

Kernel-Methoden

1 Einführung in Kernel-Methoden

Im vorherigen Kapitel haben wir die *Support Vector Machine (SVM)* kennengelernt. Sie ist ein Algorithmus, der für einen gegebenen Datensatz eine lineare Trennebene bestimmt, die die Klassen möglichst gut voneinander separiert. Formal lösen wir dazu das Optimierungsproblem

$$\min_{w,b} w^\top w \quad \text{s. d.} \quad y_i(w^\top x_i + b) \geq 1, \quad i = 1, \dots, m.$$

Als Ergebnis erhalten wir eine Entscheidungsfunktion, die jedem Datenpunkt eine Klasse zuordnet:

$$f_{w^*}(x) = \text{sign}(w^{*\top} x + b^*) = \text{sign}\left(\sum_{i=1}^m \alpha_i y_i x_i^\top x + b^*\right).$$

Diesen Algorithmus kann man in Python für konkrete Datensätze implementieren und die entstehende lineare Trennung visualisieren.

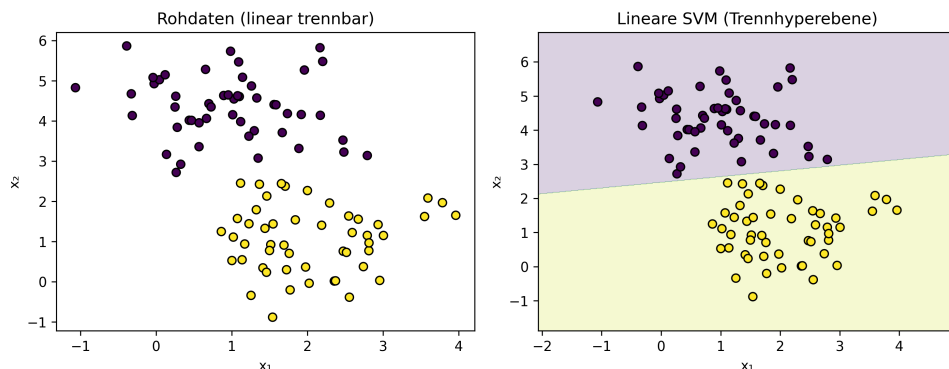


Figure 1: Motivation: linear trennbare Daten

Die SVM lässt sich nun so erweitern, dass sie auch mit *nicht* linear trennbare Daten umgehen kann. Hier kommen *Kernel-Methoden* ins Spiel. Die zentrale Idee besteht darin, die Daten durch eine *Feature Map*

$$\phi : X \rightarrow \mathcal{H}$$

in einen Hilbertraum $\mathcal{H}$ abzubilden, in dem eine lineare Trennung wieder möglich ist. Statt ϕ explizit anzugeben, nutzen wir den *Kernel-Trick*: Ein Kernel ist eine Abbildung

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}},$$

die das Skalarprodukt im höherdimensionalen Feature-Raum beschreibt. In der Entscheidungsfunktion der SVM ersetzen wir dann das Skalarprodukt $x_i^\top x$ durch den Kernel $K(x_i, x)$. So erhalten wir eine nichtlineare Entscheidungsgrenze im Ursprungsraum, die als lineare Trennung in $\mathcal{H}$ interpretiert werden kann, ohne dass $\mathcal{H}$ explizit konstruiert werden muss. Die zugrunde liegende geometrische Idee ist in Abbildung 2 dargestellt.

