Kapitel 9
Vektorklassifikatoren

Zusammenfassung In diesem Kapitel werden wir uns mit Klassifikatoren für Merkmalvektoren beschäftigen. Deren Betrachtung ist eine Voraussetzung für die Definition von Vektor*folgen*klassifikatoren, mit welchen sich Kapitel 10 befasst. Wir werden auf vier wichtige spezielle Vektorklassifikatoren eingehen: die Supportvektormaschine, den GMM-Klassifikator, künstliche neuronale Netze sowie unscharfe Klassifikatoren.

9.1 Supportvektormaschine

Wir haben in Abschnitt 8.3.4 das Neuron nach (8.126) als speziellen Linearklassifikator eingeführt. Eine Supportvektormaschine ist ein Neuron, das nach einem bestimmten Algorithmus trainiert wird. Die folgende Darstellung gibt einen Überblick über grundlegende Aspekte der Supportvektormaschine. Für Details sei auf die einschlägige Literatur verwiesen, beispielsweise [24, 25, 30, 7].

9.1.1 Duale Form des Linearklassifikators

Wir erinnern uns, dass die Trainingsaufgabe in der geeigneten Einstellung des Parametervektors $\vec{w}$ sowie der Verschiebung w_0 besteht. Dabei wird von einer klassifizierten Trainingsstichprobe nach (8.161) ausgegangen. Wir haben das Grundprinzip festgehalten, dass sich die Parameter der Trennfunktion als Linearkombination der Elemente der Lernstichprobe ergeben. Diesen Gedanken wollen wir zunächst etwas vertiefen und wiederholen dazu die Perzeptron-Lernregel nach (8.173) in leicht modifizierter Form

R. Hoffmann, M. Wolff, *Intelligente Signalverarbeitung 2,*
DOI 10.1007/978-3-662-46726-8_2, © Springer-Verlag Berlin Heidelberg 2015

$$\vec{w}^{(p+1)} = \begin{cases} \vec{w}^{(p)} + 0 \cdot \vec{o}^{(p)} c^{(p)} : c^{(p)} = s(\vec{o}^{(p)}) \\ \vec{w}^{(p)} + 2\alpha\, \vec{o}^{(p)} c^{(p)} : c^{(p)} \neq s(\vec{o}^{(p)}) \end{cases}, \tag{9.1}$$

wobei wir jedem Vektor der Lernstichprobe eine individuelle Lernkonstante $\alpha^{(p)}$ zugewiesen haben. Wir können also anstelle von (9.1) auch

$$\vec{w}^{(p+1)} = \vec{w}^{(p)} + \alpha^{(p)} c^{(p)} \vec{o}^{(p)} \qquad \text{mit } \alpha^{(p)} = \begin{cases} 0 & : c^{(p)} = s(\vec{o}^{(p)}) \\ 2\alpha & : c^{(p)} \neq s(\vec{o}^{(p)}) \end{cases} \tag{9.2}$$

schreiben. Um die $\alpha^{(p)}$ kümmern wir uns zunächst nicht weiter. Sie werden später der eigentliche Gegenstand des Trainings der Supportvektormaschine sein. Wenn wir von einem initialen Parametervektor $\vec{w}^{(0)} = \vec{0}$ ausgehen, erhalten wir nach Verbesserung mit allen Elementen der Lernstichprobe

$$\vec{w}^{(n)} = \sum_{p=1}^{n} \alpha^{(p)} c^{(p)} \vec{o}^{(p)}. \tag{9.3}$$

Diesen Ausdruck setzen wir nun in (8.122) ein und erhalten die Unterscheidungsfunktion der Supportvektormaschine

$$d(\vec{o}) = \sum_{p=1}^{n} \alpha^{(p)} c^{(p)} \vec{o}^{(p)\top} \vec{o} + w_0 = \sum_{p=1}^{n} \alpha^{(p)} c^{(p)} \left\langle \vec{o}^{(p)}, \vec{o} \right\rangle + w_0. \tag{9.4}$$

Diese Gleichung wird auch als duale Form des Linearklassifikators bezeichnet. Wir stellen fest, dass sie die Merkmalvektoren der Lernstichprobe *ausschließlich* innerhalb eines Skalarprodukts verwendet – ein Umstand, der später von großer Bedeutung sein wird.

