Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs
Dipl.-Inform. Klaus Friedrichs,
Dipl.-Stat. Julia Schiffner,
Dr. Issam Ben Khediri

Dortmund, 22.05.12

Abgabe: bis Mi, 30.05., 12.00 Uhr an
ibenkhediri@statistik.tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2012

Blatt 7

Aufgabe 7.1 (5 Punkte)
Für diese Aufgabe müssen Sie 5 Trainingsdatensätze mit je 100 Beobachtungen simulieren. Erzeugen Sie die Daten gemäß des folgenden Modells: $Y = f(X) + \epsilon = 0.5 + 0.75X + \epsilon$, wobei $\epsilon \sim N(0, 1)$ und X gleichverteilt im Intervall $[-5, 5]$ sind. R-Code zur Erzeugung der Daten liegt in der Datei Sim1.R auf der Homepage.

a) Passen Sie an die 5 Trainingsdatensätze jeweils ein lineares Modell ($Y = \beta_0 + \beta_1 X + \epsilon$) an.

b) Machen Sie für jedes der 5 geschätzten Modelle eine Vorhersage $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ in $x_0 = 1.5$.

c) Simulieren Sie 20 Beobachtungen für $x_0 = 1.5$ gemäß des oben angegebenen Simulationsmodells.

d) Schätzen Sie den erwarteten quadratischen Vorhersagefehler in x_0. Mitteln Sie dabei sowohl über die simulierten Y-Werte aus c) als auch über die 5 Trainingsmengen.

e) Berechnen Sie die Bias-Varianz-Zerlegung des quadratischen Vorhersagefehlers (Formel (10) im Skript), d.h. bestimmen Sie σ_ϵ^2, $Var(\hat{y}_0) = E_\tau((E_\tau(\hat{y}_0) - \hat{y}_0)^2 | x_0)$ und $[f(x_0) - E_\tau(\hat{y}_0)]^2$.

Aufgabe 7.2 (3 Punkte)
In dieser Aufgabe sollen Sie den naiven Vorläufer des SVM-Algorithmus ausprobieren. In der Datei Datei1.txt auf der Homepage finden Sie eine Menge von gelabelten Datenpunkten D (vgl. Abbildung 1):

Dabei bezeichnen die Komponenten jedes Tripels (x_1, x_2, y) aus D die erste und zweite Koordinate des Punktes sowie die zugehörige Klasse $y \in \{-1, 1\}$. Zur Klassifikation ist eine Hyperebene $H = \{\vec{x} | < \vec{w}, \vec{x} > + w_0 = 0\}$ gesucht.

1. Bestimmen Sie für $j \in \{-1, +1\}$ jeweils die Mittelpunkte $\vec{c}_+$ und $\vec{c}_-$ der Menge $C_j = \{(x_1, x_2, y) \in D | y = j\}$.

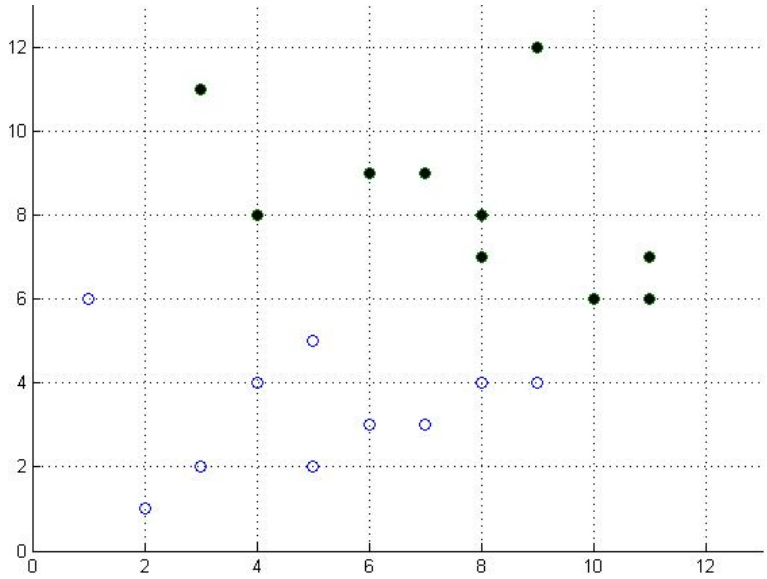

Abbildung 1: Datenpunkte im $\mathbb{R}^2$

2. Bestimmen Sie $\vec{w}$ und den Mittelpunkt $\vec{c}$.

3. Zu welchen Klassen werden anhand dieses einfachen Verfahrens die folgenden Punkte zugeordnet ? (2,10),(8,6),(4,6),(11,2)

Aufgabe 7.3 (1 Punkt)
In dieser Aufgabe sollen Sie für $\vec{x} = (x_1, ..., x_p)$ eine Funktion $f_1(\vec{x})$ unter Nebenbedingungen $h_i(\vec{x})$ optimieren. Stellen Sie dazu, wie auf den Folien, die entsprechende Lagrange-Funktion $L(\vec{x}, \vec{u})$ auf. Die Lösung ergibt sich aus dem Gleichungssystem, das durch Nullsetzen der partiellen Ableitungen von L nach $\vec{x}$ und $\vec{u}$ entsteht.

1. Maximiere $f_1(\vec{x}) = 1 - x_1^2 - x_2^2$ unter der Nebenbedingung $h(\vec{x}) = x_1 + x_2 - 1 = 0$.

Aufgabe 7.4 (2 Punkte)
In dieser Aufgabe geht es darum, den Effekt einer Φ-Transformation auf Datenpunkten zu untersuchen. Gegeben sind die Datenpunkte aus der unten stehenden Tabelle.

x_1	-1.5	-1	-0.5	0	0.5	1	1.5	-1	-0.5	0	0.5	1	1.5
x_2	-2	0	1	2	1	2	3.75	-2	-1	-3	-0.5	-2	1.5
y	+1	+1	+1	+1	+1	+1	+1	-1	-1	-1	-1	-1	-1

Transformieren Sie die Daten mit den nachfolgenden Funktionen Φ_i und geben Sie die transformierten Werte in einer Tabelle an. Welche Funktion ermöglicht eine lineare Trennung der Daten?

1. $\Phi_1(x_1, x_2) = (x_1^2, x_2)$ 2. $\Phi_2(x_1, x_2) = (x_1^3 - 2x_1, x_2)$ 3. $\Phi_3(x_1, x_2) = (x_1^3, x_2)$

Hinweis: Am besten erstellen Sie eine graphische Darstellung der Punkte.