10. Übung „Knowledge Discovery“

Sommersemester 2006

1 SVM

Gegeben sei folgender zweidimensionaler Trainingsdatensatz:

$$S = \{(x_i, y_i)\} = \{(2, 0; -1), (0, 2; -1), (2, 2; 1), (3, 2; 1)\}$$

1. Bestimmen Sie die den Abstand d der optimalen Hyperebene zum gegebenen Trainingsdatensatz.
2. Die Entscheidungsfunktion des linearen Klassifizierers sei wie im Skript angegeben:

$$f(x) = \text{sgn}(w * x + b)$$

Berechnen Sie die Gewichte w und den Schwellwert b .

3. Welche Trainingsbeispiele sind die Supportvektoren?
4. Klassifizieren Sie das Beispiel $(1, 4)$.

2 Preprocessing und Datenbeschaffenheit

1. Diskutieren Sie das Verhältnis von Preprocessing und Datenverständnis.
2. In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Nennen Sie die drei wichtigsten Skalenniveaus und beschreiben Sie sie kurz; geben Sie jeweils ein Beispiel an. Was bedingt ein Skalenniveau bei der Untersuchung von Daten?
3. Welches Ziel wird in Bezug auf die spätere Anwendung von Algorithmen mit dem Preprocessing der Daten im Data Mining verfolgt. Nennen Sie in diesem Zusammenhang zwei Beispiele, in denen Algorithmen bestimmte Preprocessingsschritte erzwingen.
4. Nennen Sie je fünf Preprocessingsschritte zu Data Cleansing und Data Manipulation und beschreiben Sie drei davon genauer.