9.1.2 Supportvektoren

Von den verschiedenen algorithmischen Möglichkeiten, die Trennfunktion aus einer Lernstichprobe zu ermitteln, haben wir die Perzeptron- und Adaline-Lernverfahren vorgestellt. Wie der Vergleich von (8.170) und (8.197) zeigt, ermitteln beide die Parameter der Trennfunktion als Linearkombination der Lernstichprobenelemente. Sie unterscheiden sich letztlich nur in der Wahl individueller Lernkonstanten $\alpha^{(p)}$.[1] Beide finden, lineare Separierbarkeit der Trainingsstichprobe vorausgesetzt, eine Trennfunktion, mit der alle Trainingsvektoren korrekt klassifiziert werden. Nach diesem Kriterium ist die Lösung der Trainingsaufgabe, falls sie existiert, aber nicht eindeutig, wie Abbildung 9.1 zeigt. Das Ergebnis des heuristischen Perzeptron-Lernalgorithmus hängt sogar von der Reihenfolge ab, in der die Elemente der Trainingsstich-

[1] nämlich beim Perzeptron $\alpha^{(p)} = -\alpha(\operatorname{sgn} \vec{w}^{(p)\top} \vec{o}^{(p)} - c^{(p)})$ und beim Adaline $\alpha^{(p)} = -2\alpha(\vec{w}^{(p)\top} \vec{o}^{(p)} + w_0^{(p)} - c^{(p)})$

probe angeboten werden. Ist die Lernstichprobe nicht linear separierbar, findet der Algorithmus überhaupt keine vernünftige Lösung. Der Adaline-Algorithmus minimiert den mittleren quadratischen Klassifikationsfehler und findet nach diesem Kriterium auch dann eine „optimale“ Lösung, wenn die Lernstichprobe nicht linear separierbar ist. Da wegen der Unstetigkeit der Vorzeichenfunktion eine modifizierte Entscheidungsfunktion verwendet werden musste, ist das Ergebnis jedoch nicht befriedigend.

