Universität Dortmund SoSe 2004
Übungen zu *Wissensentdeckung in Datenbanken*

Blatt 6. Abgabe bis Montag, den 7.6.2004

Da bei diesem Übungsblatt wieder Yale benutzt werden sollte, werden sich Michael Wurst und Timm Euler **am Dienstag, den 1.6.04 in der Zeit von 13 bis 15 Uhr** wieder in den Rechnerpools im GB V aufhalten, um Hilfestellung bei der Benutzung von Yale zu geben.

Aufgabe 1 Gegeben seien $\vec{x}_1, \ldots, \vec{x}_6 \in \mathbb{R}^2$ wie folgt: $\vec{x}_1 = (0.5, 1)$, $\vec{x}_2 = (1, 2)$, $\vec{x}_3 = (-1, 3)$, $\vec{x}_4 = (-2, 2)$, $\vec{x}_5 = (2, 2)$, $\vec{x}_6 = (-1, -0.5)$. Dabei seien $\vec{x}_1$ bis $\vec{x}_3$ positiv klassifiziert und $\vec{x}_4$ bis $\vec{x}_6$ negativ. Offensichtlich sind diese Punkte nicht linear trennbar.
Gegeben sei nun die Funktion $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^6$ mit

$$\Phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} \cdot x_1 x_2 \\ \sqrt{2} \cdot x_1 \\ \sqrt{2} \cdot x_2 \\ 1 \end{pmatrix}$$

1. Bilden Sie die Punkte $\vec{x}_1$ bis $\vec{x}_6$ mit Hilfe von Φ in den $\mathbb{R}^6$ ab und zeigen Sie, dass die Punkte dort linear trennbar sind, indem Sie eine SVM mit linearem Kernel darauf trainieren und Fehlerfreiheit des gelernten Modells nachweisen. Sie können dafür Yale benutzen.
2. Zeigen Sie, dass die Funktion $K : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ mit $K(\vec{v}, \vec{w}) = (\vec{v} * \vec{w} + 1)^2$ die zu Φ gehörige Kernfunktion ist. Dabei bezeichne "$*$" das Skalarprodukt im $\mathbb{R}^n$.

Aufgabe 2 In der zweiten Aufgabe geht es um ein Lernproblem, bei dem grundsätzlich sehr viele Attribute im Spiel sind: Textklassifikation. Yale stellt den Operator *WVTool* zur Verfügung, mit dem sich Texte direkt in Eingabedaten, also ExampleSets, umwandeln lassen. Die Datei `wvtool_example.xml` enthält ein Yale-Beispielexperiment, das die Be-

nutzung dieses Operators demonstriert. Der wichtigste Parameter ist "texts", in welchem für jede Klasse ein Verzeichnis angegeben wird, in dem die zugehörigen Texte liegen. Der zweite wichtige Parameter ist "vectorcreation". Er entscheidet darüber, wie die Texte in Vektoren umgewandelt werden. Es gibt dabei vier Möglichkeiten:

- edu.udo.cs.wvtool.generic.vectorcreation.BinaryOccurrences: Binäre Darstellung
- edu.udo.cs.wvtool.generic.vectorcreation.TermOccurrences: Absolute Häufigkeiten
- edu.udo.cs.wvtool.generic.vectorcreation.TermFrequency: Häufigkeiten normiert auf Länge
- edu.udo.cs.wvtool.generic.vectorcreation.TFIDF: Häufigkeiten normiert auf Länge und inverse Dokumentenhäufigkeiten

a. Die erste Aufgabe ist nun folgende: Laden Sie vom Vorlesungsserver die Datei `blatt6_texte.zip` herunter und entpacken Sie diese. Schreiben sie eine Yalekette, die mit 10facher Kreuzvalidierung die Leistung der Lerner:

- JMySVM
- NearestNeighbors (Parameter "distance_measure" = edu.udo.cs.yale.valueseries.measures.SimpleEuclidianDistance)
- WekaLeaener (Parameter "weka_learner_name" = weka.classifiers.trees.J48)

unter der Verwendung verschiedener Vektorisierungen (s.o.) testet. In welchen Fällen verbessert TF(/IDF) das Ergebnis? Welche Lerner schneiden am besten ab? Führen Sie empirische Tests auf den Daten durch und begründen Sie das Ergebnis qualitativ (Tipp: Welche Verfahren führen Merksmalsgewichtung durch?)
Zusatzaufgabe: Nach diesen Tests soll das Problem nun auch noch bezüglich von TCat untersucht werden. Was ist eine obere Schranke für den Fehler durch die SVM? Benutzen Sie dazu die Formeln, die in der Vorlesung vorgestellt wurden, und den *ExampleSetWriter* von Yale, mit dem Sie die gesamten Beispieldaten in eine externe Datei umwandeln können. Diese kann mit einem beliebigen externen Programm

verarbeitet werden. In der Lösung müssen allerdings die TCat-Klassen angegeben werden, welche aus den Daten erzeugt wurden, sowie die obere Schranke, die sich daraus ergibt.
Wichtiger Hinweis: Sollten Sie die Übungen nicht in den Rechnerpools in GB V durchühren, dann müssen Sie die neueste Version des wvtool-plugins von
`http://www-ai.cs.uni-dortmund.de/SOFTWARE/YALE/download.html`
herunterladen und in das Verzeichnis `<yale_home>/lib/plugins` kopieren.