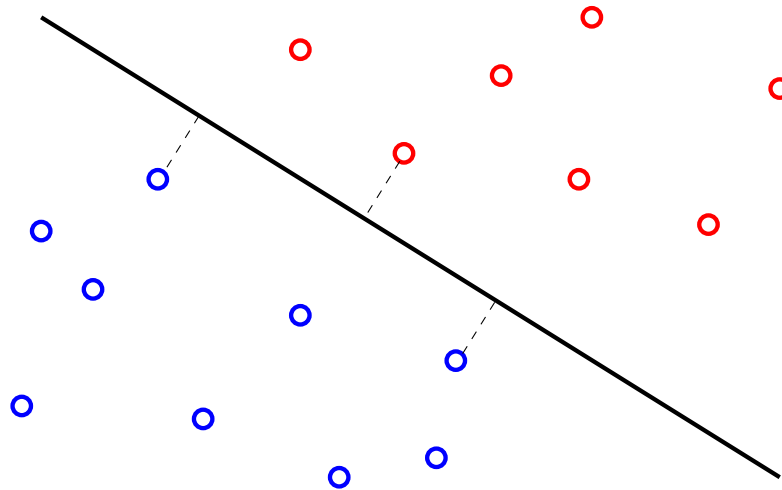


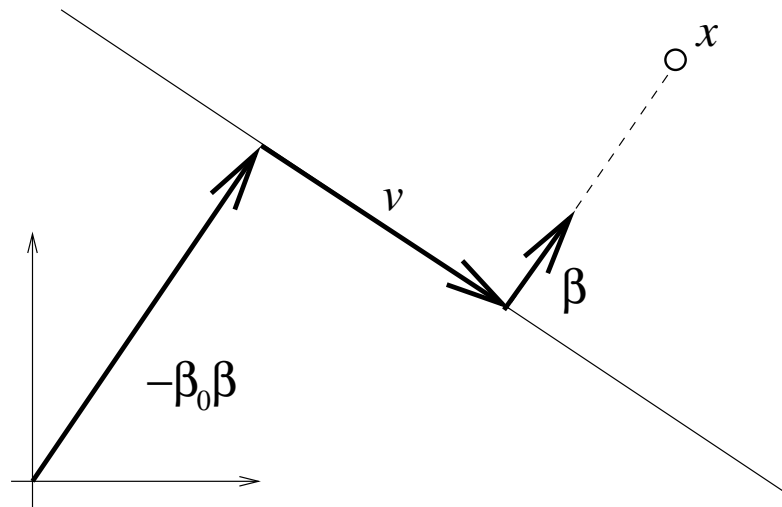
3.3 Lineare Support Vector Klassifikation

gegeben: $(x_1, y_1), \dots, (x_m, y_m)$ mit $x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$.

gesucht: Trennende Ebene $\{z \in \mathbb{R}^n \mid \langle z, \beta \rangle + \beta_0 = 0\}$, $(\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R})$, die möglichst großen Abstand zum nächstgelegenen Punkt hat, d.h.:

$$(\beta, \beta_0) = \arg \max_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}, \|\beta\|=1} \{C \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq C\}$$





$$\langle x, \beta \rangle + \beta_0 = \langle -\beta_0\beta + v + c\beta, \beta \rangle + \beta_0 = -\beta_0\langle \beta, \beta \rangle + \langle v, \beta \rangle + c\langle \beta, \beta \rangle + \beta_0 = c$$

$$\arg \max_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}, \|\beta\|=1} \{C \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq C\}$$

zu berechnen ist äquivalent zur Berechnung von

$$\arg \min_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}} \{\|\beta\| \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1\}$$

(neue (β_0, β) entsprechen alten $(\beta_0/C, \beta/C)$)

Das ist effizient lösbar, denn es ist ein

konvexes Optimierungsproblem,

d.h. eine konvexe Funktion soll auf einer konvexen Menge optimiert werden.

konvexe Menge $M \subseteq \mathbb{R}^n$:

$$x, y \in M \implies \forall p \in [0, 1] : p \cdot x + (1 - p) \cdot y \in M$$

konvexe Funktion $f : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$\forall x, y \in \mathbb{R}^n, p \in [0, 1] : f(p \cdot x + (1 - p) \cdot y) \leq p \cdot f(x) + (1 - p) \cdot f(y)$$

Ist f konvexe Funktion auf konvexer Menge, so sind alle lokalen Optima globale Optima und bilden eine konvexe Menge.

Lemma 1 Sei A eine positiv semidefinite $n \times n$ -Matrix (d.h. für alle $v \in \mathbb{R}^n$ gilt $v^T Av \geq 0$), dann ist $f : \mathbb{R}^n \rightarrow \mathbb{R}, v \mapsto v^T Av$ konvex.

Beweis: für $p \in [0, 1]$ und $v, w \in \mathbb{R}^n$ gilt:

$$\begin{aligned} & (pv + (1-p)w)^T A(pv + (1-p)w) \\ &= p^2 v^T Av + p(1-p)w^T Av + pv^T A(1-p)w + (1-p)^2 w^T Aw \\ &= pv^T Av + (1-p)w^T Aw - (1-p)p(v-w)^T A(v-w) \\ &\leq pv^T Av + (1-p)w^T Aw \end{aligned}$$

□

Spezialfall: $v \mapsto \langle v, v \rangle$ ist konvex.

