Machine Learning

Prof. Dr. Fabian Brunner

<fa.brunner@oth-aw.de>

Amberg, 29. November 2021

Übersicht

Thema heute: Support Vector Machines

- Grundidee
- Herleitung der Problemformulierung
- Vergleich mit Logistischer Regression
- Transformation in höherdimensionale Räume

Problemstellung und Notation

Problemstellung

- Binäres Klassifikationsproblem
- p numerische Features („unabhängige Variablen“)
- Binäre Zielvariable („unabhängige Variable“) mit den Klassen -1 und $+1$.
- m Trainingsdatensätze

$$(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)}) \ ,$$

wobei $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$ und $y^{(i)} \in \{-1, +1\}$.

Modell-Training

Bestimme anhand der Trainingsdaten eine Entscheidungsregel)

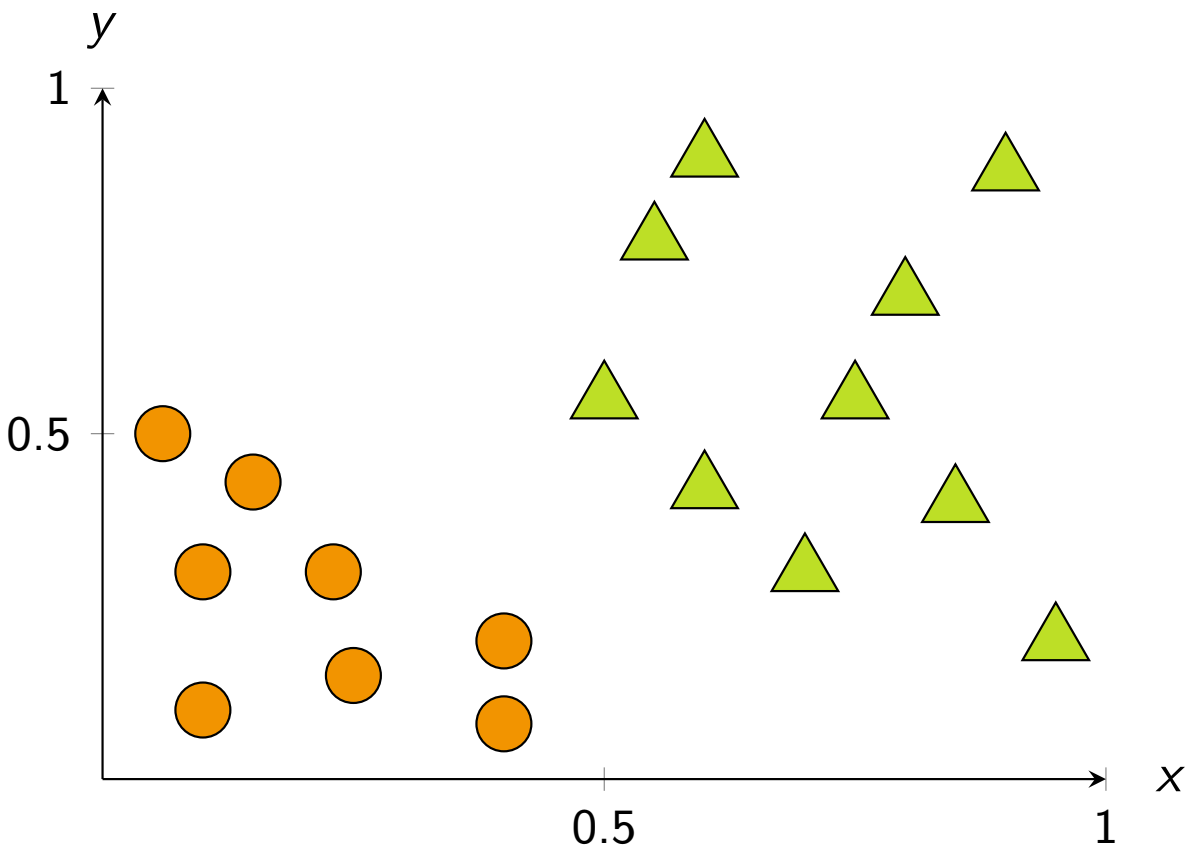
$$f : \mathbb{R}^p \to \{-1, +1\} \ .$$

Modellanwendung

Für einen Query Point $\boldsymbol{x}_q$ prognostiziere das Label $f(\boldsymbol{x}_q)$.

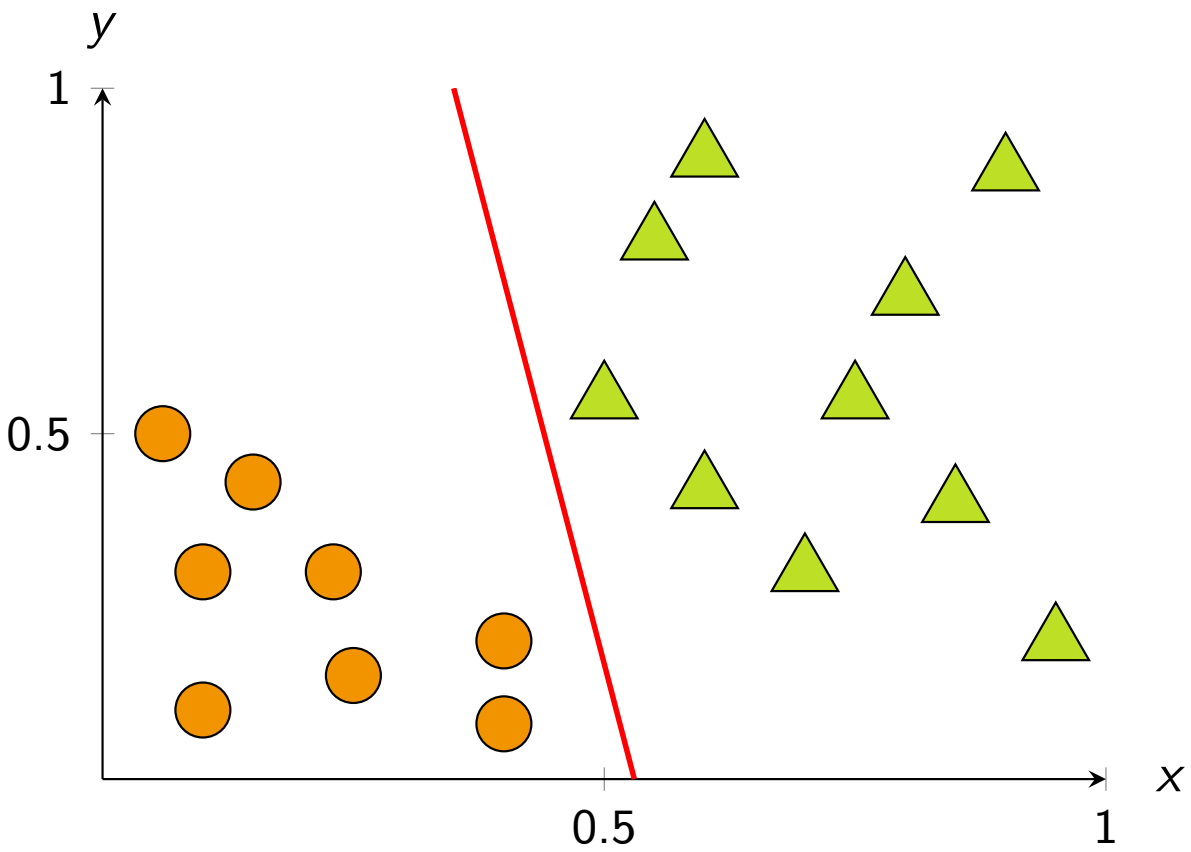
Large Margin Classification

- Ziel bei SVMs ist die Bestimmung eines Decision Boundary mit möglichst großem Randbereich („margin“), in dem keine Samples liegen. Modelle mit solchen Entscheidungsgrenzen weisen tendenziell geringere Fehler beim Verallgemeinern auf unbekannte Daten auf.
- Man spricht bei SVM daher von einem „**large margin classifier**“.



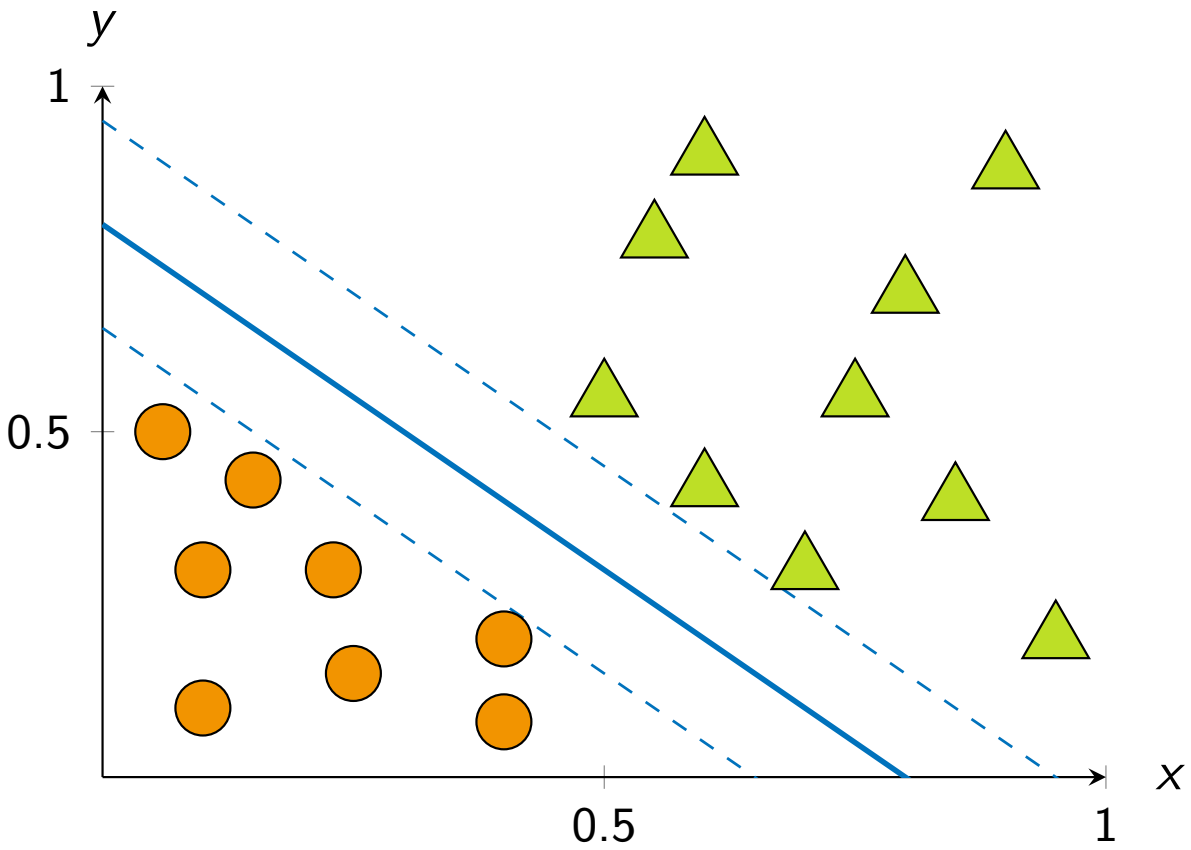
Large Margin Classification

- Ziel bei SVMs ist die Bestimmung eines Decision Boundary mit möglichst großem Randbereich („margin“), in dem keine Samples liegen. Modelle mit solchen Entscheidungsgrenzen weisen tendenziell geringere Fehler beim Verallgemeinern auf unbekannte Daten auf.
- Man spricht bei SVM daher von einem „**large margin classifier**“.



