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

Dortmund, 28.05.12
Abgabe: bis Mi, 06.06., 12.00 Uhr an
ibenkhediri@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 8

Aufgabe 8.1 (3 Punkte)
In der Vorlesung wurde zur Optimierung der SVM-Parameter das Verfahren SMO vorgestellt.

1. Warum optimiert man hierbei nicht alle Parameter gleichzeitig, und wie viele Parameter werden statt dessen gleichzeitig optimiert und warum?
2. Gehen Sie davon aus, dass $C = 0.1$ ist. Welchen Wert können die SVM-Parameter minimal und maximal annehmen?

   Lösen Sie diese Aufgabe zeichnerisch und leiten Sie Ihre Lösung daraus ab. Gehen Sie einmal von $y_1 = y_2$ und dann von $y_1 \neq y_2$ aus.

Aufgabe 8.2 (2 Punkte)

Die Regressions-SVM wurde für Zeitreihen vorgestellt. Bei Zeitreihen kann es zu Trends oder Zyklen kommen, die zu beachten sind. Geben zwei Beispiele fur Datenreihen, die Zyklen oder Trends enthalten.

Aufgabe 8.3 (5 Punkte)
Das Training einer SVM enthält eine Menge Parameter, die es sorgsam auszuwählen gilt. Beispielsweise hat der Parameter C für die Straf-Gewichte einen enormen Einfluss auf das Modell. In dieser Aufgabe sollen Sie sich etwas damit besch äftigen und den Einfluss der Parameter ausprobieren.

Auf der Homepage liegt der Spam Datensatz. Hier soll nun mit Hilfe der SVM eine Klassifikation die unerwünschte Werbe-E-mails erfolgen. Als Abgabe wird zu jeder Aufgabe jeweils ein RapidMiner-Prozess erwartet. Zudem soll ein PDF mit abgegeben werden, dass für die beiden Aufgaben jeweils kurz das Vorgehen und die Ergebnisse/ Erkenntnisse dokumentiert.

1. Erstellen Sie einen RapidMiner Prozess, der ein SVM-Modell auf den Daten lernt.

2. Testen Sie für einen linearen Kernel zunächst verschiedene Werte für C. Testen Sie danach auch eine polynomielle Kernel-Funktion. Finden Sie eine gute Kombination von Parametern?

3. Erstellen Sie ein weiteren RapidMiner Prozess, der die GridParameterOptimization nutzt um einen guten Satz von Parametern für die SVM zu finden, der das beste Ergebnis liefert. Ein Beispiel für einen derartigen Prozess finden Sie auch im samples Repository von RapidMiner.

4. Überlegen Sie sich, an welche Stelle in den Prozess sie eine Kreuzvalidierung einbinden würden um die Parameterwahl robuster zu machen. Erweitern Sie ihren Prozess um eine Kreuzvalidierung.