Prof. Dr. Claus Weihs, Prof. Dr. Katharina Morik
Julia Schiffner, Marco Stolpe

Dortmund, 26. Mai
Abgabe: bis zum 03.06. um 10:00 Uhr an
marco.stolpe@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2009

Blatt 7

Aufgabe 7.1 (4 Punkte)
Sei eine beliebige Hyperebene H gegeben als

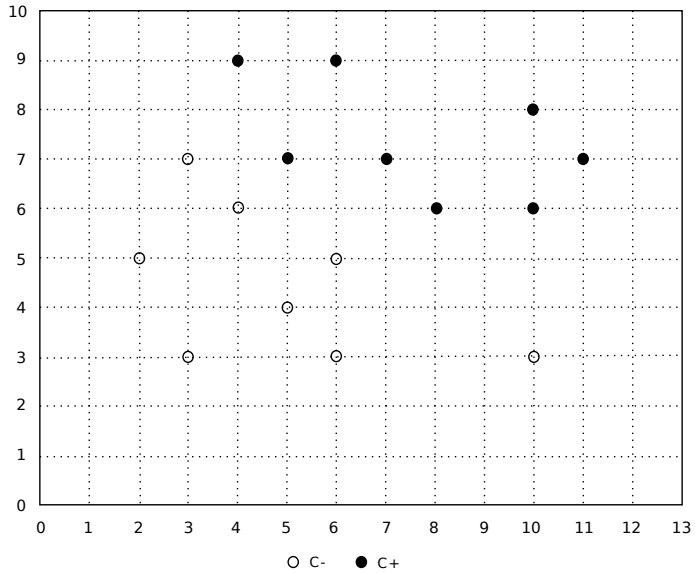
$$H = \left\{ \vec{x} \mid \langle \vec{\beta}, \vec{x} \rangle + \beta_0 = 0 \right\} .$$

(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 7.2 (3 Punkte)



Gegeben sei eine Menge von Tupeln

$$\begin{aligned} D \; = \{ \; & (2,3,-1),(6,3,-1),(10,3,-1),(5,4,-1), \\ & (2,5,-1),(6,5,-1),(4,6,-1),(3,7,-1), \\ & (8,6,+1),(10,6,+1),(5,7,+1),(7,7,+1), \\ & (11,7,+1),(10,8,+1),(4,9,+1),(6,9,+1) \; \}, \end{aligned}$$

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 7.3 (3 Punkte)
Nachdem wir für die Klassifikation von Spam mit Naive Bayes und Entscheidungsbäumen sehr bescheidene Ergebnisse erhalten haben, möchten wir es noch einmal mit der Stützvektormethode versuchen. Lesen Sie mit Hilfe von `ExampleSource` erneut die Spam-Daten von Blatt 5 in RapidMiner ein. Normalisieren Sie die Daten mittels `Normalization` 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 `JMySVMLearner` mit Default-Einstellungen für die Parameter. Geben Sie — wie auf dem letzten Übungsblatt — 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 `ParameterIteration` zum Einsatz kommen.

Erstellen Sie — wie in Aufgabe 1, Blatt 3 (siehe dort) — ein Experiment, bei dem mit Hilfe des Operators `ParameterIteration` 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 `ClassificationPerformance` als Fehlermaß `Accuracy` ein und loggen Sie diese zusammen mit Gamma über `ProcessLog`. Denken Sie daran, dass Sie dafür den zweiten Performance-Wert des Operators `XValidation` loggen müssen. Geben Sie bitte den Scatterplot von Gamma und Accuracy zusammen mit Ihrem Experiment ab! Achtung: Das Experiment dauert sehr lange!