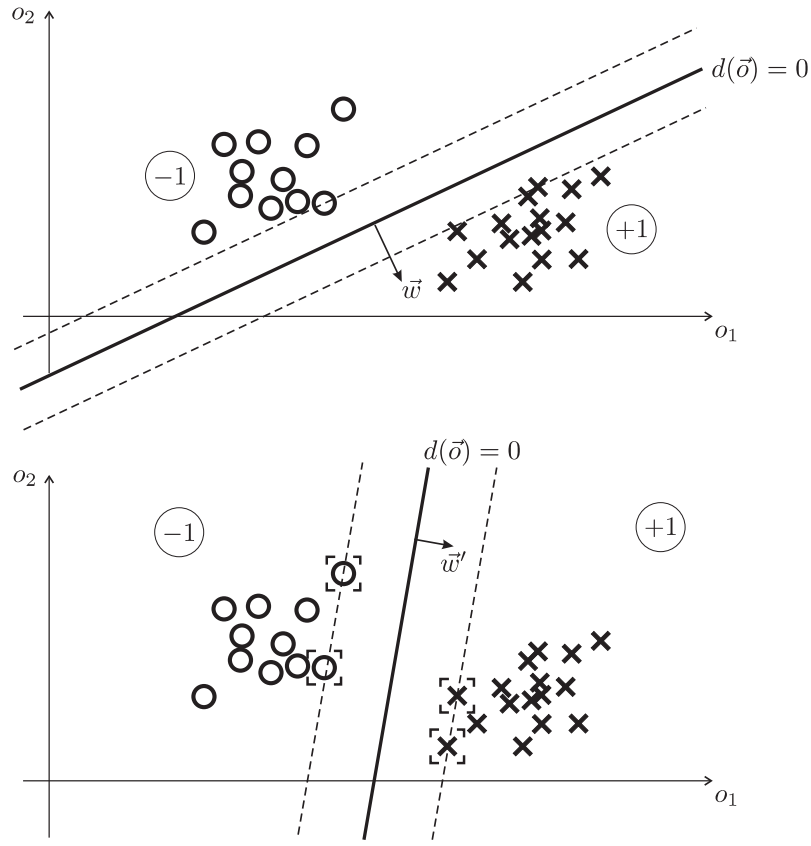


Abbildung 9.1 Klassifizierte Vektorstichprobe mit zwei möglichen linearen Trennfunktionen. Im unteren Diagramm wird ein maximaler Abstand der Trenngeraden von den Vertretern beider Klassen realisiert.

Der Ausgangspunkt der Herleitung der Supportvektormaschine ist die Definition eines leistungsfähigen Zielkriteriums für das Lernen. Wir betrachten dazu noch einmal Abbildung 9.1. Sowohl das obere als auch das untere Diagramm stellen gültige Lösungen für das Trainingsproblem dar. Wir müssen aber davon ausgehen, dass unsere Trainingsstichprobe unvollständig ist. Es kann also gut sein, dass weitere Vertreter der Klassen *neben* und nicht *in* den

sichtbaren Ballungsgebieten liegen. Wir wünschen uns daher eine Trennfunktion, die eine möglichst große Sicherheit für die Klassifikation von Ausreißern beider Klassen bietet. Diese würden aber nur dann ein Problem für den Klassifikator darstellen, wenn sie in Richtung der anderen Klasse lägen. Genaueres lässt sich nicht sagen, da wir nur unsere Trainingsstichprobe kennen. So ist die einzige Möglichkeit, unseren Wunsch nach möglichst großer Klassifikationssicherheit zu erfüllen, die Wahl einer Trennfunktion, die in möglichst großem Abstand von den bekannten Klassenvertretern verläuft und somit Ausreißern soviel Platz wie möglich bietet. Im Sinne der statistischen Lerntheorie ist dieses Kriterium sogar optimal, was bewiesen werden kann. Wir wollen diesen Beweis hier aber nicht führen und verweisen anstelle dessen auf die weiterführende Literatur (zum Beispiel [29, 30, 7]).

Wenn wir einen möglichst großen Abstand der Trennfunktion von den bekannten Klassenvertretern fordern, so fließen in dessen Ermittlung offensichtlich nur sehr wenige von ihnen ein. Von Bedeutung sind ausschließlich diejenigen, die am weitesten in Richtung der jeweils anderen Klasse liegen. Diese Elemente der Lernstichprobe heißen Supportvektoren. Im unteren Diagramm von Abbildung 9.1 sind sie mit Markierungen versehen. Genau genommen, können wir alle anderen Elemente der Lernstichprobe beim Training weglassen, da sie auf den Wert unserer Zielfunktion keinen Einfluss haben. Mathematisch bedeutet das, dass die meisten Lernkonstanten Null sind. Das einzige Problem besteht nun darin, die Supportvektoren in der Lernstichprobe zu finden. Das ist aber leider nicht trivial.

9.1.3 Geometrischer und funktionaler Rand, Randmaximierung

Zur Definition der Zielfunktion für das Training suchen wir also die Abstände $\gamma^{(p)}$ der Lernvektoren $\vec{o}^{(p)}$ von der Trennfunktion $d(\vec{o}) = \vec{w}^\top \vec{o} + w_0 = 0$. Diese werden auch als *geometrische* Abstände bezeichnet. Abbildung 9.2 veranschaulicht das Problem für je einen Vektor beider Klassen. Zur Berechnung des geometrischen Abstandes eines Vektors fällen wir das Lot von dessen Endpunkt auf die Trennfunktion. Die Länge des Lots ist genau der gesuchte Wert. Wir erhalten einen Fußpunkt

$$\vec{q}^{(p)} = \vec{o}^{(p)} - \gamma^{(p)} \cdot \frac{\vec{w}}{\|\vec{w}\|}, \tag{9.5}$$

wenn der Vektor links, und

$$\vec{q}^{(p)} = \vec{o}^{(p)} + \gamma^{(p)} \cdot \frac{\vec{w}}{\|\vec{w}\|}, \tag{9.6}$$

