Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

29.1.2009

11. Übung „Knowledge Discovery“

Wintersemester 2008/2009

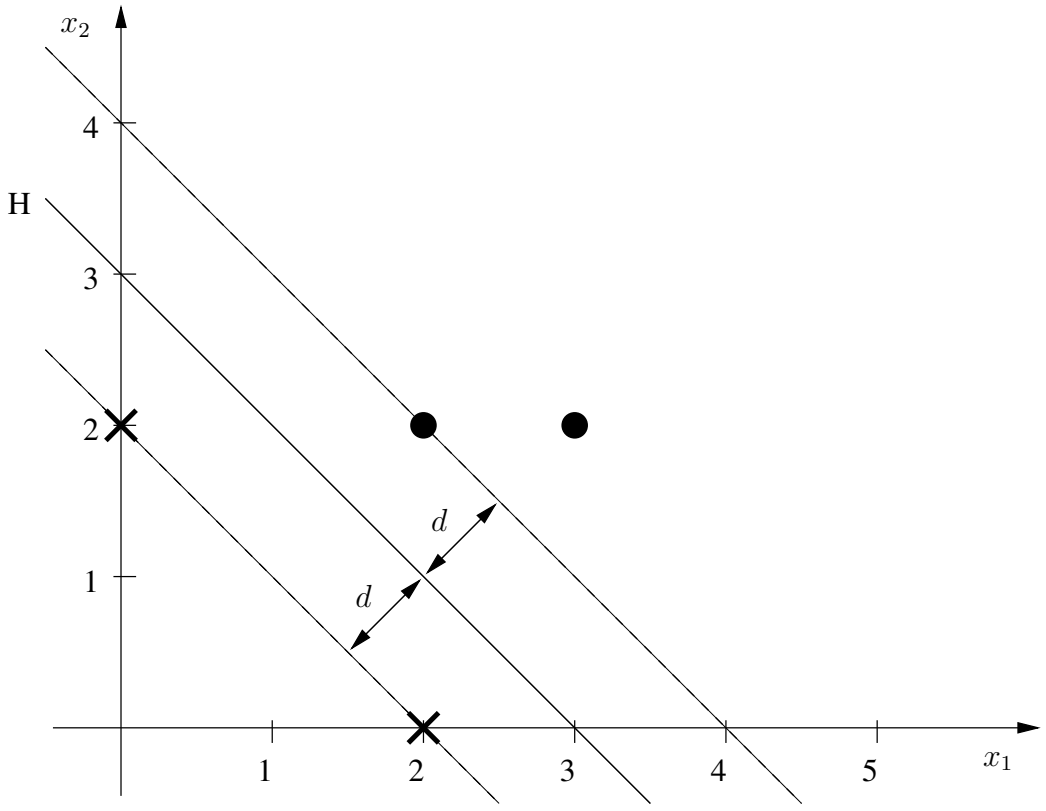
1 SVM

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.

Lösung der Aufgabe durch Zeichnung und „scharfes Hinschauen“:



Ergibt $d = \frac{1}{\sqrt{2}}$.

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

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

Ermitteln Sie (mittels „scharfem Hinschauen“) die Gewichte w und den Schwellwert b.

$$w = \begin{pmatrix}1\\1\end{pmatrix}, b = -3$$

3. Welche Trainingsbeispiele sind die Supportvektoren?

Die Datensätze $\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}$.

4. Klassifizieren Sie das Beispiel $(1, 4)$.

$$f\left(\begin{pmatrix}1\\4\end{pmatrix}\right) = \text{sgn}\left(w * \begin{pmatrix}1\\4\end{pmatrix} + b\right) = \text{sgn}\left(\begin{pmatrix}1\\1\end{pmatrix} * \begin{pmatrix}1\\4\end{pmatrix} - 3\right) = \text{sgn}(1 + 4 - 3) = 1$$

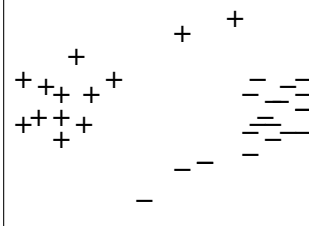
2 SVM

1. Was sagt ihre Intuition: Warum ist ein großer Abstand d gut?

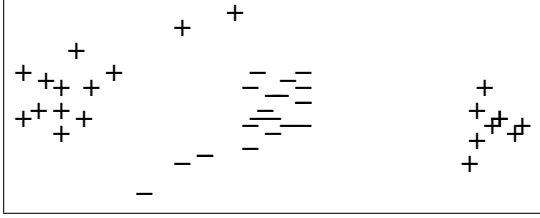
Er erhöht die „Trennschärfe“ der SVM – Ausreißer haben mehr „Platz zum Ausreißen“und die Gefahr einer Falschklassifikation ist niedriger.

2. Überlegen Sie sich ein Beispiel für einen Datensatz, für den eine (lineare) SVM neue Daten gut klassifizieren können wird, bei dem der kNN-Algorithmus jedoch versagen wird. Finden Sie auch ein Beispiel für den umgekehrten Fall.

Diesen Datensatz kann eine SVM gut lernen und neue Instanzen korrekt klassifizieren:



Folgender Datensatz ist dagegen für eine (lineare) SVM nicht brauchbar zu lernen:



3. Interpretieren Sie die geometrische Bedeutung der Konstante C in der Problembeschreibung

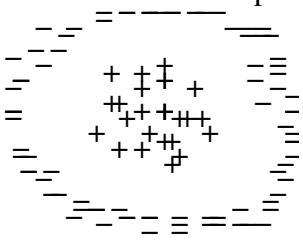
$$\min \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

für die weich trennende Hyperebene.

Mit C kann man beeinflussen, wie hoch der Klassifikationsfehler sein darf. Insbesondere legt man damit fest, wie weit eine Instanz auf der „falschen“ Seite der Hyperebene liegen darf.

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?

Bei diesem Beispiel ist eine weich trennende Hyperebene wenig hilfreich:



Jedoch kann man hier mit einer quadratischen Funktion die Ebene so in den dreidimensionalen Raum „verbiegen“, dass sich eine trennende Hyperebene im dreidimensionalen Raum finden läßt. Mehr dazu in der nächsten Vorlesung.

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?

Ein weiteres Maß ist der sogenannte „gain-ratio“:

$$\text{gain-ratio}(T, A) = \frac{\text{informationsgewinn}(T, A)}{\text{split-info}(T, A)}$$

mit

$$\text{split-info}(T, A) = -\sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} \log_2 \left(\frac{|T_i^A|}{|T|} \right),$$

eine Abwandlung des Informationsgewinns, welcher vom C4.5-Algorithmus verwendet wird. Ein Nachteil des Informationsgewinns ist, dass er Attribute mit vielen Ausprägungen bevorzugt. Diesen Nachteil kann man mit dem gain-ratio beheben.