Prof. Dr. Katharina Morik, Prof. Dr. Claus Weihs
Gero Szepannek, Martin Scholz

Dortmund, 22. Juni
Abgabe: bis 28.6., 20:00 Uhr an
scholz@ls8.cs.uni-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2006

Blatt 10

Aufgabe 10.1

Bei dieser Aufgabe sollen Texte in für Klassifikationsaufgaben geeignete Repräsentationen überführt werden. Beziehen Sie dazu zunächst die folgenden Publikationen aus geeigneter Quelle, beispielsweise aus dem Internet. Alle Artikel sind online frei verfügbar.

- H. Manilla, H. Toivonen, A. I. Verkamo (1997). *Discovery of frequent episodes in event sequences* (Technical Report C-1997-15). University of Helsinki, Finland.
- F. Höppner (2001). Discovery of Temporal Patterns - Learning Rules about the Qualitative Behaviour of Time Series. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Lecture Notes in Artificial Intelligence 2168, Springer.
- J.-F. Boulicaut, A. Bykowski, C. Rigotti (2003). Free-Sets: A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. In *Data Mining and Knowledge Discovery*, 7(1): 5-22.
- B. Goethals (2003). *Survey on frequent pattern mining*, Technical Report.

(a) Überführen Sie die ersten drei Sätze der Abstracts dieser Artikel in Wortvektoren, bei (Goethals, 2003) verwenden Sie bitte die ersten drei Sätze des Abschnitts "Introduction". Der Einfachheit halber beschränken Sie sich dabei bitte auf Substantive nach geeigneter Normierung in Stammformen.

(b) Berechnen Sie für das erste Dokument die Euklidsche Länge des Wortvektors.

(c) Berechnen Sie die IDF Werte der Dokumente, sowie einen beliebigen TFIDF Wert größer 0.

(d) Betrachten Sie die ersten beiden Publikationen als positive Beispiele des Zielkonzepts "Data Mining unter Berücksichtigung zeitlicher Aspekte", die letzten beiden als negative Beispiele. Berechnen Sie die odds ratios der 3 häufigsten Terme, die in beiden Konzepten mindestens einmal vorkommen.

Aufgabe 10.2

In der Vorlesung wurde die strukturelle Risikominimierung vorgestellt (Folien-SVM2, Folie 31 ff.), die dazu dient, die Ausdrucksstärke von Modellklassen zu messen. Sie basiert auf der in den Folien als η notierten VC Dimension.

(a) Das empirische Risiko zweier jeweils bester Modelle aus unterschiedlichen Modellklassen sei identisch, die erste Modellklasse besitze endliche, die zweite unendliche VC Dimension. Welches Modell empfiehlt sich hinsichtlich der strukturellen Risikominimierung?

(b) Wie groß ist die VC Dimension der Modellklasse unbeschränkt tiefer Entscheidungsbäume für binäre Klassifikationsprobleme bei mindestens einem kontinuierlichen Merkmal? Begründen Sie bitte Ihre Ansicht.

(c) Wir betrachten die Modellklasse der Kreise im $\mathbb{R}^2$ (für Zweiklassenprobleme). Die Klasse besitzt die Parameter "Mittelpunkt" und "Radius". Jedes Modell klassifiziert genau die Punkte innerhalb des Kreises als positiv. Beweisen Sie, zum Beispiel graphisch, dass diese Hypothesenklasse eine VC Dimension von mindestens 3 besitzt.

Aufgabe 10.3

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 $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$ mit

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

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.