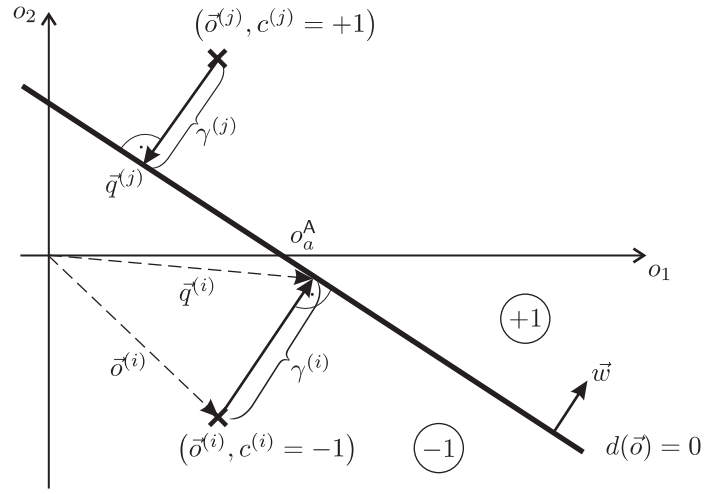


Abbildung 9.2 Zur Berechnung der geometrischen Abstände $\gamma^{(i)}$ und $\gamma^{(j)}$ für zwei Merkmalvektoren $\vec{o}^{\,(i)}$ und $\vec{o}^{\,(j)}$ von unterschiedlicher Klasse.

wenn der Vektor rechts der Trenngerade liegt. Da im ersten Fall $c^{(p)} = +1$ und im zweiten Fall $c^{(p)} = -1$ gilt, können wir beide Gleichungen zusammenfassen:

$$\vec{q}^{(p)} = \vec{o}^{(p)} - c^{(p)}\gamma^{(p)}\frac{\vec{w}}{\|\vec{w}\|}. \tag{9.7}$$

Der Fußpunkt liegt auf der Trenngeraden. Wir können $\gamma^{(p)}$ also wie folgt berechnen:

$$0 = \vec{w}^\top \vec{q}^{(p)} + w_0 = \vec{w}^\top \left(\vec{o}^{(p)} - c^{(p)}\gamma^{(p)}\frac{\vec{w}}{\|\vec{w}\|}\right) + w_0 \tag{9.8}$$

$$0 = \vec{w}^\top \vec{o}^{(p)} + w_0 - c^{(p)}\gamma^{(p)}\frac{\overbrace{\vec{w}^\top \vec{w}}^{=\|\vec{w}\|^2}}{\|\vec{w}\|}. \tag{9.9}$$

Durch Umstellen erhalten wir

$$\vec{w}^\top \vec{o}^{(p)} + w_0 = c^{(p)}\gamma^{(p)}\|\vec{w}\| \tag{9.10}$$

sowie schließlich

$$\gamma^{(p)} = \frac{1}{c^{(p)}\|\vec{w}\|}\left(\vec{w}^\top \vec{o}^{(p)} + w_0\right), \tag{9.11}$$

was sich wegen $c^{(p)} = \frac{1}{c^{(p)}}$ zum gesuchten Ausdruck für den geometrischen Abstand vereinfacht:

$$\gamma^{(p)} = c^{(p)} \frac{\vec{w}^\top \vec{o}^{(p)} + w_0}{\|\vec{w}\|} = c^{(p)} \frac{d(\vec{o}^{(p)})}{\|\vec{w}\|}. \tag{9.12}$$