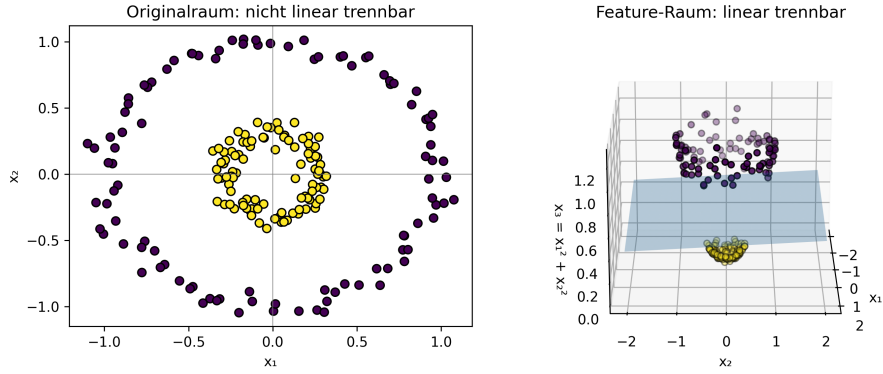


Figure 2: Nichtlineare Daten und lineare Trennung nach der Feature Map

Um diese Idee präziser zu fassen, führen wir im nächsten Schritt einige zentrale Begriffe und Definitionen ein, die wir im weiteren Verlauf benötigen

2 Kerne und reproduzierende Kernel-Hilberträume

Grundlegende Begriffe wie Skalarprodukt, Norm und Cauchyfolgen werden vorausgesetzt und hier nicht erneut definiert. Dabei sei die Norm stets durch $\|f\|_H := \sqrt{\langle f, f \rangle_H}$ gegeben.

Definition 2.1. Sei $(H, \|\cdot\|_H)$ ein normierter Raum. Eine Abbildung $T: H \rightarrow \mathbb{R}$ heißt linearer Operator, wenn für alle $u, v \in H$ und alle $\lambda \in \mathbb{R}$ gilt:

$$T(u + \lambda v) = T(u) + \lambda T(v).$$

T heißt beschränkt, wenn eine Konstante $M > 0$ existiert, mit

$$|T(u)| \leq M \|u\|_H \quad \text{für alle } u \in H.$$

Definition 2.2 (Hilbertraum). Ein normierter Raum $(H, \|\cdot\|_H)$ heißt *vollständig*, wenn jede Cauchy-Folge in H bezüglich $\|\cdot\|_H$ gegen ein Element von H konvergiert.

Ein innerer Produktraum heißt *Hilbertraum*, wenn er vollständig bezüglich der durch das Skalarprodukt induzierten Norm ist.

Satz 2.3 (Rieszscher Darstellungssatz). Sei H ein Hilbertraum und $T: H \rightarrow \mathbb{R}$ ein beschränkter linearer Operator. Dann existiert genau ein $g \in H$ mit

$$T(f) = \langle f, g \rangle_H \quad \text{für alle } f \in H.$$

Der Beweis dieses Satzes wird hier weggelassen, da er technisch aufwendig ist und für den weiteren Vortrag nicht von zentraler Bedeutung ist.

Definition 2.4 (Kernel). Sei $\mathcal{X}$ eine nichtleere Menge. Eine symmetrische Funktion $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ heißt *positiv definiten Kernel*, wenn für alle $n \in \mathbb{N}$, $x_1, \dots, x_n \in \mathcal{X}$ und $c_1, \dots, c_n \in \mathbb{R}$ gilt:

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) \geq 0.$$

Diese Bedingung ist gleichbedeutend damit, dass für alle n und $x_1, \dots, x_n$ die *Gram-Matrix* $K_{ij} = K(x_i, x_j)$ positiv semidefinit ist.

Wir haben nun einen Kernel K eingeführt. Die zentrale Frage ist jedoch, unter welchen Voraussetzungen dieser Kernel tatsächlich als Skalarprodukt eines Hilbertraums aufgefasst werden kann. Genau diese Frage führt zur folgenden Definition.

Definition 2.5 (RKHS). Sei X eine Menge und $H \subset \mathbb{R}^X$ ein Hilbertraum reellwertiger Funktionen $f: X \rightarrow \mathbb{R}$. Ein Kernel $K: X \times X \rightarrow \mathbb{R}$ heißt *reproduzierender Kernel von H* , wenn für alle $f \in H$ und $x \in X$ gilt:

$$f(x) = \langle f, K_x \rangle_H,$$

wobei $K_x := K(x, \cdot) \in H$. In diesem Fall heißt H ein *reproduzierender Kernel-Hilbertraum (RKHS)* mit Kernel K . Insbesondere gilt für alle $x, y \in X$:

$$K(x, y) = \langle K_x, K_y \rangle_H.$$

Wir haben damit eine Voraussetzung formuliert, unter der sich ein gegebener Kernel als Skalarprodukt schreiben lässt und in der SVM verwendet werden kann. Die folgende Bemerkung gibt eine hinreichende Bedingung dafür, dass ein Hilbertraum von Funktionen überhaupt einen solchen reproduzierenden Kernel besitzt.