Large Margin Classification

- Ziel bei SVMs ist die Bestimmung eines Decision Boundary mit möglichst großem Randbereich („margin“), in dem keine Samples liegen. Modelle mit solchen Entscheidungsgrenzen weisen tendenziell geringere Fehler beim Verallgemeinern auf unbekannte Daten auf.
- Man spricht bei SVM daher von einem „**large margin classifier**“.



Das Euklidische Skalarprodukt

Gegeben seien zwei Vektoren

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

- Dann berechnet sich das Skalarprodukt von $\boldsymbol{x}$ und $\boldsymbol{y}$ durch

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y} = x_1 y_1 + x_2 y_2 .$$

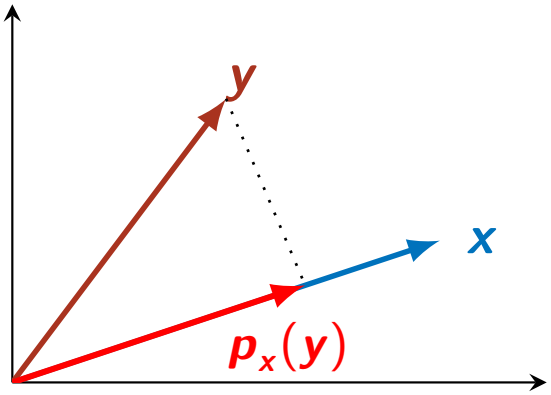
- Die Länge des Ortsvektors von $\mathbf{0}$ zu $\boldsymbol{x}$ lautet

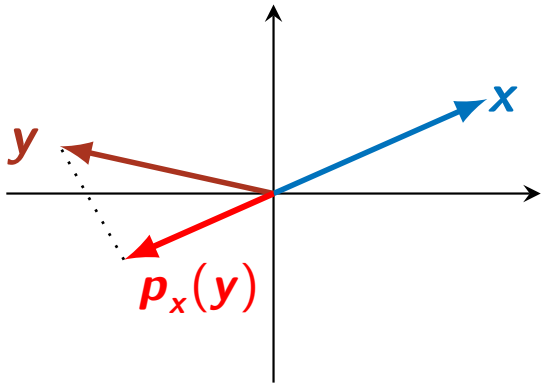
$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} = \sqrt{x_1^2 + x_2^2} .$$

- Mit Hilfe des Skalarprodukts $p := \boldsymbol{x} \cdot \boldsymbol{y}$ lautet die Projektion von $\boldsymbol{y}$ auf $\boldsymbol{x}$:

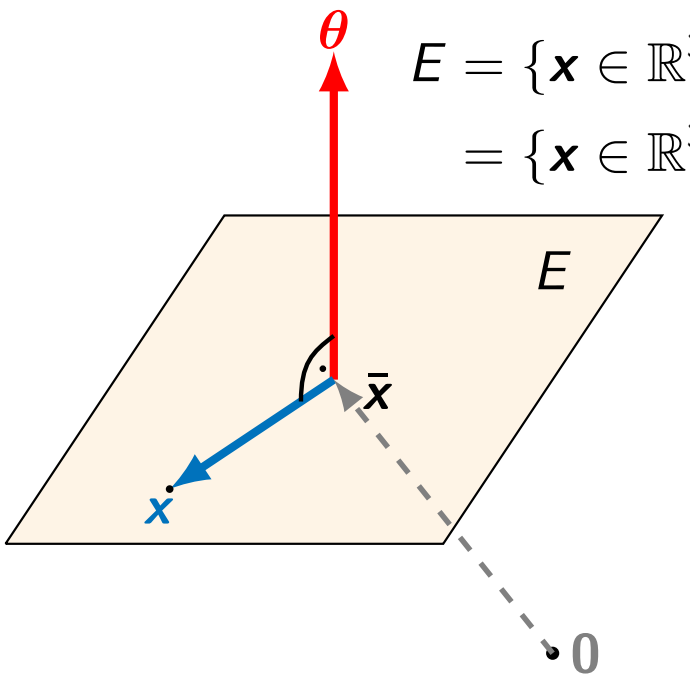
$$\boldsymbol{p}_x(\boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|^2} \cdot \boldsymbol{x} \Rightarrow |\boldsymbol{x} \cdot \boldsymbol{y}| = \|\boldsymbol{x}\| \cdot \|\boldsymbol{p}_x(\boldsymbol{y})\| .$$

- Ist $\boldsymbol{x}$ normiert, gibt das Skalarprodukt $\boldsymbol{x} \cdot \boldsymbol{y}$ also die (signierte) Länge der Projektion von $\boldsymbol{y}$ auf $\boldsymbol{x}$ an.





Ebenen



$$E = \{\boldsymbol{x} \in \mathbb{R}^3 \ : \ (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \boldsymbol{\theta} = 0\}$$
$$= \{\boldsymbol{x} \in \mathbb{R}^3 \ : \ \boldsymbol{x} \cdot \boldsymbol{\theta} = \underbrace{\bar{\boldsymbol{x}} \cdot \boldsymbol{\theta}}_{=:b}\}$$

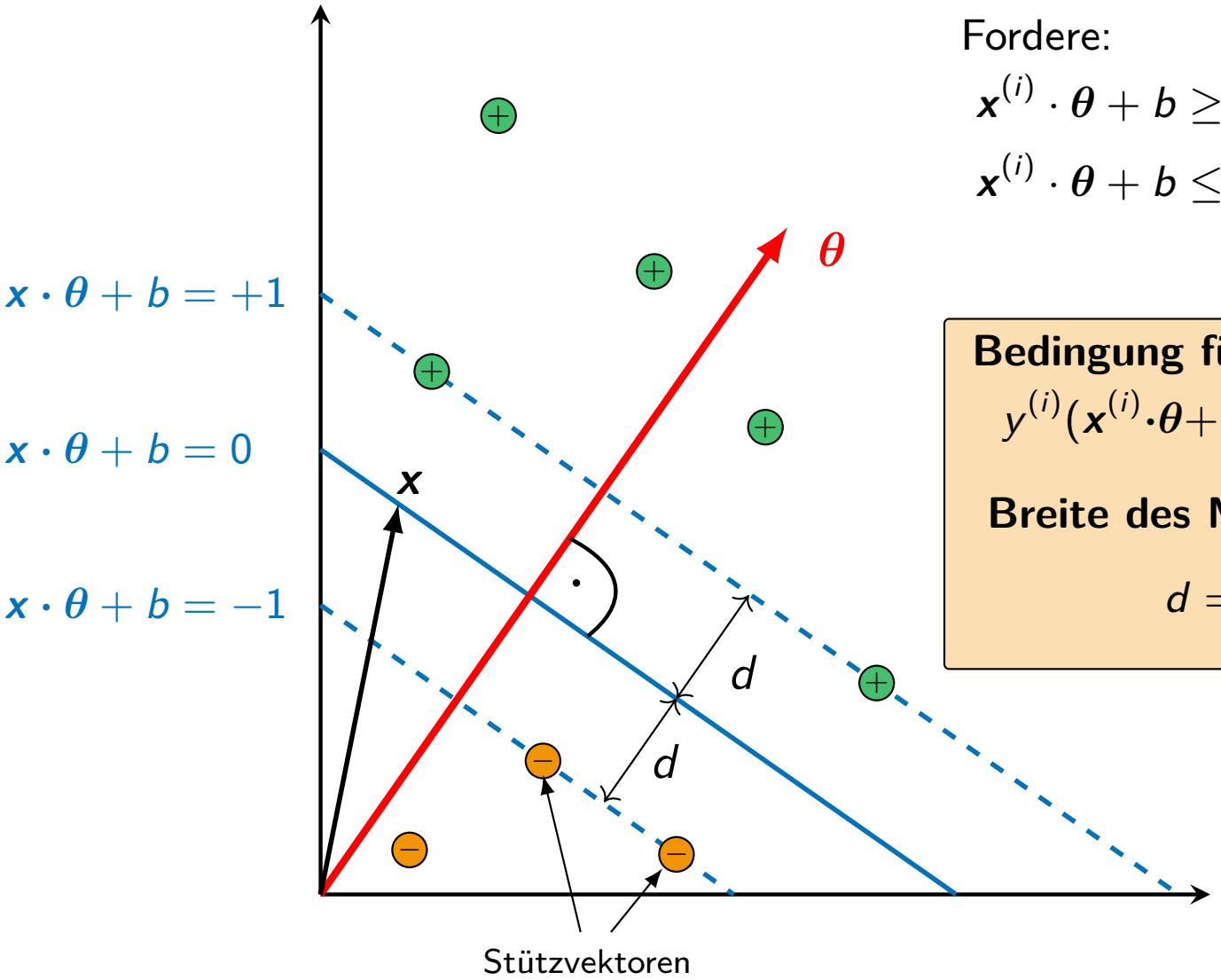
Verständnisfragen:

- Welchen Wert hat b, wenn die Ebene durch den Nullpunkt verläuft?
- Welche relative Lage haben zwei Ebenen mit den Darstellungen

$$E_1 = \{\boldsymbol{x} \in \mathbb{R}^3 \ : \ \boldsymbol{x} \cdot \boldsymbol{\theta} = b_1\},$$
$$E_2 = \{\boldsymbol{x} \in \mathbb{R}^3 \ : \ \boldsymbol{x} \cdot \boldsymbol{\theta} = b_2\} \ ?$$

- Wie berechnet man den Abstand von E zum Nullpunkt?

