Prof. Dr. Stephan Trahasch
Betriebliche Kommmunikationssysteme
und IT-Sicherheit

Hochschule Offenburg
offenburg.university

# Übungsblatt 3

## SVM 25 Punkte

Data Mining

Wintersemester 2016/17

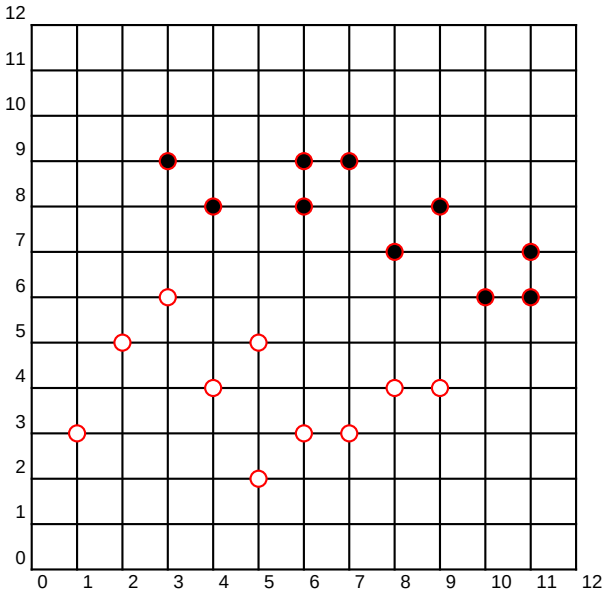
**Abgabe: 02.12.2016 9:45 Uhr**

## Aufgabe 1 SVM-Vorläufer 7 Punkte

In dieser Aufgabe sollen Sie den naiven Vorläufer des SVM-Algorithmus „von Hand“ ausprobieren. Gegeben ist die folgende Menge D von Datenpunkten:

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

Dabei bezeichnen die Komponenten jedes Tripels $(x_1, x_2, y)$ aus D die erste und zweite Koordinate des Punktes sowie die zugehörige Klasse $y \in \{-1, 1\}$.



**Aufgabe 1.1** 4 Punkte

Bestimmen Sie für $j \in \{-1, +1\}$ jeweils die Mittelpunkte $\vec{c_+}$ und $\vec{c_-}$ der Menge

$$C_j = \{(x_1, x_2, y) \in D | y = j\}$$

**Aufgabe 1.2** 2 Punkte

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

**Aufgabe 1.3** 1 Punkt

Zu welchen Klassen werden anhand dieses einfachen Verfahrens die folgenden Punkte zugeordnet?

$$(4,6), (7,6), (12,4), (-1,8), (-4,11)$$

## Aufgabe 2 Kernel-Funktionen 5 Punkte

Eine Stärke der SVM ist die Verwendung von Kernel-Funktionen, die eine implizite Transformation $\phi$ der Daten in einen anderen Raum ermöglichen, so dass ursprünglich linear nicht-trennbare Daten in diesem neuen Raum schließlich trennbar sind. In dieser Aufgabe geht es darum, den Effekt der $\phi$-Transformation der Datenpunkten zu untersuchen. Gegeben sind die Datenpunkte aus der Tabelle.

Transformieren Sie die Daten mit den nachfolgenden Funktionen $\phi_i$ und geben Sie die transformierte Tabelle sowie eine grafische Darstellung der neuen Punkte an. Zeichnen Sie die transformierten Punkte in die Koordinatensysteme auf dem Aufgabenblatt ein.

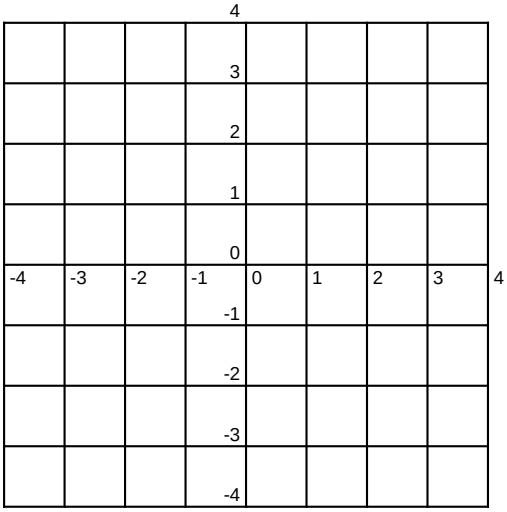
Welche Funktion ermöglicht eine Trennung der Daten?