Bemerkung 2.6. Sei H ein Hilbertraum von Funktionen $f: X \rightarrow \mathbb{R}$. Für jedes $x \in X$ sei das Auswertungsfunktional

$$L_x: H \rightarrow \mathbb{R}, \quad L_x(f) = f(x),$$

linear und beschränkt d.h. es existiert ein $M_x > 0$ mit

$$|f(x)| = |L_x(f)| \leq M_x \|f\|_H \quad \forall f \in H.$$

Dann existiert nach dem Rieszschen Darstellungssatz für jedes $x \in X$ ein eindeutig bestimmtes $K_x \in H$ mit

$$f(x) = \langle f, K_x \rangle_H \quad \forall f \in H.$$

Damit ist

$$K(x, y) := \langle K_x, K_y \rangle_H$$

ein Kernel auf $X \times X$ und H ist eine RKHS mit reproduzierendem Kernel K .

Beispiel 2.7. Sei $X = \mathbb{R}^2$ und definiere für jedes $x \in \mathbb{R}^2$ die Funktion

$$\phi_x: X \rightarrow \mathbb{R}, \quad \phi_x(y) = x_1 y_1 + x_2 y_2 + x_1 x_2 y_1 y_2, \quad y \in \mathbb{R}^2.$$

Wir betrachten den Hilbertraum

$$H = \text{span}\{\phi_x \mid x \in \mathbb{R}^2\}$$

mit Skalarprodukt, das auf den erzeugenden Funktionen durch

$$\langle \phi_x, \phi_y \rangle_H := x_1 y_1 + x_2 y_2 + x_1 x_2 y_1 y_2$$

festgelegt ist, damit ist

$$K(x, y) := \langle \phi_x, \phi_y \rangle_H = x_1 y_1 + x_2 y_2 + x_1 x_2 y_1 y_2$$

ein Kernel auf $X \times X$.

Für $u = (u_1, u_2, u_3) \in \mathbb{R}^3$ definieren wir

$$f_u(x) = u_1 x_1 + u_2 x_2 + u_3 x_1 x_2, \quad x \in \mathbb{R}^2.$$

Man überprüft leicht, dass f_u als Linearkombination von ϕ_x geschrieben werden kann, etwa

$$f_u = \phi_{(u_3, 1)} + \phi_{(u_1 - u_3, 0)} + \phi_{(0, u_2 - 1)} \in H.$$

Für $x \in X$ setzen wir $K_x := K(x, \cdot) = \phi_x$. Dann gilt

$$\begin{aligned} \langle f_u, K_x \rangle_H &= \langle \phi_{(u_3, 1)}, \phi_x \rangle_H + \langle \phi_{(u_1 - u_3, 0)}, \phi_x \rangle_H + \langle \phi_{(0, u_2 - 1)}, \phi_x \rangle_H \\ &= u_3 x_1 + x_2 + u_3 x_1 x_2 + u_1 x_1 - u_3 x_1 + u_2 x_2 - x_2 \\ &= u_1 x_1 + u_2 x_2 + u_3 x_1 x_2 = f_u(x). \end{aligned}$$

Damit ist K reproduzierender Kernel von H .

Für Anwendungen ist entscheidend, dass wir uns bei gegebenem Kernel nicht explizit um den zugrunde liegenden Hilbertraum kümmern müssen, sondern sicher sein können, dass es einen geeigneten Raum gibt, in dem der Kernel als Skalarprodukt realisiert wird. Diese Existenz wird im folgenden Satz formuliert.

Satz 2.8 (Moore–Aronszajn). *Sei $K : X \times X \rightarrow \mathbb{R}$ ein Kernel. Dann existiert ein eindeutig bestimmter RKHS $H \subset \{f : X \rightarrow \mathbb{R}\}$ mit reproduzierendem Kernel K . Insbesondere existiert eine Abbildung $\Phi : X \rightarrow H$, so dass*

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_H \quad \forall x, y \in X.$$

Damit können wir in der SVM den gegebenen Kernel verwenden und im zugehörigen Feature-Raum nach einer linearen Trennung suchen. Bevor wir den folgenden Satz beweisen, betrachten wir zunächst einige Beispiele, um die Konstruktion besser zu verstehen.

