Sie die Gesamtentscheidungsregel, d.h. generieren Sie aus den in a) berechneten decision stumps einen Entscheidungsbaum.



Hinweis: Man muss nicht unbedingt alle Splits berechnen!