Der kleinste geometrische Abstand wird als geometrischer Rand bezeichnet:

$$\gamma = \min_p \gamma^{(p)}. \tag{9.13}$$

Wenn der geometrische Abstand eines Lernvektors positiv ist, wird dieser korrekt klassifiziert. Ist er negativ, wird er falsch klassifiziert:

$$\gamma^{(p)} \begin{cases} > 0 : \text{korrekte Klassifikation}, \\ < 0 : \text{falsche Klassifikation}. \end{cases} \tag{9.14}$$

Ist der geometrische Rand positiv, kann die Lernstichprobe linear separiert werden. Ist er negativ, kann die Lernstichprobe nicht linear separiert werden:

$$\gamma \quad \begin{cases} > 0 : \text{Lernstichprobe linear separierbar}, \\ < 0 : \text{Lernstichprobe nicht linear separierbar}. \end{cases} \tag{9.15}$$

Ziel des Trainings ist, wie oben erläutert, die Maximierung des geometrischen Randes

$$\gamma \rightarrow \ \text{Max}\,(\vec{w}, w_0)! \tag{9.16}$$

Leider ist diese Extremwertaufgabe nicht eindeutig lösbar, wie Abbildung 9.3 an einem eindimensionalen Beispiel verdeutlicht. Wir finden unendlich viele Trennfunktionen

$$C\vec{w}^\top \vec{o} + Cw_0 \qquad \text{mit } C \neq 0 \in \mathbb{R} \tag{9.17}$$

mit maximalem geometrischen Rand, die sich alle nur durch ihren Anstieg unterscheiden. Wir betrachten daher den Wert $d\left(\vec{o}^{(p)}\right)$ der Unterscheidungsfunktion über den Merkmalvektoren. Wie beim geometrischen Rand multiplizieren wir diesen Wert mit dem Klassenindex und erhalten den so genannten funktionalen Abstand

$$\hat{\gamma}^{(p)} = c^{(p)} d(\vec{o}^{(p)}) = c^{(p)} \left(\vec{w}^\top \vec{o}^{(p)} + w_0\right). \tag{9.18}$$

Mit (9.12) können wir den funktionalen Abstand auch als

$$\hat{\gamma}^{(p)} = \|\vec{w}\| \cdot \gamma^{(p)} \tag{9.19}$$

ausdrücken. Wie oben nennen wir den kleinsten funktionalen Abstand aller Lernvektoren funktionaler Rand:

$$\hat{\gamma} = \min_p \hat{\gamma}^{(p)}. \tag{9.20}$$

Abbildung 9.4 veranschaulicht den Zusammenhang zwischen Unterscheidungsfunktion $d(\vec{o})$, Trennfunktion $d(\vec{o}) = 0$, geometrischem und funktio-

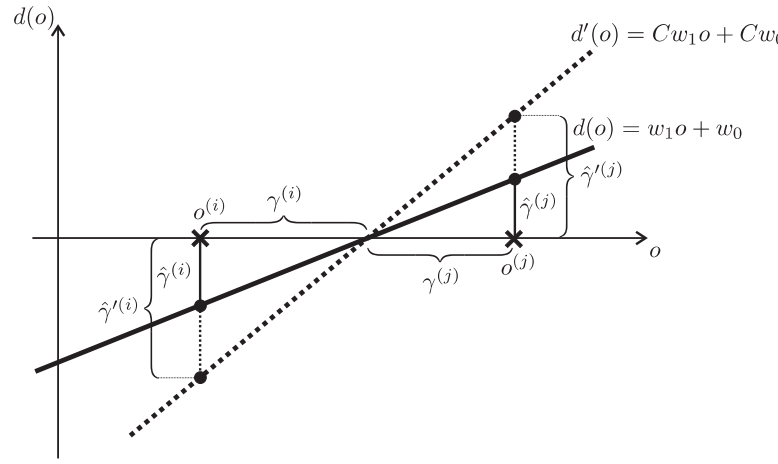


Abbildung 9.3 Zum Zusammenhang zwischen funktionalem und geometrischem Rand.

