JProf. Dr. Uwe Ligges, Prof. Dr. Katharina Morik
Julia Schiffner, Felix Jungermann

Dortmund, 15. Juni
Abgabe: bis zum 22.06. um 10:00 Uhr an
felix.jungermann@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 9

Aufgabe 9.1 (4 Punkte)
Sei eine beliebige Hyperebene H gegeben als

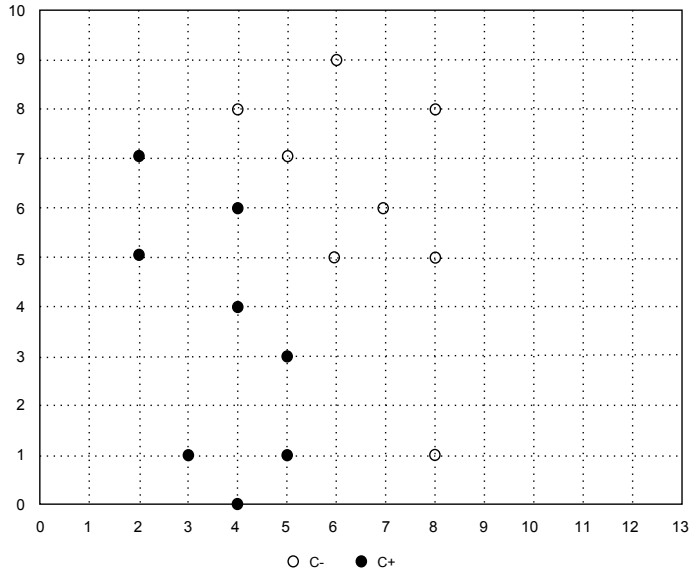
$$H = \left\{ \vec{x} \mid \langle \vec{\beta}, \vec{x} \rangle + \beta_0 = 0 \right\} .$$

(a) Leiten Sie her, wie sich der Abstand der Hyperebene zum Ursprung berechnen lässt.

(b) Sei $y(\vec{x}) = \langle \vec{\beta}, \vec{x} \rangle + \beta_0$. Zeigen Sie, dass die vorzeichenbehaftete Distanz $d(\vec{x}, H)$ eines Punktes $\vec{x}$ zur Hyperebene H (in Hesse'scher Normalform) gegeben ist durch

$$d(\vec{x}, H) = \frac{y(\vec{x})}{||\vec{\beta}||} .$$

Aufgabe 9.2 (3 Punkte)



Gegeben sei eine Menge von Tupeln

$$\begin{aligned} D \ = \{ \ & (2,5,+1),(2,7,+1),(3,1,+1),(4,0,+1), \\ & (4,4,+1),(4,6,+1),(5,1,+1),(5,3,+1), \\ & (4,8,-1),(5,7,-1),(6,5,-1),(6,9,-1), \\ & (7,6,-1),(8,1,-1),(8,5,-1),(8,8,-1) \ \}, \end{aligned}$$

wobei ein Tripel (x_1, x_2, y) jeweils aus der ersten und zweiten Koordinate eines Punktes aus der Abbildung und seiner Klassenzuordnung $y \in \{-1, +1\}$ besteht. Sei C_- die Menge aller Punkte mit $y = -1$ und C_+ die Menge aller Punkte mit $y = +1$.

(a) Wählen Sie aus C_- und C_+ geeignete Stützvektoren aus und stellen Sie die dazugehörigen Geradengleichungen auf. Überlegen Sie sich in diesem Zusammenhang, wie viele Stützvektoren zur eindeutigen Bestimmung dieser Geraden mindestens benötigt werden.

(b) Ermitteln Sie die optimale separierende Hyperebene (hier eine Gerade) zwischen den gewählten Stützvektoren und geben Sie diese Gerade in Hesse'scher Normalform an.

Aufgabe 9.3 (3 Punkte)
In dieser Aufgabe sollen Sie die bereits bekannten Spam-Daten in RapidMiner mit Hilfe einer SVM analysieren. Lesen Sie die schon für Blatt 8 benutzte `spam.txt`-Datei mittels `Read CSV` ein. Achten Sie darauf, dass Sie mit `Set Role` die Rolle `label` vergeben. Normalisieren Sie die Daten mittels `Normalize` auf einen Bereich (Methode `range transformation`) von 0 bis 1.

(a) Testen Sie in einer Kreuzvalidierung die Performanz der SVM. Verwenden Sie dafür den Operator `Support Vector Machine` mit Default-Einstellungen für die Parameter. Geben Sie die Konfusionsmatrix an. Wie bewerten Sie das Ergebnis?

(b) Wir wollen nun eine SVM mit einem Radial-Basis-Kernel (kernel type `radial`) verwenden. Dabei ist unklar, wie wir Gamma (`kernel gamma`) wählen sollen. Aus diesem Grund soll der Ihnen bereits bekannte Operator `Loop Parameters` zum Einsatz kommen.

Erstellen Sie — wie in Aufgabe 1, Blatt 4 (siehe dort) — ein Experiment, bei dem mit Hilfe des Operators `Loop Parameters` der Parameter Gamma von 0 bis 4 in 10 Schritten linear durchlaufen wird. Innerhalb der Schleife soll jeweils die zuvor beschriebene SVM mit einer 10-fachen Kreuzvalidierung für das jeweilige Gamma evaluiert werden. Stellen Sie im Operator `Performance (Classification)` als Fehlermaß `accuracy` ein und loggen Sie diese zusammen mit Gamma über `Log`. Denken Sie daran, dass Sie dafür den Performance-Wert des Operators `X-Validation` loggen müssen. Geben Sie bitte den Scatterplot von Gamma und Accuracy zusammen mit Ihrem Experiment ab!
Achtung: Benutzen Sie vor `Loop Parameters` den Operator `Sample (Stratified)` mit sample ratio 0.5, um die Laufzeit zu verringern.