Universität Dortmund SoSe 2004
Übungen zu *Wissensentdeckung in Datenbanken*

Blatt 5. Abgabe bis Montag, den 31.5.2004

Teil der Aufgabenstellung dieses Übungsblattes ist es, sich in die Experimentierumgebung YALE einzuarbeiten. Yale kann als freie Software unter `http://yale.cs.uni-dortmund.de` bezogen werden; es befinden sich Beispieldateien bei der Distribution. Außerdem ist Yale auf den Rechnern in den Räumen 009 und 010 im Gebäude GB V am Campus Süd installiert, für die Sie Accounts erhalten haben. Nach Einloggen dort geben Sie bitte ein:

```
module add yale
```

und anschließend:

```
YaleGUI &
```

Damit startet die graphische Benutzeroberfläche von YALE. Hinweise zur Benutzung von YALE finden sich auf der Webseite der Vorlesung (Datei `yale_einfuehrung.pdf`) und auf den Yale-Seiten. Weiterhin enthält die Datei `Daten/yale_samples.zip` auf der Vorlesungswebseite Beispiele für Yale-Experimente. Diese Beispiele enthalten die Grundkonstruktionen, die für die Lösung der folgenden Aufgaben benötigt werden.

Zusätzlich werden sich Michael Wurst und Timm Euler am Dienstag, den 25.5.04 in der Zeit von 13 bis 15 Uhr in den Rechnerpools im GB V aufhalten, um Hilfestellung bei der Benutzung von Yale zu geben.

Zur Lösung der folgenden Aufgaben verwenden Sie bitte den in Yale bereitgestellten SVM-Operator *JMySVMLearner*. Beachten Sie, dass die α-Werte, die dieser Operator als Resultat ausgibt, bereits mit dem Label y (-1 oder 1) des jeweiligen Beispiels multipliziert sind. Hinweise: Wenn Sie diesen Operator in einer Kette einsetzen, können Sie seine Ausgabe inspizieren, indem Sie einen Breakpoint nach diesem Operator setzen und nach Ausführung des Experiments bis zum Breakpoint auf "Results" klicken. Die Ausgabewerte des Operators sollten evtl. gerundet werden.

Auf der Webseite der Vorlesung finden Sie die Daten für dieses Übungsblatt als Zip-Datei. Nach dem Auspacken gibt es für die Aufgaben 1 bis 3 jeweils eine Datei für die Daten (`aufgabeX_daten.txt`) und eine Attributbeschreibungsdatei (`aufgabeX_attribs.xml`). Die letztere ist als Eingabeparameter des Yale-Operators *ExampleSource* geeignet.

Aufgabe 1 Wenden Sie die SVM auf die Daten zu dieser Aufgabe an. Setzen Sie dabei den Parameter C auf 1.0. Berechnen Sie aus der SVM-Ausgabe die Gleichung einer Geraden, die die Punkte trennt. Zeichnen Sie die Datenpunkte sowie die trennende Gerade in ein Koordinatensystem ein.

Aufgabe 2 Wenden Sie die SVM auf die Daten zu dieser Aufgabe an; verwenden Sie den linearen Kernel ("dot"). Setzen Sie diesmal den Parameter C auf 0.5. Werden alle Datenpunkte (Beispiele) richtig klassifiziert? Wenn nicht, welche werden falsch klassifiziert? Wie muss der Parameter C verändert werden, damit alle Beispiele richtig klassifiziert werden, und warum?

Aufgabe 3 Wenden Sie die SVM auf die Daten zu dieser Aufgabe an. Sind die Datenpunkte linear trennbar? Wenn nicht, welcher Kernel bietet sich an? Verwenden Sie die Parameteroptimierung von Yale, um eine günstige Einstellung des Parameters C zu finden. Setzen Sie dabei Kreuzvalidierung mit Leave-one-out ein. Als Leistungskriterium bei der Optimierung ist *absolute* zu wählen.

Geben Sie bitte, zusätzlich zu den Antworten auf die Fragen, Ihre Yale-Experimentdateien (als XML-Dateien) mit ab.