nalem Rand. Wir hatten oben festgestellt, dass die Maximierung des geometrischen Randes keine eindeutige Lösung für die Parameter der Unterscheidungsfunktion ergibt, sondern dass eine Schar von Funktionen nach (9.17) optimal ist. Wir können den störenden Freiheitsgrad C mit einem einfachen Trick beseitigen: Wir fordern einfach, dass der funktionale Rand genau Eins ist

$$\hat{\gamma} := 1. \tag{9.21}$$

Damit ist C eindeutig bestimmt und die Maximierung erbringt eine einzige Lösung. Nach der Definition (9.20) des funktionalen Randes gilt mit (9.21) für alle funktionalen Abstände

$$\hat{\gamma}^{(p)} \geq 1 \quad \forall p. \tag{9.22}$$

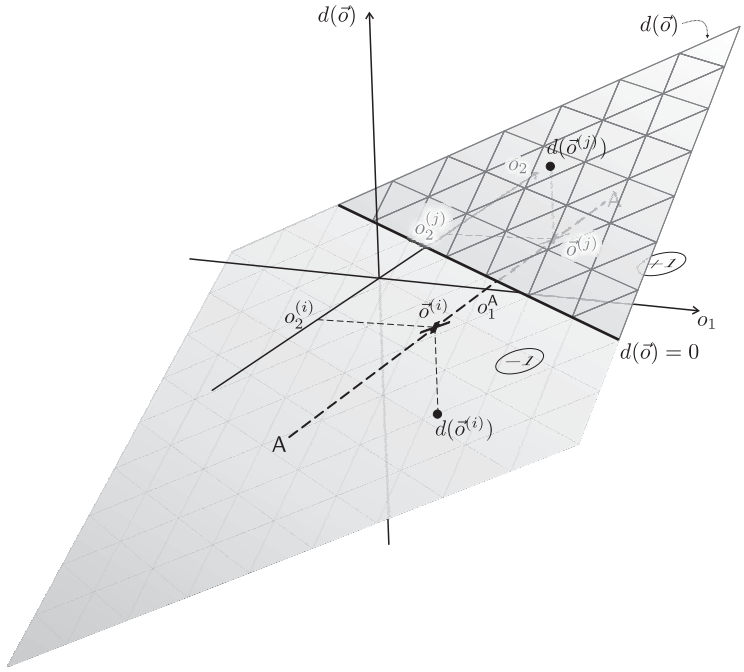
Da aus (9.19)

$$\gamma = \frac{\hat{\gamma}}{\|\vec{w}\|} \tag{9.23}$$

folgt und wir $\hat{\gamma}$ bei Eins festhalten wollen, ist die Maximierung des geometrischen Randes gleichbedeutend mit der Minimierung der Länge des Parametervektors $\|\vec{w}\|$. Wiederum gleichbedeutend damit, aber einfacher berechenbar, ist die Minimierung des Quadrates dieser Länge, so dass sich schließlich die Extremwertaufgabe

$$\frac{1}{2}\|\vec{w}\|^2 = \frac{1}{2}\vec{w}^\top\vec{w} \rightarrow \text{ Min}\,(\vec{w}, w_0)! \tag{9.24}$$

ergibt. Wir müssen dabei beachten, dass diese den Nebenbedingungen (9.22)



Draufsicht: geometrischer Rand

Schnitt A – A: funktionaler Rand

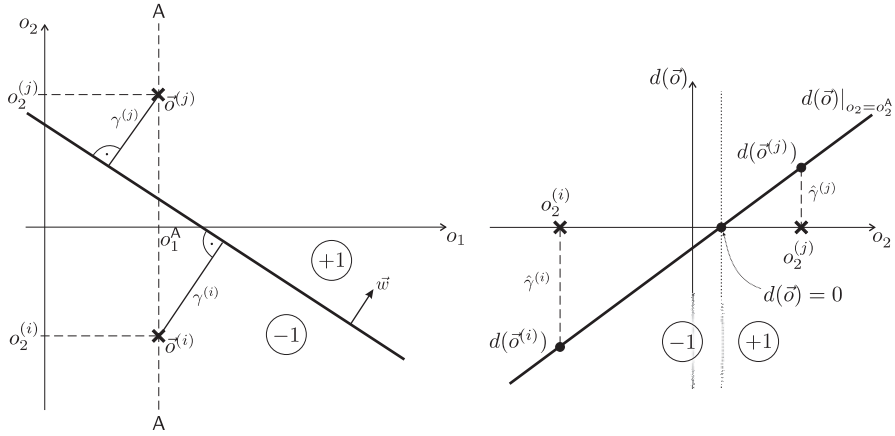


Abbildung 9.4 Zum Zusammenhang zwischen funktionalem Rand, geometrischem Rand und Unterscheidungsfunktion.

