

# Grundlagen des maschinellen Lernens

## Aufgabenblatt 06

Prof. Dr. David Spieler  
Hochschule München

22. November 2018

**Abgabe bis 29. November 2018, 23:59 Uhr**

**Aufgabe 1 (Maximal Margin Klassifikator)** Gegeben seien die Trainingsdaten in Tabelle 1.

Datenpunkt	$\mathbf{x}_1^{(i)}$	$\mathbf{x}_2^{(i)}$	$y^{(i)}$
$\mathbf{x}^{(1)}$	3	4	-1
$\mathbf{x}^{(2)}$	2	2	-1
$\mathbf{x}^{(3)}$	4	4	-1
$\mathbf{x}^{(4)}$	1	4	-1
$\mathbf{x}^{(5)}$	2	1	1
$\mathbf{x}^{(6)}$	4	3	1
$\mathbf{x}^{(7)}$	4	1	1

Tabelle 1: Trainingsdaten für den Maximal Margin Klassifikator.

1. Zeichnen Sie die Trainingsdatenpunkte in ein Koordinatensystem ein. Verwenden Sie schwarze (gefüllte) Kreise für positiv klassifizierte Punkte und weiße (ungefüllte) Kreise für negative klassifizierte Punkte. Schreiben Sie neben jeden Punkt den Index  $i$ .
2. Zeichnen Sie die optimale separierende Hyperebene ein und erstellen Sie eine Gleichung, die für alle Punkte gilt, die genau auf dieser Hyperebene liegen.
3. Geben Sie eine mathematische Definition des Maximum Margin Klassifikators  $f : \mathbb{R}^2 \rightarrow \{-1, 1\}$  an, welcher die Trainingsdatenpunkte perfekt klassifiziert.
4. Zeichnen Sie im Koordinatensystem den maximalen Abstand (Margin) ein.

5. Geben Sie alle Indizes der Datenpunkte an, welche Supportvektoren von  $f$  darstellen.
6. Zeichnen Sie einen neuen Datenpunkt  $\mathbf{x}^{(8)}$  in einer Klasse Ihrer Wahl ein, sodass die Datenpunkte nicht mehr linear trennbar sind.

**Aufgabe 2 (Praxisbeispiel SVM)** In dieser Aufgabe erstellen Sie ein Klassifikationsmodell mit Hilfe von SVM auf den *Auto* Datensatz aus dem R Begleitpaket des Buches *An Introduction to Statistical Learning, with applications in R*, G. James, D. Witten, T. Hastie and R. Tibshirani, Springer, 2013. Der Datensatz enthält Charakteristiken verschiedener Automodelle.

1. Laden Sie den Auto-Datensatz *Auto.csv* in einen *DataFrame*.
2. Schauen Sie sich den kompletten Datensatz mit Hilfe von `print(df.to_string())` an. Was fällt Ihnen auf, wenn Sie sich das Feature *horsepower* genauer anschauen? Korrigieren Sie das Problem.
3. Erstellen Sie den Zielvektor  $\mathbf{y}$  welcher den Wert 1 annehmen soll für alle Autos, deren Verbrauch über dem Median liegt und 0 falls nicht.
4. Erstellen Sie die Feature-Matrix  $\mathbf{X}$  indem Sie aus dem *DataFrame* die Spalten *mpg* und *name* tilgen.
5. Bei SVM müssen Sie für das Fitten den Hyperparameter  $C$  festlegen, welcher ebenso optimiert werden sollte. Wir machen das wie im letzten Übungsblatt mit einer Rastersuche auf dem Gitter

[0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0].

Verwenden Sie einen linearen Kernel für die SVM und fitten Sie mit Hilfe 5-facher Kreuzvalidierung. Berechnen Sie dabei für jeden Wert von  $C$  den Mittelwert der Genauigkeit und plotten Sie am Ende die Genauigkeit über  $C$ . Verwenden Sie mit Hilfe von `matplotlib.pyplot.semilogx` eine logarithmische Skala für die X-Achse.

6. Für welches Modell entscheiden Sie sich und warum?
7. Wiederholen Sie das Training mit polynomialen Kernel zweiten Grades auf dem Raster

[0.01, 0.05, 0.1, 0.2, 0.5, 1.0].

Was fällt Ihnen auf?

8. Wiederholen Sie das Training mit radialem Basis-Kernel auf dem Raster

[0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0]

und spielen Sie mit dem Hyperparameter  $\gamma$ . Können Sie ein besseres Modell als vorher finden?