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

Dortmund, 19.06.12
Abgabe: bis Mi, 27.06., 12.00 Uhr an
friedrichs@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 11

Aufgabe 11.1 (6 Punkte)
In dieser Aufgabe sollen die in der Vorlesung vorgestellten Methoden zur Fehlerratenschätzung miteinander verglichen werden. Dazu sollen Sie wie in Übungsblatt 6 erneut den Datensatz *Sonar* und die Lineare Diskriminanzanalyse verwenden.

a) Erstellen Sie einen RapidMiner Prozess, der die folgenden Methoden zur Fehlerratenschätzung durchführt:

   1) *Train-and-test Methode* (Teilen Sie hierfür den Datensatz im Verhältnis 3:1 in einen Trainings- und einen Testdatensatz auf)
   2) *Leave-one-out Methode*
   3) *10-fache Kreuzvalidierung*
   4) *e0-Bootstrap Schätzer* (mit $B = 200$)
   5) *.632-Bootstrap Schätzer* (mit $B = 200$)
   6) *Repeated Cross Validation* (mit $R = k = 10$)
   7) *Bootstrap Cross Validation* (mit $B = k = 10$)

b) Wie hoch sind die jeweiligen Fehlerraten?

c) Welche Methode halten Sie für das gegebene Problem am geeignetsten? Begründen Sie ihre Meinung!

d) Unter welchen Umständen wären andere Methoden, als die in c) genannte, sinnvoller?

Hinweis: Nicht alle der oben genannten Methoden sind in RapidMiner direkt vorhanden, für diese müssen sie geeignete Operatoren miteinander verknüpfen. Beispielsweise könnte einer der *Loop*-Operatoren nützlich sein.

Aufgabe 11.2 (4 Punkte)
In dieser Aufgabe soll noch einmal das bereits in der Vorlesung vorgestellte Verfahren *SVMstruct* betrachtet werden.

a) Welche Aufgabe löst die *SVMstruct*?

b) Erläutern Sie die grundlegende Idee des *SVMstruct*-Ansatzes!

c) Betrachten Sie nun die folgende Abbildung. Es sind Beziehungen zwischen Nutzergruppen des sozialen Netzwerkes *facebook* dargestellt. Die Betreiber haben vier Nutzergruppen definiert, die sich durch bestimmte Beziehungen untereinander und gewissen eigene Attributen auszeichnen. Die eigenen Attribute sind in Tabelle 1 angegeben.
Stellen Sie den Ψ-Vektor auf, um alle Informationen, die die *facebook*-Betreiber zusammengestellt haben, zu codieren!

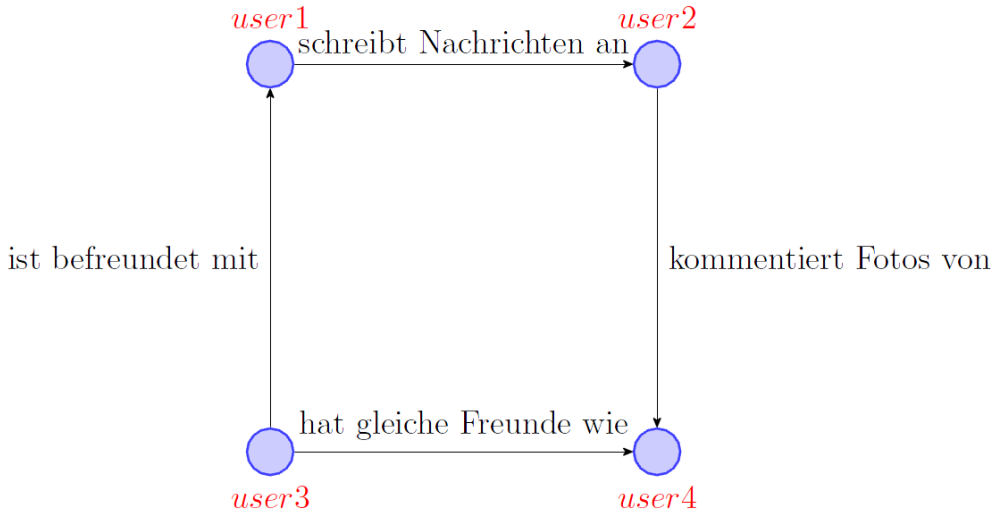

Tabelle 1: Eigenschaften der vier Nutzergruppen von *facebook*

Nutzergruppe	Eigenschaften
user1	hat viele Freunde und kommentiert viel
user2	spielt Spiele und hat wenige Freunde
user3	kommentiert nie
user4	hat viele Fotos