$$\hat{\gamma}^{(p)} = c^{(p)} \left(\vec{w}^\top \vec{o}^{(p)} + w_0 \right) \geq 1 \quad \forall p \tag{9.25}$$

beziehungsweise

$$1 - c^{(p)} \left(\vec{w}^\top \vec{o}^{(p)} + w_0 \right) \leq 0 \quad \forall p \tag{9.26}$$

unterliegt. Dieses Problem kann mit Hilfe der in Abschnitt 8.4.4 vorgestellten konvexen Optimierung gelöst werden, da sowohl die Zielfunktion als auch die Nebenbedingungen konvex sind. Wir setzen also die LAGRANGE-Funktion

$$L(\vec{w}, w_0, \vec{\alpha}) = \frac{1}{2} \vec{w}^\top \vec{w} + \sum_{p=1}^{n} \alpha^{(p)} \left[1 - c^{(p)} \left(\vec{w}^\top \vec{o}^{(p)} + w_0 \right) \right] \tag{9.27}$$

an, verfolgen die duale Lösung (siehe Abschnitt 8.4.4)

$$(\vec{w}^*, w_0^*, \vec{\alpha}^*)^d : \underbrace{\left(L(\vec{w}, w_0, \vec{\alpha}) \to \mathrm{Min}(\vec{w}, w_0) \right)}_{L(\vec{w}^*, w_0^*, \vec{\alpha})} \to \mathrm{Max}(\vec{\alpha})! \tag{9.28}$$

und führen die innere Minimierung durch partielles Ableiten und Nullsetzen aus. Für den Parametervektor $\vec{w}$ erhalten wir

$$\begin{aligned} &\left. \frac{\partial L(\vec{w}, w_0, \vec{\alpha})}{\partial \vec{w}} \right|_{\vec{w}^*} = \vec{w} - \textstyle\sum_p \alpha^{(p)} c^{(p)} \vec{o}^{(p)} = 0 \\ \rightsquigarrow \quad &\vec{w}^* = \textstyle\sum_p \alpha^{(p)} c^{(p)} \vec{o}^{(p)}. \end{aligned} \tag{9.29}$$

Die optimale Lösung $\vec{w}^*$ entspricht also genau der dualen Form des Linearklassifikators! Die LAGRANGE-Multiplikatoren sind *identisch* mit den Lernkonstanten, welche also „nebenbei“ mitbestimmt werden. Das Ableiten nach der Verschiebung w_0 und Nullsetzen ergibt die Bedingung

$$\left. \frac{\partial L(\vec{w}, w_0, \vec{\alpha})}{\partial w_0} \right|_{w_0^*} = - \sum_p \alpha^{(p)} c^{(p)} = 0. \tag{9.30}$$

Der zweite Schritt der dualen Lösung ist die äußere Maximierung

$$L(\vec{w}^*, w_0^*, \vec{\alpha}) \ \to \mathrm{Max}(\vec{\alpha})! \tag{9.31}$$