Beispiel 2.9 (Polynomkernel). Sei $X \subset \mathbb{R}^2$, $d \in \mathbb{N}$ und $c \geq 0$. Wir betrachten den Polynomkernel

$$K(x, y) = (1 + x^\top y)^d.$$

Für $d = 2$ und $x, y \in \mathbb{R}^2$ gilt:

$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ c \end{pmatrix}^\top \begin{pmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \\ \sqrt{2c} y_1 \\ \sqrt{2c} y_2 \\ c \end{pmatrix}.$$

Dies zeigt, dass der Kernel K als Skalarprodukt in einem höherdimensionalen Raum dargestellt werden kann.

Beispiel 2.10 (Gauß-Kernel / RBF-Kernel). Sei $X \subset \mathbb{R}^n$ und $\sigma > 0$. Für $x, y \in X$ definieren wir

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

Für $\sigma = 1$ und $X = \mathbb{R}$ definieren wir

$$\phi(x) := \left(\frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right)_{n \geq 0} \in \ell^2(\mathbb{N}_0).$$

Dann gilt:

$$\begin{aligned} \langle \phi(x), \phi(y) \rangle &= \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right) \left(\frac{1}{\sqrt{n!}} e^{-\frac{y^2}{2}} y^n \right) \\ &= e^{-\frac{x^2+y^2}{2}} \sum_{n=0}^{\infty} \frac{(xy)^n}{n!} \\ &= e^{-\frac{x^2+y^2}{2}} e^{xy} = e^{-\frac{(x-y)^2}{2}} = K(x, y). \end{aligned}$$

Damit kann der Gauß-Kernel als Skalarprodukt eines unendlichdimensionalen Raumes interpretiert werden.

Beispiel 2.11. Wir erinnern uns an die Lösung der SVM bzw. die Entscheidungsfunktion

$$f(x) = \text{sign}\left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b\right).$$

Für einen gegebenen Kernel K entspricht dies einer linearen Trennung in dem zugehörigen Raum H , die im ursprünglichen Eingaberaum als nichtlineare Entscheidungsgrenze erscheint.

Beweisidee von Satz 2.8.

Ziel: Aus einem Kernel $K : X \times X \rightarrow \mathbb{R}$ eine eindeutige RKHS H mit reproduzierendem Kernel K konstruieren. Sei

$$G = \left\{ \sum_{i=1}^n a_i K_{x_i} \mid n \in \mathbb{N}, x_i \in X, a_i \in \mathbb{R} \right\}.$$

Für

$$f = \sum_{i=1}^n a_i K_{x_i} \in G, \quad g = \sum_{j=1}^m b_j K_{y_j} \in G$$

definieren wir

$$\langle f, g \rangle_G = \sum_{i=1}^n \sum_{j=1}^m a_i b_j K(x_i, y_j).$$

Da K symmetrisch und positiv definit ist, gilt:

- $\langle \cdot, \cdot \rangle_G$ ist symmetrisch und bilinear.
- Aus der positiven Definitheit von K folgt $\langle f, f \rangle_G \geq 0$ für alle $f \in G$ und aus $\langle f, f \rangle_G = 0$ folgt $f = 0$.

Damit ist $\langle \cdot, \cdot \rangle_G$ ein Skalarprodukt auf G .

Reproduktionseigenschaft: Für $f = \sum_{i=1}^n a_i K_{x_i} \in G$ und $x \in X$ gilt:

$$f(x) = \sum_{i=1}^n a_i K(x_i, x) = \sum_{i=1}^n \sum_{j=1}^1 a_i K(x_i, y_j) = \left\langle \sum_{i=1}^n a_i K_{x_i}, K_x \right\rangle_G = \langle f, K_x \rangle_G,$$

wobei $y_1 := x$ gesetzt wurde. Wir haben damit G als Prähilbertraum mit reproduzierendem Kernel konstruiert.