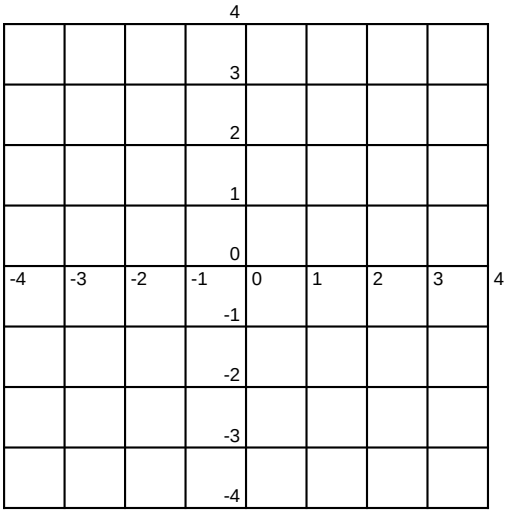
$$\begin{aligned} \phi_1(x_1, x_2) &= (x^2{}_1, x_2) \\ \phi_2(x_1, x_2) &= (x^3{}_1 - 2x_1, x_2) \\ \phi_3(x_1, x_2) &= (x^3{}_1, x_2) \end{aligned}$$

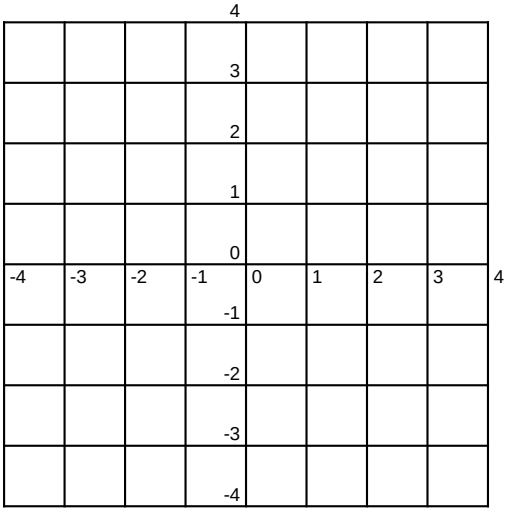
Hinweis: Die graphische Darstellung ist als Hilfe für Sie gedacht.

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| -1.5 | -2.0 | +1 |
| -1.0 | 0.0 | +1 |
| -0.5 | 1.0 | +1 |
| 0.0 | 2.0 | +1 |
| 0.5 | 1.0 | +1 |
| 1.0 | 2.0 | +1 |
| 1.5 | 3.75 | +1 |
| -1.0 | -2.0 | -1 |
| -0.5 | -1.0 | -1 |
| 0.0 | -3.0 | -1 |
| 0.5 | -0.5 | -1 |
| 1.0 | -2.0 | -1 |
| 1.5 | 1.5 | -1 |

Tabelle 1: Instanzen


4
3
2
1
0
-4
-3
-2
-1
-1
0
1
2
3
4
-2
-3
-4


4
3
2
1
0
-4
-3
-2
-1
-1
0
1
2
3
4
-2
-3
-4


4
3
2
1
0
-4
-3
-2
-1
-1
0
1
2
3
4
-2
-3
-4

## Aufgabe 3 SVM mit R 8 Punkte

Sie untersuchen in dieser Aufgabe Daten zum Verkauf von Orangensaft. Der Datensatz enthält 1070 Beobachtungen mit 18 Variablen. Die Erläuterung der Variablen finden Sie im R-Skript. Es soll ein Model mit SVM trainiert werden, das aufgrund einer Beobachtung vorhesagt, ob der Kunde die Marke Citrus Hill (CH) oder Minute Maid Orange Juice (MM) kauft. Der Datensatz ist im Package ISLR unter OJ enthalten. Die Datei 03_SVM_OrangeJuice_Student.R enthält das Grundgerüst für die Aufgabe. Integrieren Sie Ihre Lösungen in diese Datei.

Trainieren Sie SVM-Modelle mit folgenden Kernel-Funktionen

a) linear

b) radial

c) poly

und führen Sie ein Tuning durch.

Bewerten Sie die Ergebnisse. Welche Parameter ergeben das beste Modell?

## Aufgabe 4 Vergleich von Verfahren 5 Punkte

Sie untersuchen in dieser Aufgabe wieder den Datensatz vehicle.dat, der bereits beim ersten und zweiten Übungsblatt verwendet wurde. Informationen dazu finden Sie unter
http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29

Dieser umfasst 846 Beobachtungen und 19 Variablen. Ziel ist es, eine Silhouette dem richtigen von 4 Autotypen zuzuordnen. Die Silhouette ist dabei gegeben durch 18 verschiedene Merkmale wie Kompaktheit, Längenverhältnisse, usw. Die Zielvariable heisst Class und umfasst die Typen bus, opel, saab und van.

Erstellen Sie ein neues Modell mit SVM und vergleichen Sie das Ergebnis mit RandomForest (Übungsblatt 1) sowie k-NN (Übungsblatt 2). Zeichnen Sie eine ROC-Diagramm mit allen drei Modellen und bewerten Sie die Verfahren.