Wir setzen die Lösung (9.29) der inneren Minimierung bezüglich $\vec{w}$ in die LAGRANGE-Funktion ein und vereinfachen den Ausdruck so weit wie möglich:

$$L(\vec{w}^*, w_0^*, \vec{\alpha}) = \frac{1}{2} \vec{w}^{*\top} \vec{w}^* + \sum_p \alpha^{(p)} \left[1 - c^{(p)} \left(\vec{w}^{*\top} \vec{o}^{(p)} + w_0^* \right) \right] \tag{9.32}$$

$$L(\vec{w}^*, w_0^*, \vec{\alpha}) = \tfrac{1}{2} \underbrace{\vec{w}^{*\top} \vec{w}^*}_{\mathsf{A}} + \textstyle\sum_p \alpha^{(p)} - \sum_p \alpha^{(p)} c^{(p)} \underbrace{\vec{w}^{*\top}}_{\mathsf{B}} \vec{o}^{(p)} - w_0^* \underbrace{\sum_p \alpha^{(p)} c^{(p)}}_{=0 \text{ wg. } (9.30)}. \tag{9.33}$$

In die Terme A und B setzen wir jeweils das Ergebnis von (9.29) ein:

$$\mathsf{A} = \left(\sum_i \alpha^{(i)} c^{(i)} \vec{o}^{(i)} \right)^\top \cdot \sum_j \alpha^{(j)} c^{(j)} \vec{o}^{(j)} \quad \text{und} \quad \mathsf{B} = \sum_q \alpha^{(q)} c^{(q)} \vec{o}^{(q)}. \tag{9.34}$$

Damit wird (9.33) nach Umsortieren zu

$$L(\vec{w}^*, w_0^*, \vec{\alpha}) = \frac{1}{2} \sum_i \sum_j \alpha^{(i)} \alpha^{(j)} c^{(i)} c^{(j)} \vec{o}^{(i)\top} \vec{o}^{(j)} + \sum_p \alpha^{(p)} \tag{9.35}$$

$$- \sum_p \sum_q \alpha^{(p)} \alpha^{(q)} c^{(p)} c^{(q)} \vec{o}^{(q)\top} \vec{o}^{(p)}$$

$$= \sum_p \alpha^{(p)} - \frac{1}{2} \sum_p \sum_q \alpha^{(p)} \alpha^{(q)} c^{(p)} c^{(q)} \vec{o}^{(q)\top} \vec{o}^{(p)}. \tag{9.36}$$

Nach Ausklammern der Summe in p erhalten wir schließlich wegen $\vec{o}^{(q)\top} \vec{o}^{(p)} = \vec{o}^{(p)\top} \vec{o}^{(q)} = \langle \vec{o}^{(p)}, \vec{o}^{(q)} \rangle$ folgenden Ausdruck zur Maximierung:

$$L(\vec{w}^*, w_0^*, \vec{\alpha}) = \sum_p \left(\alpha^{(p)} - \frac{1}{2} \sum_q \alpha^{(p)} \alpha^{(q)} c^{(p)} c^{(q)} \langle \vec{o}^{(p)}, \vec{o}^{(q)} \rangle \right). \tag{9.37}$$

9.1.4 Hard-Margin-Supportvektormaschine

Eine weitere Herleitung ist allgemein nicht möglich. Wir definieren die einfachste Supportvektormaschine daher also wie folgt:

Ein Linearklassifikator mit $d(\vec{o}) = \vec{w}^{*\top} \vec{o} + w_0^*$ realisiert einen maximalen Rand auf einer Lernstichprobe $\left\{ (\vec{o}^{(p)}, c^{(p)}) \right\}$, wenn gilt

$$\vec{w}^* = \sum_{p=1}^{n} \alpha^{*(p)} c^{(p)} \vec{o}^{(p)} \tag{9.38}$$

und

$$w_0^* = -\frac{1}{2} \vec{w}^{*\top} \left(\vec{o}^{(j)} + \vec{o}^{(i)} \right) \tag{9.39}$$

mit zwei beliebigen Supportvektoren $\vec{o}^{(i)} \in SV_{+1}$ und $\vec{o}^{(j)} \in SV_{-1}$, wobei $\alpha^{*(p)}$ die Ergebnisse sind aus

http://www.springer.com/978-3-662-46725-1

Intelligente Signalverarbeitung 2
Signalerkennung
Hoffmann, R.; Wolff, M.
2015, XVI, 345 S. 107 Abb., Hardcover
ISBN: 978-3-662-46725-1