Erzeugung von Margins



Fordere:

$$x^{(i)} \cdot \theta + b \geq 1 \qquad \text{für } y^{(i)} = +1$$

$$x^{(i)} \cdot \theta + b \leq -1 \qquad \text{für } y^{(i)} = -1$$

Bedingung für Margin:

$$y^{(i)}(x^{(i)} \cdot \theta + b) \geq 1 \text{ für alle } x^{(i)}, y^{(i)}$$

Breite des Margins:

$$d = \frac{1}{\|\theta\|}$$

SVM - linear separierbarer Fall

Um einen möglichst breiten Margin zu bekommen, ist also $\frac{1}{\|\boldsymbol{\theta}\|}$ unter den Nebenbedingungen $y^{(i)}(\boldsymbol{x}^{(i)} \cdot \boldsymbol{\theta} - b) \geq 1$ zu maximieren. Alternativ kann auch $\|\boldsymbol{\theta}\|^2$ minimiert werden. Es ergibt sich ein restringiertes Optimierungsproblem:

Optimierungsproblem im Fall linear separierbarer Samples

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p,\, b \in \mathbb{R}} \quad \frac{1}{2}\|\theta\|^2 \qquad \textbf{(P)}$$

$$\text{u.d.N.} \quad y^{(i)}(\boldsymbol{\theta} \cdot \boldsymbol{x}^{(i)} + b) \geq 1\,, \quad i = 1, \ldots, m$$

- Es handelt sich um ein (konvexes) quadratisches Problem mit linearen Ungleichungs-Nebenbedingungen.
- Es ist nur dann lösbar, wenn die Daten linear trennbar sind. Andernfalls können die Nebenbedingungen nicht erfüllt werden.
- Ausweg (s. später): modifiziere das Problem, damit auch Fehler (z.B. Punkte innerhalb des Margins oder auf der falschen Seite) toleriert werden können. Dies wird durch Einführung von Schlupfvariablen erreicht.

Konvexe Optimierung mit Nebenbedingungen

- Anstelle des oben hergeleiteten Optimierungsproblems wird beim Modelltraining üblicherweise das dazu **duale Problem** gelöst, da es eine vorteilhafte Struktur im Hinblick auf die Verwendung nichtlinearer Kernels (s. später) aufweist.
- Seien $f : \mathbb{R}^n \to \mathbb{R}$ und $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, l$ stetig differenzierbare konvexe Funktionen. Betrachte das folgende Optimierungsproblem mit Nebenbedingungen:

$$\begin{aligned} &\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \\ &\quad \text{u.d.N. } g_i(\boldsymbol{x}) = 0 \text{ für } i = 1, \ldots, l \ . \end{aligned}$$

- Das Lagrange-Funktional lautet

$$\mathcal{L}(\boldsymbol{x}, \lambda_1, \ldots, \lambda_l) := f(\boldsymbol{x}) + \sum_{i=1}^{l} \lambda_i g_i(\boldsymbol{x})$$

- Das Optimierungsproblem ist äquivalent zum Problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \max_{\boldsymbol{\lambda} \in \mathbb{R}^l} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \ .$$

Methode der Lagrange-Multiplikatoren

Methode der Lagrange-Multiplikatoren

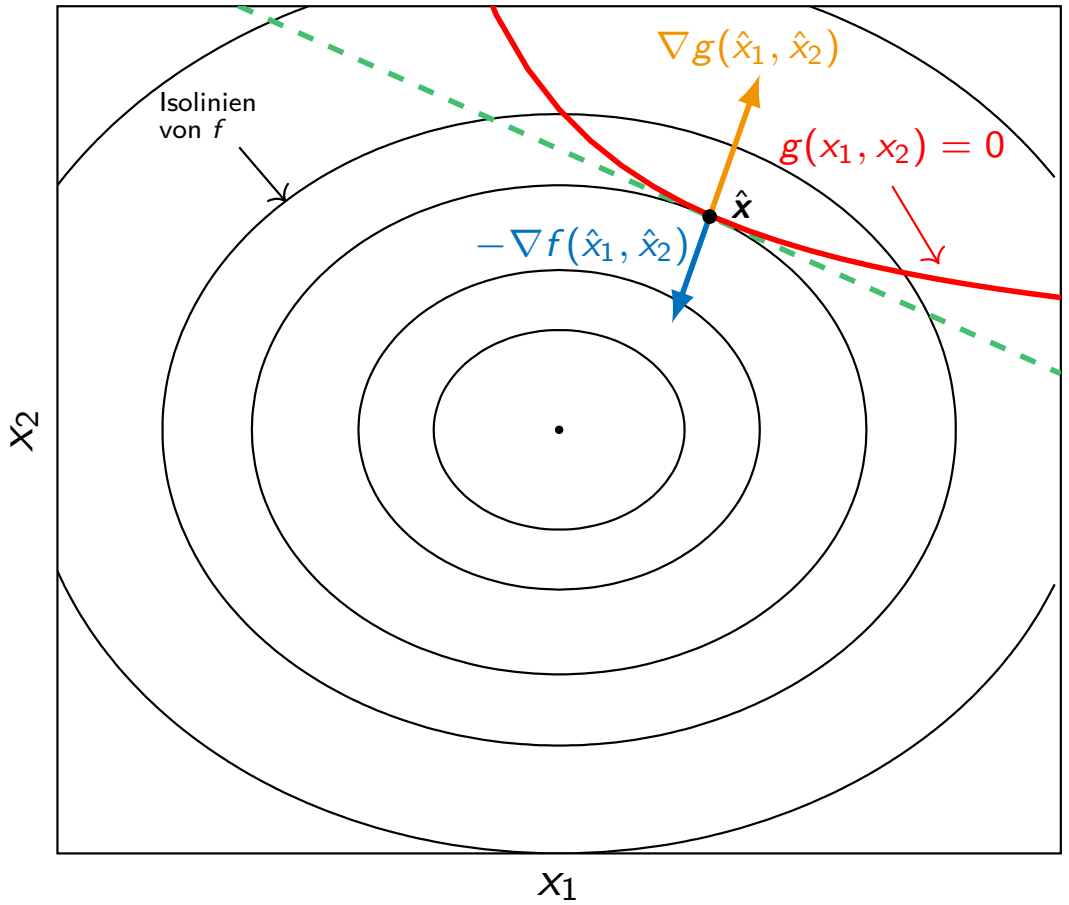
Ist $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ eine Lösung des obigen (konvexen) Optimierungsproblems und sind bestimmte Regularitätsvoraussetzungen (z.B. Slater-Bedingung) erfüllt, dann gibt es Zahlen $\lambda_1, \ldots, \lambda_l \in \mathbb{R}$ (**Lagrange-Multiplikatoren**), sodass

$$\nabla f(\boldsymbol{x}) = \sum_{i=1}^{l} \lambda_i \nabla g_i(\boldsymbol{x}) \ ,$$

$$g_i(\boldsymbol{x}) = 0 \text{ für } i = 1, \ldots, l \ .$$

- Die obigen Bedingungen sind gleichbedeutend damit, dass $\nabla \mathcal{L}(\boldsymbol{x}, \lambda_1, \ldots, \lambda_l) = \boldsymbol{0}$ gilt.
- Es handelt sich allgemein nur um notwendige Bedingungen. Sind sie in einem Punkt erfüllt, dann heißt er **kritischer Punkt** und ist damit ein „Kandidat“ für ein Minimum.

Lagrange-Multiplikatoren - Anschauung



Im Minimum $\hat{\boldsymbol{x}}$ gilt:

$$\nabla f(\hat{x}_1, \hat{x}_2) = \lambda \nabla g(\hat{x}_1, \hat{x}_2)$$

für ein $\lambda \in \mathbb{R}$, d.h. die beiden Vektoren sind kollinear.

Verallgemeinerte Lagrange-Multiplikatoren

Wir erweitern unser Problem nun um Ungleichungsnebenbedingungen:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \; & f(\boldsymbol{x}) \\ \text{u.d.N.} \quad & g_i(\boldsymbol{x}) = 0 \quad \text{für } i = 1, \ldots, l \;, \\ & h_i(\boldsymbol{x}) \leq 0 \text{ für } \quad i = 1, \ldots, k \;. \end{aligned}$$

Das Lagrange-Funktional dazu lautet:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\boldsymbol{x}) + \sum_{i=1}^{l} \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^{k} \alpha_i h_i(\boldsymbol{x}) \;.$$

Primales Problem

Mit Hilfe des Lagrange-Funktionals kann man das obige Optimierungsproblem in folgendes äquivalentes Optimierungsproblem umformulieren:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; \max_{\boldsymbol{\lambda},\boldsymbol{\alpha};\, \boldsymbol{\alpha}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \;.$$

Duales Problem

Durch Vertauschung von max und min erhält man das sog. **duale Problem**

Duales Problem

$$\max_{\boldsymbol{\lambda},\boldsymbol{\alpha};\,\boldsymbol{\alpha}\geq\mathbf{0}} \; \min_{\boldsymbol{x}\in\mathbb{R}^n} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\alpha})$$

- Es gilt stets $d^* \leq p^*$, wobei d^* die Lösung des dualen und p^* die Lösung des primalen Problems bezeichnen.
- Sei d^* die Lösung des dualen Problems und es gelte $d^* = \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\alpha}^*)$. Dann kann $h_i(\boldsymbol{x}^*) < 0$ nur gelten, falls $\alpha_i = 0$.
- Unter bestimmten Voraussetzungen (die bei SVM vorliegen!), gilt auch $d^* = p^*$. Statt des primalen Problems kann man also auch das duale Problem lösen.

