Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

29.1.2009

# 11. Übung „Knowledge Discovery“

Wintersemester 2008/2009

## 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 gegeben Trainingsdatasatz.
2. Die Entscheidungsfunktion des linearen Klassifizieres sei wie im Skript angegeben:

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

   Ermitteln Sie (mittels „scharfem Hinschauen“) die Gewichte $w$ und den Schwellwert $b$.
3. Welche Trainingsbeispiele sind die Supportvektoren?
4. Klassifizieren Sie das Beispiel $(1, 4)$.

## 2 SVM

1. Was sagt ihre Intuition: Warum ist ein großer Abstand $d$ gut?
2. Überlegen Sie sich ein Beispiel für einen Datensatz, für den eine (lineare) SVM neue Daten gut klassifizieren können wird, bei dem der $k$NN-Algorithmus jedoch versagen wird. Finden Sie auch ein Beispiel für den umgekehrten Fall.
3. Interpretieren Sie die geometrische Bedeutung der Konstante $C$ in der Problembeschreibung

   $$\min \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

   für die weich trennende Hyperebene.
4. Eine weich trennende Hyperebene erlaubt Ausreißer und kann daher, selbst wenn die Trainings-Menge nicht linear trennbar ist, zum Lernen eines Klassifikators genutzt werden. Geben Sie ein Beispiel an, bei dem dieser Ansatz nicht funktioniert. Wie kann man solche Probleme dennoch mit einer SVM lösen?

## 3 Nachschlag: Entscheidungsbäume

Was könnte – neben dem Informationsgewinn oder Gini-Index – ein brauchbares Maß sein, um passende Attribute für Splits zu finden? Überlegen Sie sich ein eigenes Maß. Welche Vor- bzw. Nachteile gegenüber dem Informationsgewinn hätte dieses?