Vervollständigung: Um eine RKHS zu erhalten, müssen wir den Raum vervollständigen. Dazu betrachten wir den Raum aller Cauchy-Folgen in G und definieren eine Äquivalenzrelation

$$(f_n) \sim (g_n) :\iff \|f_n - g_n\|_G \xrightarrow{n \rightarrow \infty} 0.$$

Die Vervollständigung H ist dann der Quotientenraum

$$H := \{(f_n)_{n \geq 1} \text{ Cauchy in } G\} / \sim.$$

Für Äquivalenzklassen schreiben wir der Einfachheit halber $f = [f_n]$ und definieren das Skalarprodukt durch

$$\langle f, g \rangle_H = \lim_{n \rightarrow \infty} \langle f_n, g_n \rangle_G.$$

Im Folgenden überprüft man, dass H mit diesem Skalarprodukt tatsächlich ein Hilbertraum ist. Dazu müssen die folgenden Punkte gezeigt werden:

- 1) $\langle \cdot, \cdot \rangle_H$ ist wohldefiniert: Der Grenzwert $\lim_{n \rightarrow \infty} \langle f_n, g_n \rangle_G$ existiert und ist unabhängig von der Wahl der Repräsentanten.
- 2) $\langle \cdot, \cdot \rangle_H$ ist ein Skalarprodukt auf H .
- 3) $(H, \langle \cdot, \cdot \rangle_H)$ ist vollständig, also ein Hilbertraum.
- 4) $\langle f, K_x \rangle_H = f(x)$ lässt sich von G auf H übertragen
- 5) H ist eindeutig.

Damit ist H ein Hilbertraum mit reproduzierendem Kernel K . □

In Abbildung 3 sind die Entscheidungsgebiete einer SVM mit linearem, polynomialen und RBF-Kernel auf einem Ring-Datensatz dargestellt. In Abbildung 4 sieht man dieselben Kernel auf einem Halbmond-Datensatz. Abbildung 5 zeigt die durch den RBF-Kernel entstehende gekrümmte Entscheidungsfläche im dreidimensionalen Raum

3 Satz von Mercer

Wir haben nun gesehen, wie man mit nichtlinear trennbaren Daten umgehen kann. Durch einen geeigneten Kernel wird der Suchraum der SVM erweitert, sodass wieder eine lineare Trennung möglich wird, ohne dass der zugrunde liegende Raum explizit beschrieben werden muss. Dennoch stellt sich die Frage, wie dieser Raum strukturiert ist. Eine Antwort darauf liefert der Satz von Mercer. Zuvor formulieren wir einen Hilfssatz.

Satz 3.1. *Sei $B \subset \mathcal{X}$ kompakt und μ ein endliches Maß auf B . Zum Kernel K definieren wir*

$$T_K : L^2(B, \mu) \rightarrow L^2(B, \mu),$$

durch

$$T_K f : B \rightarrow \mathbb{R}, \quad x \mapsto T_K f(x) := \int_B K(x, x') f(x') \mu(dx').$$

Eine Abbildung $e : B \rightarrow \mathbb{R}$ heißt *Eigenfunktion* von T_K zum *Eigenwert* $\lambda \in \mathbb{R}$, wenn

$$(T_K e)(x) = \lambda e(x) \quad \forall x \in B.$$

Damit können wir den folgenden Satz formulieren.

Satz 3.2 (Mercer). *Sei K ein stetiger positiv definiter Kernel und μ ein endliches Maß auf einer kompakten Teilmenge $B \subset \mathcal{X}$. Dann gibt es eine Orthonormalbasis $(e_i)_{i \geq 1}$ von $L^2(B, \mu)$, die aus Eigenfunktionen von T_K mit nichtnegativen Eigenwerten $(\lambda_i)_{i \geq 1}$ besteht. Für alle i mit $\lambda_i > 0$ ist e_i stetig, und für alle $x, y \in B$ gilt*