KKT-Bedingungen

Ist f konvex und sind die Funktionen h_i und f_i linear, so gilt $d^* = p^*$ und die folgenden KKT-Bedingungen sind notwendige und hinreichende Bedingungen, damit ein Punkt $\hat{\boldsymbol{x}}$ ein Optimum ist:

KKT-Bedingungen

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha})}{\partial x_i}\Big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} &= 0 \ , \\
\alpha_i h_i(\hat{\boldsymbol{x}}) &= 0 \ , \\
\alpha_i &\geq 0 \ , \\
h_i(\hat{\boldsymbol{x}}) &\leq 0 \ , \\
g_i(\hat{\boldsymbol{x}}) &= 0 \ ,
\end{aligned}$$

Duales Optimierungsproblem

Durch Anwenden der KKT-Bedingungen erhält man das zu (P) duale Problem:

Duales Problem

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \, \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$$

$$\text{u.d.N.} \quad \alpha_i \geq 0 \quad \text{für } i = 1, \ldots, m \,,$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \,.$$

- Es handelt sich um ein (konvexes) quadratisches Optimierungsproblem mit Nebenbedingungen.
- Man kann zeigen, dass es eindeutig lösbar ist, sofern die Daten linear separierbar sind.
- Für die Lösung solcher Probleme existieren numerische Löser.

Das duale Problem

- Für die Entscheidungsfunktion kann man zeigen:

$$\boldsymbol{\theta} \cdot \boldsymbol{x}_q + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b \ ,$$

d.h. die zum dualen Problem gehörende Entscheidungsregel lautet:

Entscheidungsregel bei Lösung des dualen Problems

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b \geq 0 \quad \text{Zuordnung zur Klasse } +1 \ ,$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b < 0 \quad \text{Zuordnung zur Klasse -1} \ .$$

- Die Entscheidungsregel hängt, im Gegensatz zum Primalen Problem, von den Trainingsdatenpunkten ab.
- Die resultierenden Gewichte α_i sind allerdings nur für die Stützvektoren ungleich Null, d.h. obige Summe enthält nur sehr wenige Summanden.

Struktur des dualen Problems

Minimierungsproblem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \, \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$$

$+$ Nebenbedingungen

Entscheidungsregel

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b \geq 0 \quad \rightarrow \text{Zuordnung zur Klasse } +1 \,,$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b < 0 \quad \rightarrow \text{Zuordnung zur Klasse -1} \,.$$

Struktur des dualen Problems

Minimierungsproblem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \, \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$$

$+$ Nebenbedingungen

Entscheidungsregel

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b \geq 0 \quad \rightarrow \text{Zuordnung zur Klasse } +1 \,,$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}_q + b < 0 \quad \rightarrow \text{Zuordnung zur Klasse -1} \,.$$

Struktur des dualen Problems

- Das Minimierungsproblem hängt nur von Skalarprodukten zwischen den Feature-Vektoren ab.
- Die Entscheidungsregel hängt nur von Skalarprodukten der Feature-Vektoren mit dem Query Point ab.

SVM - nicht linear separierbarer Fall

Häufig sind die Daten nicht exakt linear separierbar. Dann besitzt das obige Problem keine Lösung. Durch Einführung von **Schlupfvariablen** wird das Verfahren robust gegen Trainings-Fehler.

Optimierungsproblem im Fall nicht linear separierbarer Samples

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p,\ \boldsymbol{\xi}\in\mathbb{R}^m,\ b\in\mathbb{R}} \quad \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{u.d.N.} \quad y^{(i)}(\boldsymbol{\theta}\cdot\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i\ ,\ \xi_i \geq 0 \quad i = 1,\ldots,m\ .$$

- Das obige Problem ist ein quadratisches Optimierungsproblem mit Ungleichungsnebenbedingungen.
- Jede Nebenbedingung kann erfüllt werden, wenn ξ_i hinreichend groß ist.
- Ist $\xi_i > 1$, so liegt $\boldsymbol{x}^{(i)}$ auf der falschen Seite des Decision Boundary.
- Ist $0 \leq \xi_i \leq 1$, so liegt $\boldsymbol{x}^{(i)}$ im Margin (bzw. auf den Ebenen, die diesen abgrenzen)

SVM als unrestringiertes Optimierungsproblem

Ostbayerische Technische Hochschule Amberg-Weiden

Bisher hatten wir SVM als restringiertes Optimierungsproblem formuliert:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p,\boldsymbol{\xi}\in\mathbb{R}^m,b\in\mathbb{R}} \quad \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i$$

unter den Nebenbed. $y^{(i)}(\boldsymbol{\theta}\cdot\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i\ , \quad \xi_i \geq 0\ , \quad i = 1,\ldots,m\ .$

Umformulierung der Ungleichungs-Nebenbedingungen

Wir erweitern nun $\boldsymbol{x}^{(i)}$ und $\boldsymbol{\theta}$ um eine Komponente, indem wir $x_0^{(i)} := 1$ und $\theta_0 := b$ setzen und definieren $f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \boldsymbol{\theta}\cdot\boldsymbol{x}$. Damit lauten die **Ungleichungs**-Nebenbedingungen

$$y^{(i)} f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \geq 1 - \xi_i \quad \wedge \quad \xi_i \geq 0\ .$$

Diese lassen sich wie folgt als äquivalente **Gleichungen** schreiben:

$$\xi_i = \max(0, 1 - y^{(i)} f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))\ .$$

SVM als unrestringiertes Optimierungsproblem

Ostbayerische Technische Hochschule Amberg-Weiden

Durch die Einführung der Maximum-Funktion ist ein Minimierungsproblem ohne

SVM als unrestringiertes Optimierungsproblem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min}\ J(\boldsymbol{\theta}) := \frac{1}{2} \sum_{i=1}^{p} \theta_i^2 + C \sum_{i=1}^{m} \max(0, 1 - y^{(i)} f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})) \ .$$

Nachdem die Parameter $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$ durch Lösen dieses Optimierungsproblems bestimmt worden sind, wird die Klassenzuordnung eines neuen Datenpunkts $\boldsymbol{x}_q$ wie folgt definiert:

Klassenzuordnung bei SVM

Falls $\boldsymbol{\theta} \cdot \boldsymbol{x}_q \geq 0 \rightarrow$ Zuordnung zur Klasse $y = +1$

Falls $\boldsymbol{\theta} \cdot \boldsymbol{x}_q < 0 \rightarrow$ Zuordnung zur Klasse $y = -1$

Regularisierung

Der Parameter C in der zu minimierenden Zielfunktion

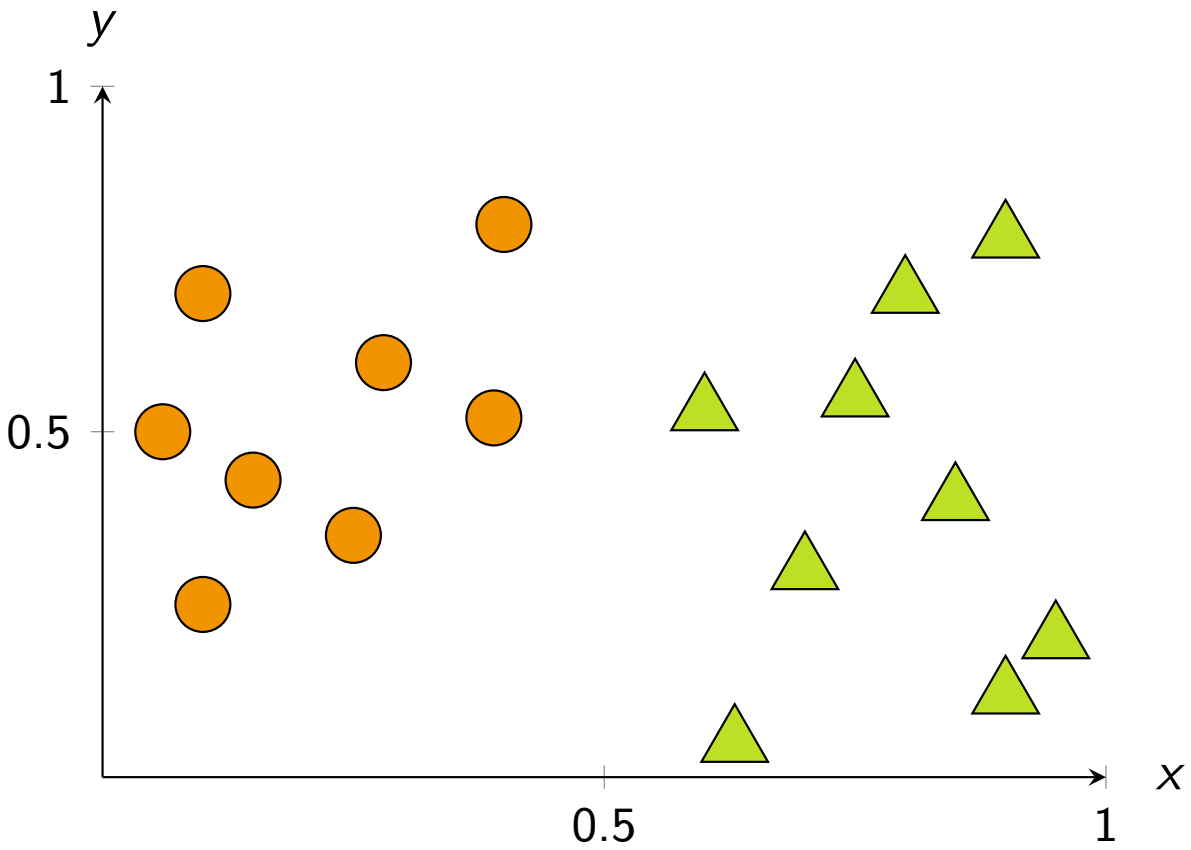
$$\frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i$$

kann als Regularisierungsparameter aufgefasst werden:

- Ist C klein, so erhalten die Nebenbedingungen weniger Gewicht und es werden Trainingsfehler zugelassen („soft margin“). Im Gegenzug wird versucht, die Parameter so zu bestimmen, dass der Margin möglichst breit wird.
- Ist C groß, werden die Parameter so bestimmt, dass die Nebenbedingungen möglichst strikt eingehalten werden. Dafür wird der Margin tendenziell schmäler.
- $C \to \infty$ erzwingt alle Nebenbedingungen („hard margin“)
- C klein: geringere Varianz, höhere Verzerrung
- C groß: höhere Varianz, geringere Verzerrung
- C ist ein Hyperparameter und muss, z.B. mit Hilfe von Kreuzvalidierung, festgelegt werden.

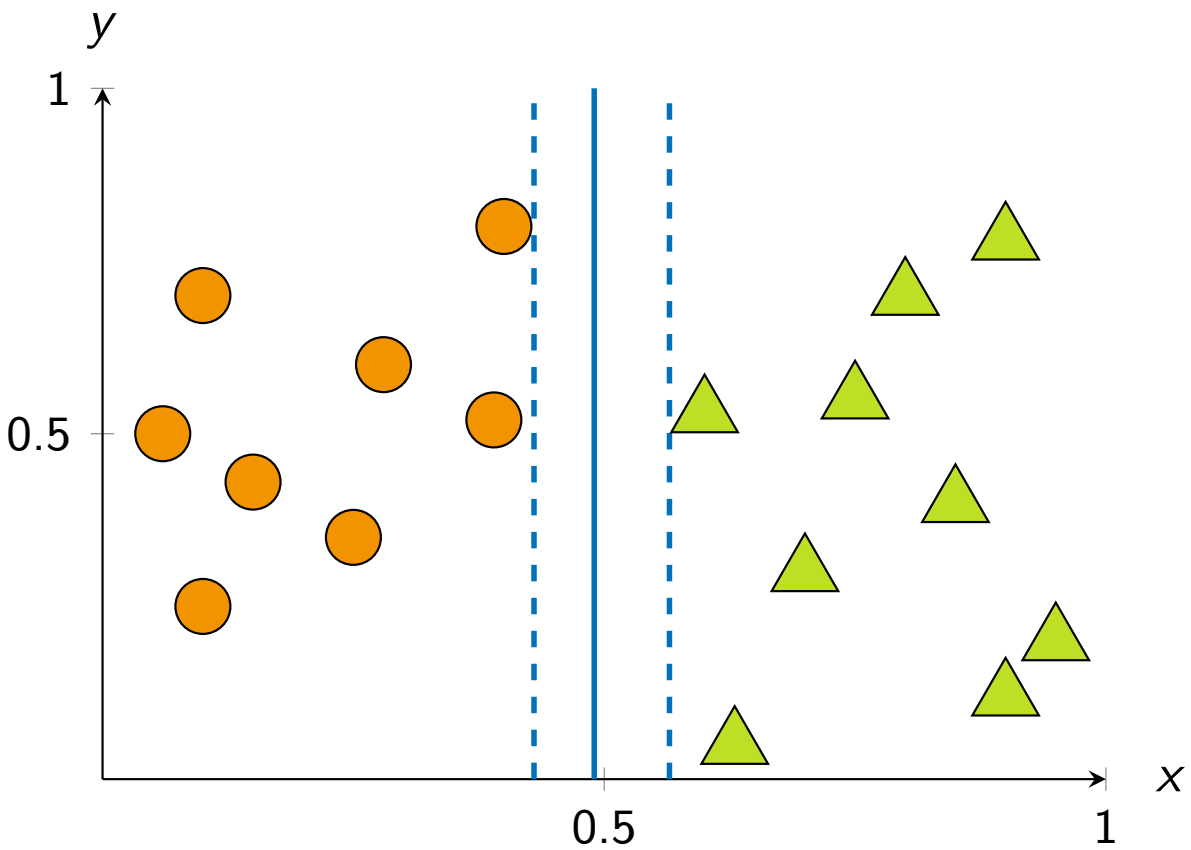
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



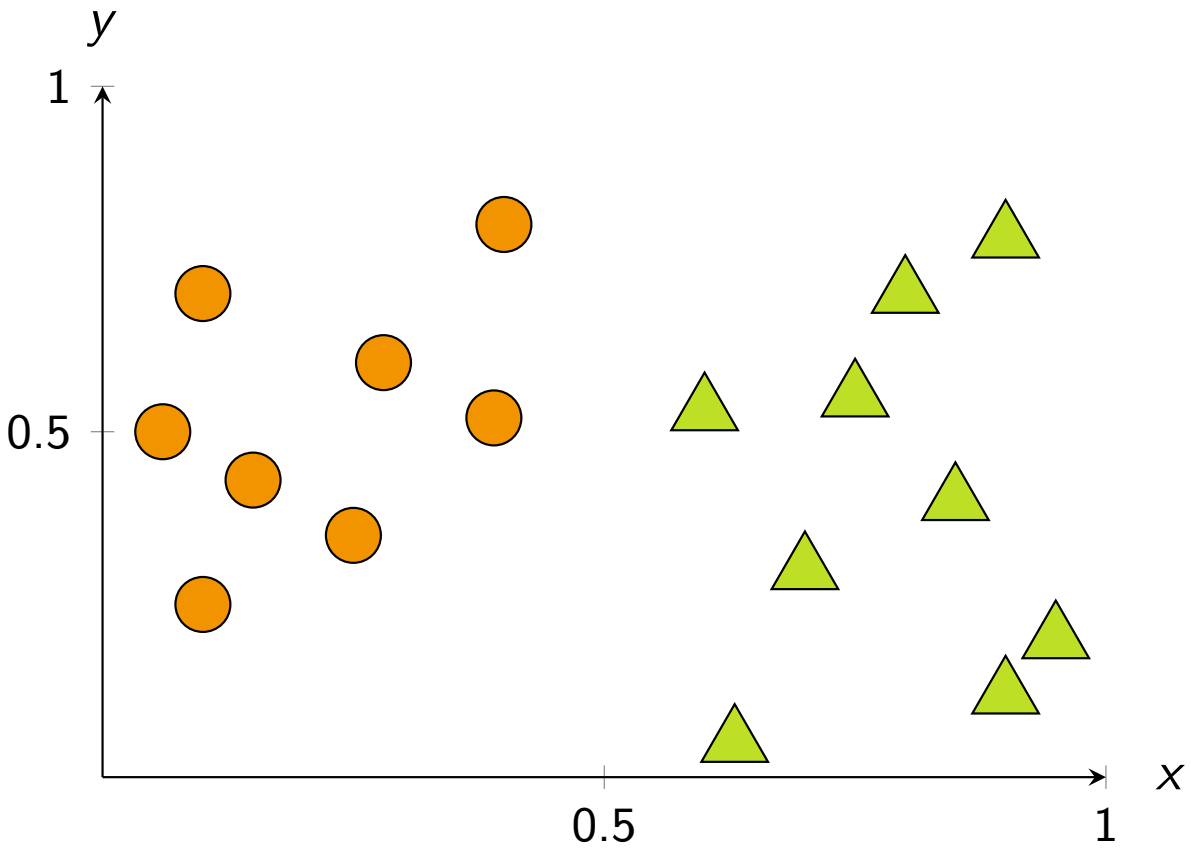
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



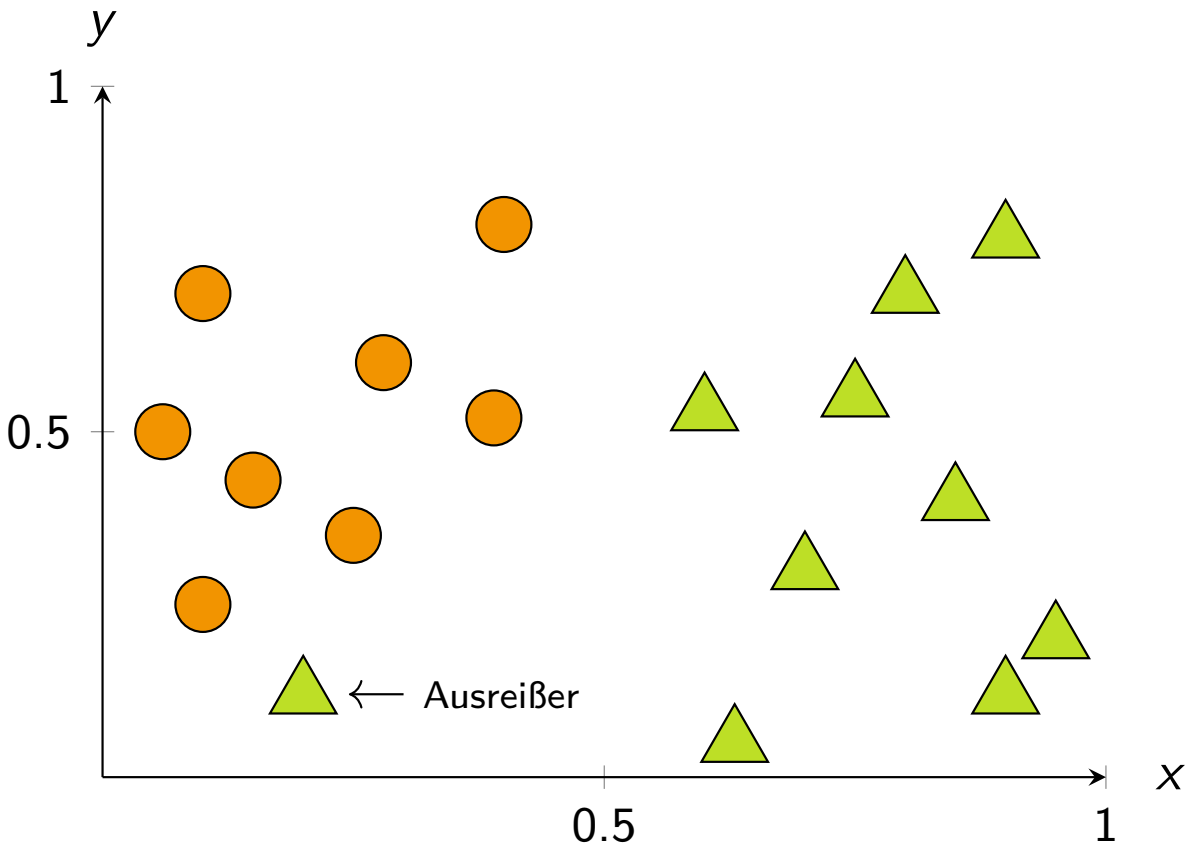
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