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(y).$$

Hat T_K nur endlich viele von 0 verschiedene Eigenwerte $\lambda_1, \dots, \lambda_n$, so können wir K als Skalarprodukt in $\mathbb{R}^n$ schreiben:

$$K(x, y) = \sum_{i=1}^n \lambda_i e_i(x) e_i(y) = \langle e_\lambda(x), e_\lambda(y) \rangle_{\mathbb{R}^n},$$

wobei für $x \in B$

$$e_\lambda(x) := (\sqrt{\lambda_1} e_1(x), \dots, \sqrt{\lambda_n} e_n(x)) \in \mathbb{R}^n.$$

Zusammenfassung

Abschließend lässt sich sagen, dass Kernel-Methoden es ermöglichen, komplexe, im ursprünglichen Datenraum nichtlinear trennbare Strukturen mit linearen Verfahren in geeigneten Hilberträumen zu bearbeiten, die häufig hochdimensional oder sogar unendlichdimensional sind. Die Sätze von Moore–Aronszajn und Mercer bilden dabei die theoretische Grundlage, indem sie sicherstellen, dass jeder positiv definite Kernel als Skalarprodukt in einem eindeutig bestimmten reproduzierenden Kernel-Hilbertraum aufgefasst werden kann und eine zugehörige Eigenfunktionszerlegung besitzt.

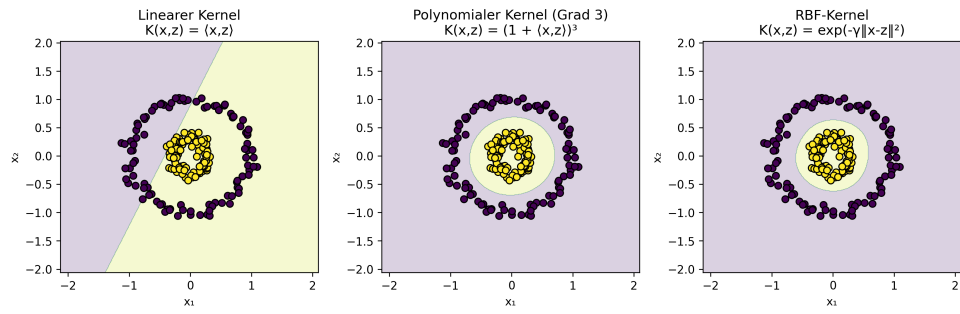


Figure 3: Verschiedene Kernel auf Ring-Datensatz

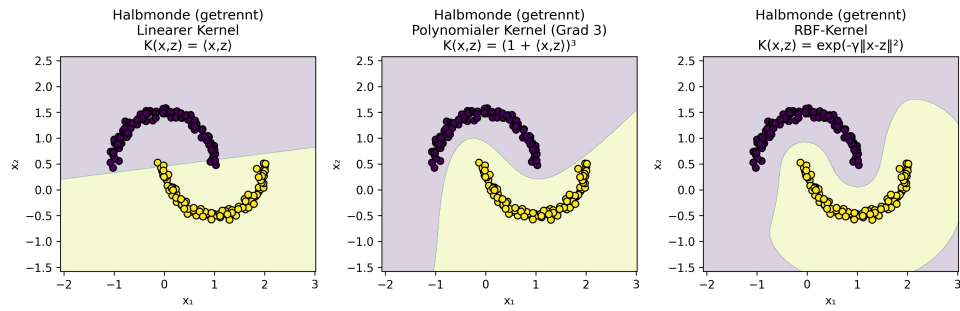


Figure 4: Verschiedene Kernel auf Halbmond-Datensatz

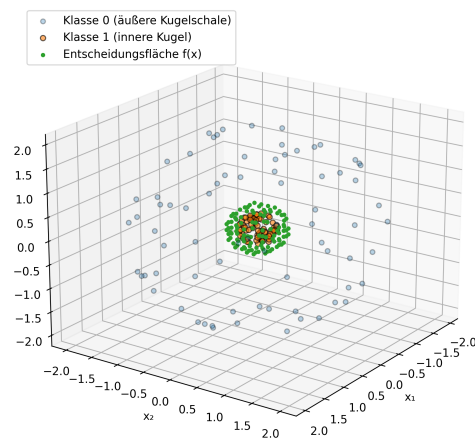


Figure 5: Entscheidungsfläche im 3-Dimensionalen Raum mit RBF-Kernel