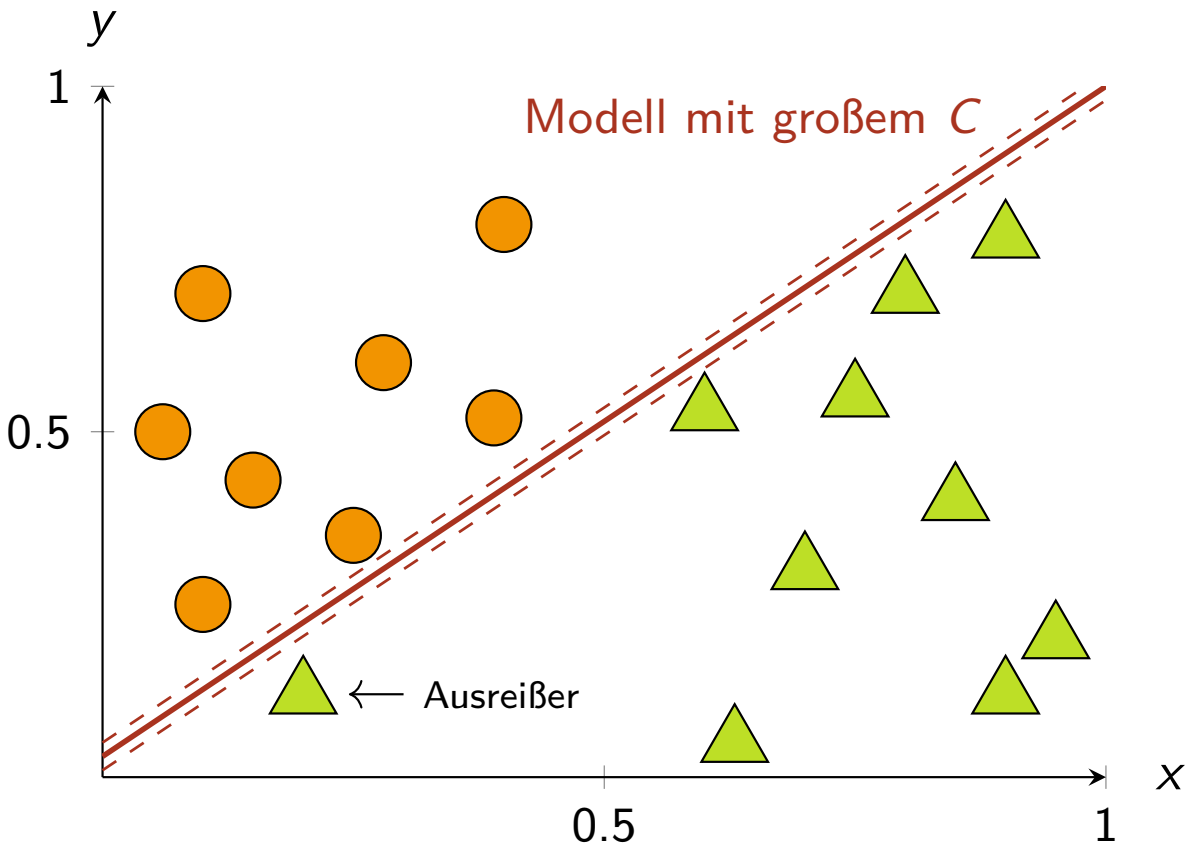
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



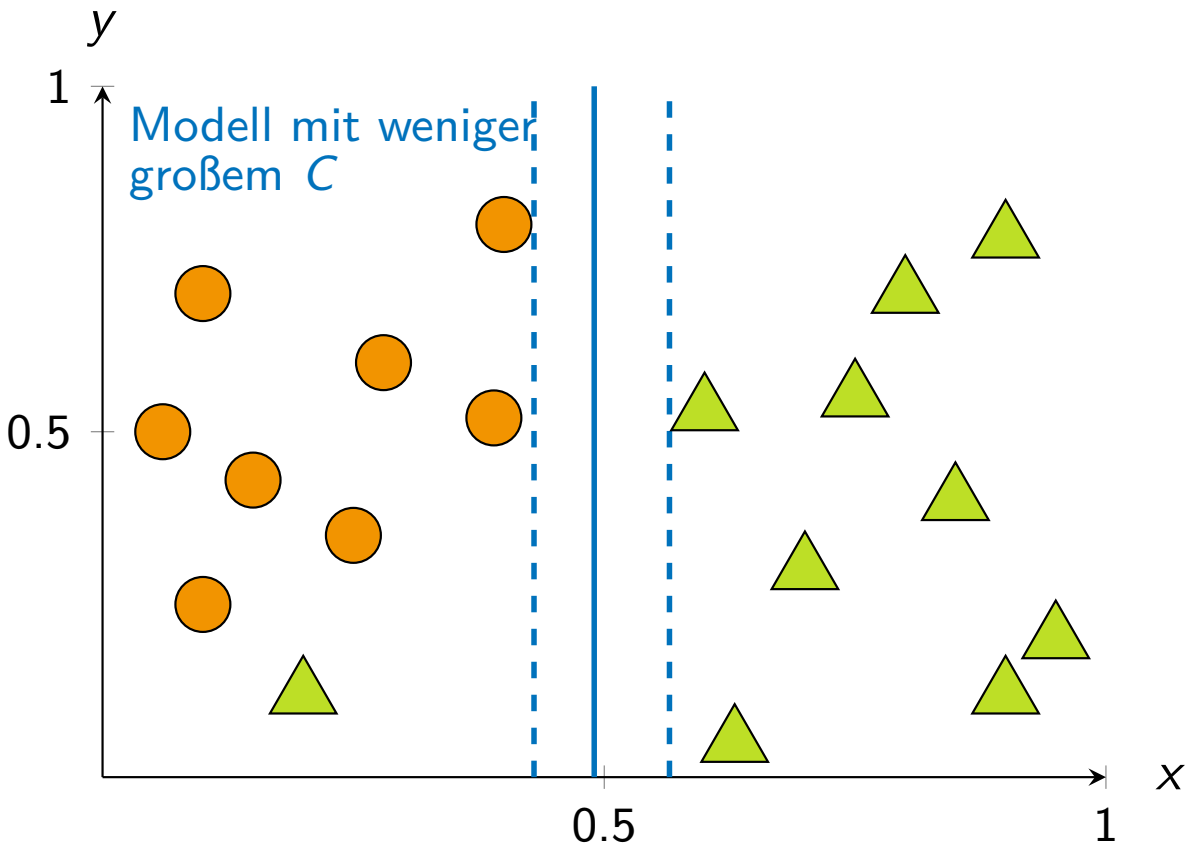
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



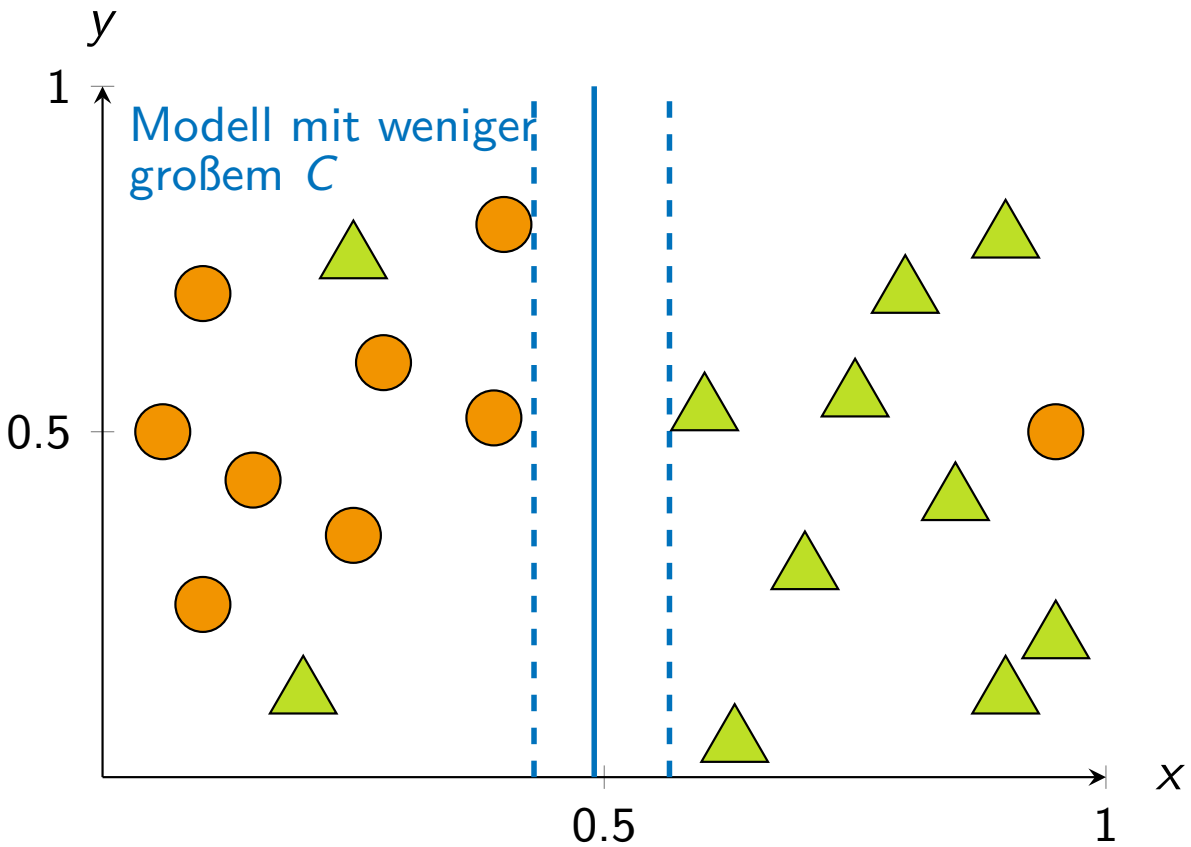
Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



Large Margin classification bei Ausreißern

- Je größer C ist, desto stärker wird das Modell an die Daten angepasst, d.h. die Gerade wird so bestimmt, dass möglichst viele Datenpunkte korrekt klassifiziert werden. Ausreißer haben dann ggf. großen Einfluss.
- Für kleineres C ist das Verfahren „robuster “gegenüber Ausreißern und einzelnen Fehlern (entspricht stärkerer Regularisierung).



Logistische Regression revisited

Ansatz bei der Logistischen Regression

Bei der Logistischen Regression hat die Modellfunktion die folgende Gestalt:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta}^T \boldsymbol{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p) \ .$$

wobei

$$g(x) = \frac{1}{1 + e^{-x}} \ .$$

Die Parameter werden durch Minimierung der Cross Entropy bestimmt:

Cross Entropy als Kostenfunktional bei der Logistischen Regression

$$J(\boldsymbol{\theta}) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \right)$$

SVM vs. Logistische Regression

Regularisierte Logistische Regression

$$J_{LR}(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) + (1-y^{(i)}) \log\left(1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) + \frac{\lambda}{2m}\sum_{i=1}^{p} \theta_i^2 \ .$$

SVM

$$J_{SVM}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{p} \theta_i^2 + C \sum_{i=1}^{m} \max(0, 1 - y^{(i)} f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))$$

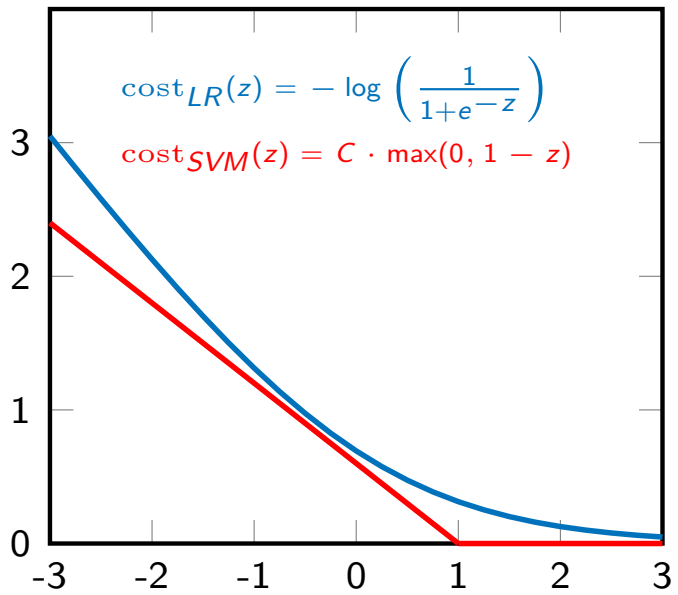
- Der Term zur Beschreibung der Breite des Margins entspricht einem Regularisierungsterm (Varianzreduzierung).
- Klassifikationsfehler auf dem Trainingsdatensatz werden mit Kosten versehen (Reduzierung der Verzerrung)
- Worin unterscheiden sich die Terme, die die Klassifikationsfehler bestrafen?

Vergleich mit Logistischer Regression

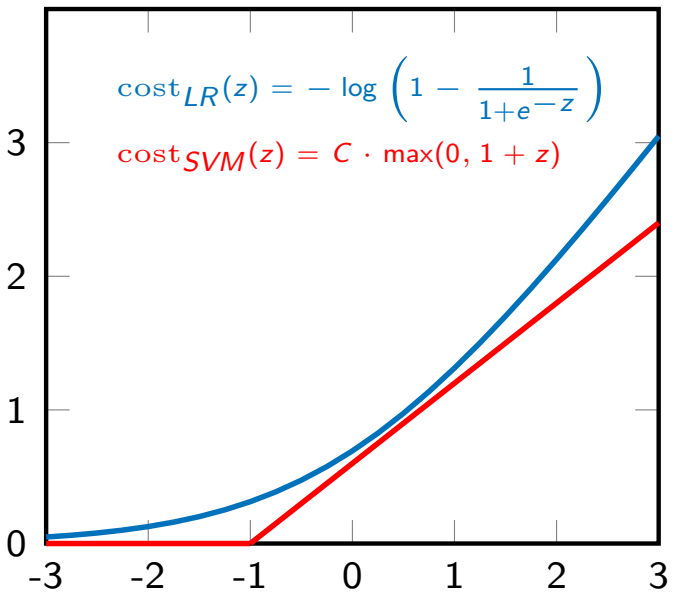
Kosten eines Samples bei der logistischen Regression

$$
\begin{aligned}
&- y \log(f_\theta(\boldsymbol{x})) - (1-y)\log(1 - f_\theta(\boldsymbol{x})) \\
&= -y \log\left(\frac{1}{1+e^{-\theta^T \boldsymbol{x}}}\right) - (1-y)\log\left(1 - \frac{1}{1+e^{-\theta^T \boldsymbol{x}}}\right)
\end{aligned}
$$

$y = 1$

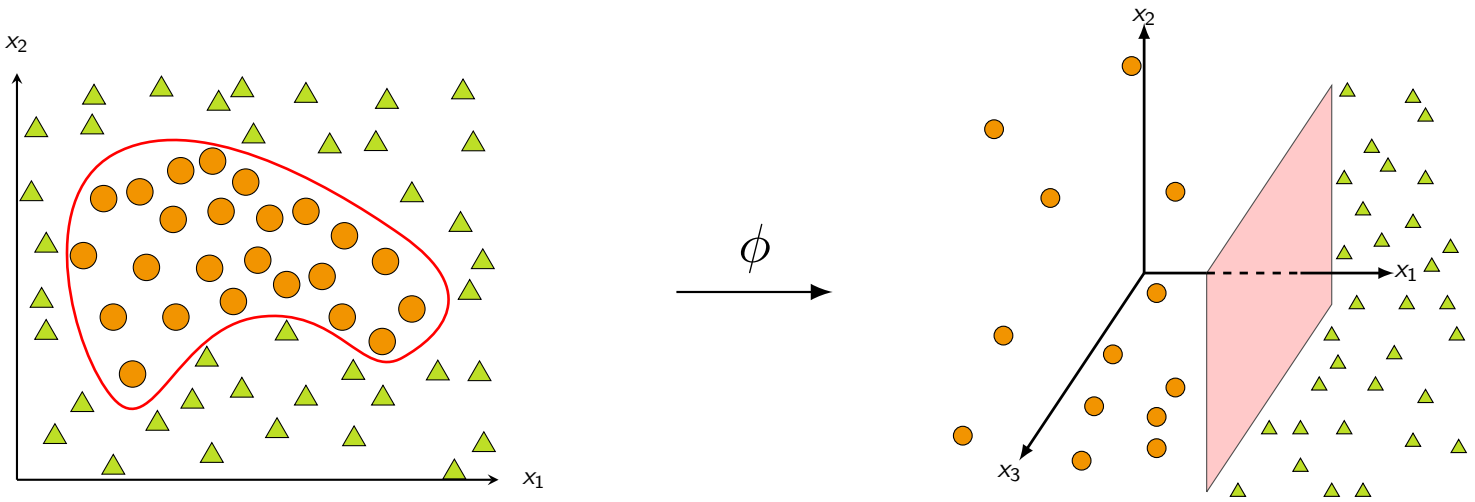


$y = 0$ bzw. $y = -1$



Nicht linear separierbare Daten

Idee: Nichtlineare Transformation des Problems



- Nichtlineare Transformation: $\phi : \mathbb{R}^p \to \mathbb{R}^P$ mit $P \geq p$.
- Fitte ein Modell der Form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta} \cdot \phi(\boldsymbol{x}) + b$.
- Die transformierten Daten $\phi(\boldsymbol{x}^{(1)}), \ldots, \phi(\boldsymbol{x}^{(m)})$ sind ggf. linear trennbar.
- Das transformierte Problem ist weiterhin linear in $\boldsymbol{\theta}$.
- Beispiel: Übergang von der univariaten linearen Regression zur polynomialen Regression:

$$\phi : \mathbb{R} \to \mathbb{R}^{d+1} , \quad x \mapsto \phi(x) = (1, x, x^2, x^3, \ldots, x^d)^T .$$

SVM im transformierten Raum

Minimierungsproblem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,\, j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \, \phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}^{(j)})$$

$+$ Nebenbedingungen

Entscheidungsregel

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}_q) + b \geq 0 \quad \rightarrow \text{Zuordnung zur Klasse } +1$$

SVM im transformierten Raum

Minimierungsproblem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \, \phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}^{(j)})$$

$+$ Nebenbedingungen

Entscheidungsregel

$$\sum_{i=1}^{m} \alpha_i y^{(i)} \phi(\boldsymbol{x}^{(i)}) \cdot \phi(\boldsymbol{x}_q) + b \geq 0 \quad \rightarrow \text{Zuordnung zur Klasse } +1$$

Kernel Trick

- Definiere den **Kernel** $K(\boldsymbol{x}, \boldsymbol{z}) := \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{z})$.
- Wenn man K kennt, kann der SVM-Klassifikator trainiert werden, ohne dass ϕ explizit berechnet werden muss. Dies ist häufig effizienter.
- Es muss kein nichtlineares Problem gelöst werden. Die mathematischen Eigenschaften (Konvexität) des SVM-Klassifikators bleiben erhalten.

Beispiel

Betrachte die Transformation

$$\phi : \mathbb{R}^2 \to \mathbb{R}^3 , \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \end{pmatrix} .$$

Dann gilt

$$\begin{aligned} \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{z}) &= (x_1^2, x_2^2, \sqrt{2} x_1 x_2) \cdot \begin{pmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2} z_1 z_2 \end{pmatrix} \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2 \\ &= (x_1 z_1 + x_2 z_2)^2 \\ &= (\boldsymbol{x} \cdot \boldsymbol{z})^2 . \end{aligned}$$

Die direkte Berechnung des Kernels $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x} \cdot \boldsymbol{z})^2$ ist effizienter als die Transformation und anschließende Berechnung des Skalarprodukts.

Beispiele für Kernels

Linearer Kernel

$$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{z}$$

Entspricht der linearen SVM ohne Verwendung eines Kernels.

Polynomialer Kernel

$$K(\boldsymbol{x}, \boldsymbol{z}) = (1 + \boldsymbol{x} \cdot \boldsymbol{z})^d \text{ für } d > 0 \ .$$

Der polynimiale Kernel enthält alle polynomialen Terme bis zum Grad d.

Gauß'scher Kernel

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2}\right) \text{ für } \sigma > 0 \ .$$

Der Gauß'sche Kernel führt auf einen unendlichdimensionalen Feature Space.

SVM-Klassifikator mit Gauß-Kernel

Entscheidungsfunktion:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i y^{(i)} \underbrace{K(\boldsymbol{x}^{(i)}, \boldsymbol{x})}_{=:f_i(\boldsymbol{x})} + b \ .$$

- Die Funktion $f_i(\boldsymbol{x}) = K(\boldsymbol{x}^{(i)}, \boldsymbol{x})$ ist eine radiale Basisfunktion (RBF) um den Punkt $\boldsymbol{x}^{(i)}$. Man erhält also für jedes Item des Trainingsdatensatzes eine Basisfunktion (m sollte nicht zu groß sein!)
- Für $\boldsymbol{x} \approx \boldsymbol{x}^{(i)}$ ist $f_i(\boldsymbol{x}) \approx 1$. Falls $\boldsymbol{x}$ weit von $\boldsymbol{x}^{(i)}$ entfernt ist, ist $f_i(\boldsymbol{x}) \approx 0$. Der Parameter $\sigma > 0$ gibt an, wie eng f_i um $\boldsymbol{x}^{(i)}$ zentriert ist.

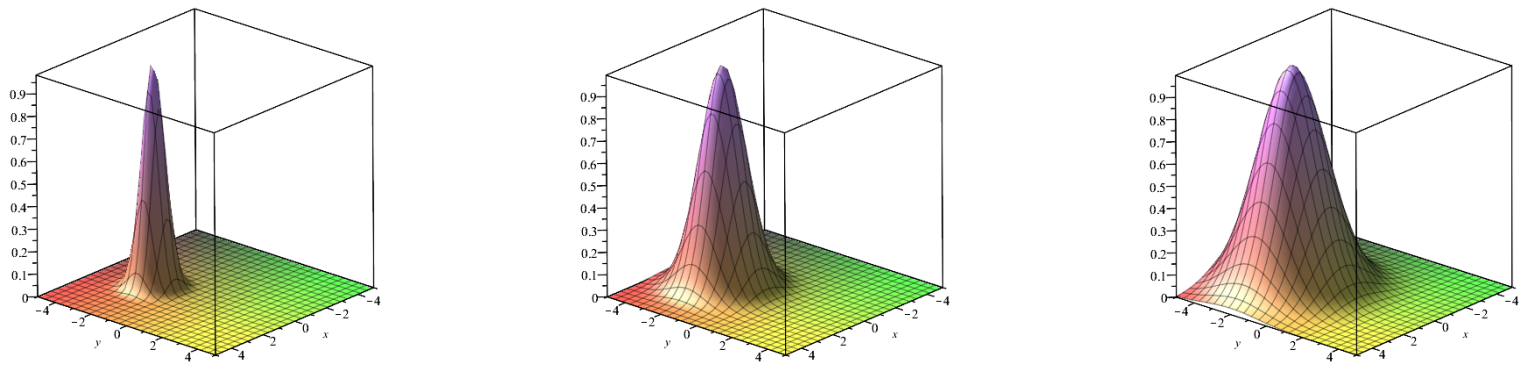
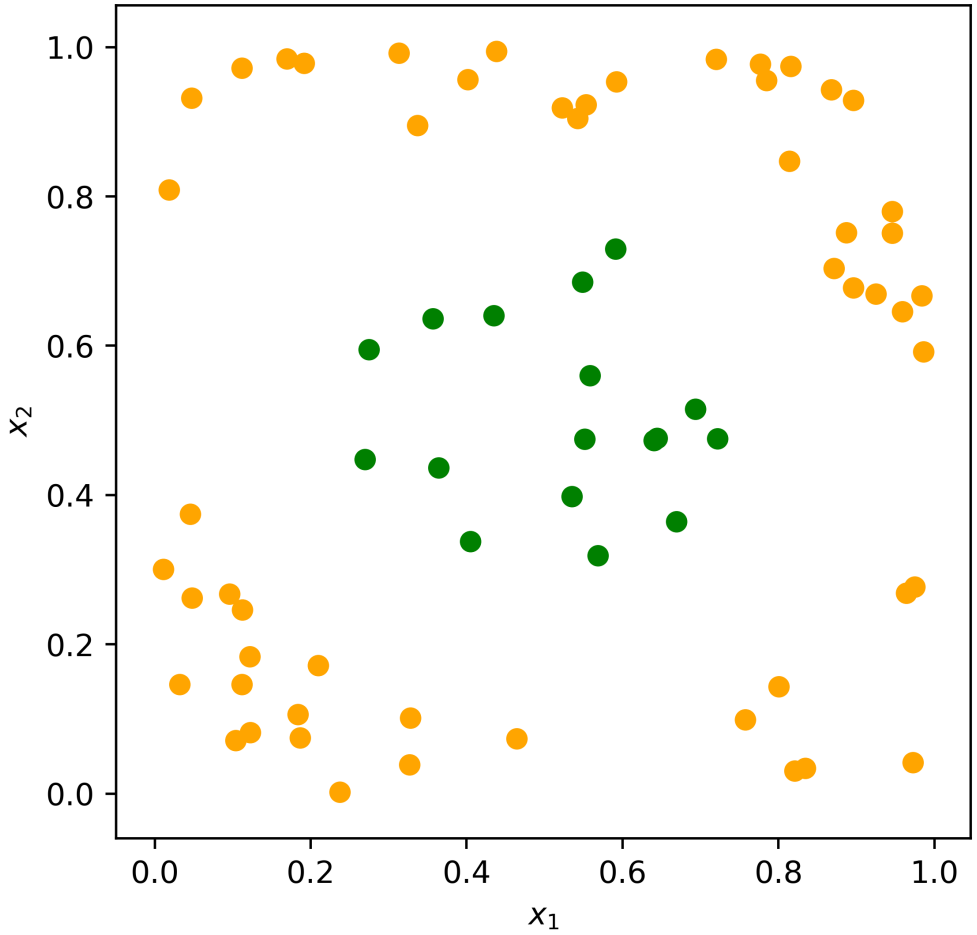
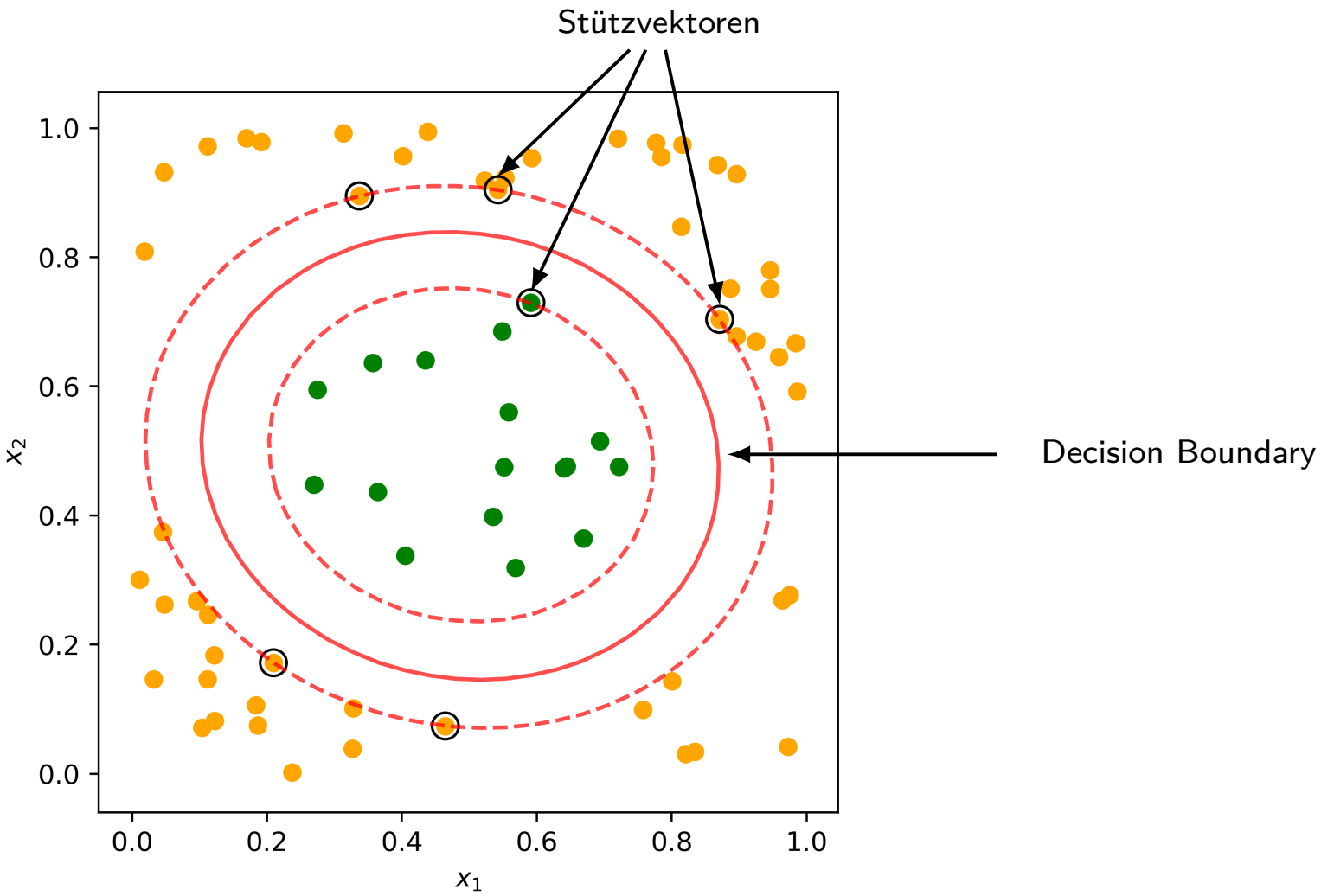


Abbildung: Gauß-Kernel für $\sigma = 0.5$ (links) $\sigma = 1$ (Mitte) und $\sigma = 1.5$ (rechts)

SVM-Klassifikator mit Gauß-Kernel - Beispiel



SVM-Klassifikator mit Gauß-Kernel - Beispiel



Zusammenfassung

- SVM als Large-Margin-Klassifikator
- Hard margin vs. Soft margin
- Formulierung des primalen und dualen Optimierungsproblems (für den linear separierbaren Fall)
- Struktur des dualen Optimierungsproblems
- Einführung von Schlupfvariablen zur Behandlung des nicht linear separierbaren Falls, Regularisierung
- Vergleich mit logistischer Regression
- Transformation in höherdimensionale Feature-Räume unter Verwendung von Kernels