$\{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^n \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1\}$ ist konvexe Menge,
denn:

für $p \in [0, 1]$ und $(\alpha_0, \alpha), (\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^n$ mit $y_i \cdot (\langle x_i, \alpha \rangle + \alpha_0) \geq 1$ und
 $y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1$ gilt:

$$\begin{aligned} & y_i \cdot (\langle x_i, p\alpha + (1-p)\beta \rangle + (p\alpha_0 + (1-p)\beta_0)) \\ &= py_i \cdot (\langle x_i, \alpha \rangle + \alpha_0) + (1-p)y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \\ &\geq p \cdot 1 + (1-p) \cdot 1 = 1 \end{aligned}$$

Also ist die Suche nach

$$\begin{aligned} & \arg \min_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}} \{ \|\beta\| \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1 \} \\ &= \arg \min_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}} \left\{ \frac{1}{2} \|\beta\|^2 \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1 \right\} \end{aligned}$$

ein konvexes Optimierungsproblem.

Satz 8 (Karush-Kuhn-Tucker) Seien f, h_1, \dots, h_m konvexe Funktionen auf \mathbb{R}^n . x^* ist genau dann ein Minimum von f unter der Nebenbedingung $\forall_j h_j(x^*) \leq 0$, wenn es ein $\xi = (\xi_1, \dots, \xi_m) \in \mathbb{R}_{\geq 0}^m$ gibt, so dass gilt:

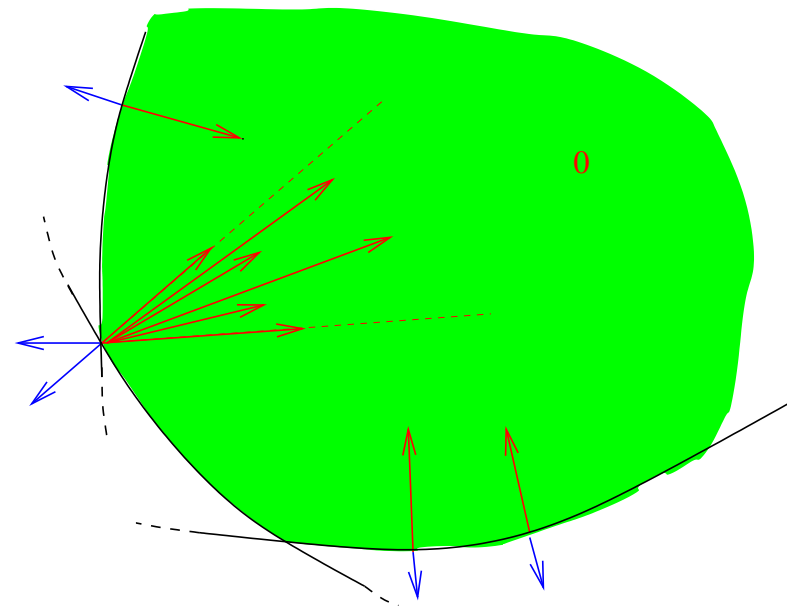
1. $\forall j : \xi_j \cdot h_j(x^*) = 0$

2. $df(x^*) = - \sum_{j=1}^m \xi_j \cdot dh_j(x^*)$

3. $\forall j : h_j(x^*) \leq 0$

(df ist der Gradient von f , d.h.: $df(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right)$)

Beweisersatz: Die zweite Bedingung besagt, dass der **Gradient von f eine Linearkombination der Gradienten der h_j** ist, und zwar nach der ersten Bedingung nur von solchen Richtungen, in denen wir auf dem Rand des **zulässigen Bereichs** sind.



Die **Gradienten der h_j** zeigen aus **diesem Bereich** heraus und stehen darauf senkrecht, und da $\xi_j \geq 0$ gefordert ist, gilt das auch für $-df(x^*)$. Also müsste man den Bereich verlassen, um den Funktionswert zu verringern. □

Anwendung zur Berechnung von

$\arg \min_{\beta \in \mathbb{R}^n, \beta_0 \in \mathbb{R}} \left\{ \frac{1}{2} \|\beta\|^2 \mid \forall i = 1, \dots, m : y_i \cdot (\langle x_i, \beta \rangle + \beta_0) \geq 1 \right\}$:

$$f(\beta_0, \beta_1, \dots, \beta_n) = \frac{1}{2} \|\beta\|^2 = \frac{1}{2} \sum_{i=1}^n \beta_i^2$$

$$h_j(\beta_0, \beta_1, \dots, \beta_n) = 1 - y_j (\langle x_j, \beta \rangle + \beta_0) = 1 - y_j \left(\sum_{i=1}^n x_{ji} \beta_i + \beta_0 \right)$$

\implies

$$df(\beta_0, \beta_1, \dots, \beta_n) = (0, \beta_1, \dots, \beta_n)$$

$$dh_j(\beta_0, \beta_1, \dots, \beta_n) = (-y_j, -y_j x_{j1}, \dots, -y_j x_{jn})$$

Anwendung der Karush-Kuhn-Tucker-Bedingungen auf

$$df(\beta_0, \beta_1, \dots, \beta_n) = (0, \beta_1, \dots, \beta_n)$$

$$dh_j(\beta_0, \beta_1, \dots, \beta_n) = (-y_j, -y_j x_{j1}, \dots, -y_j x_{jn})$$

zum finden des optimalen $(\beta_0^*, \beta_1^*, \dots, \beta_n^*)$:

1. für j mit $h_j(\beta_0^*, \beta_1^*, \dots, \beta_n^*) > 0$ gilt $\xi_j = 0$
2. $0 = \sum_{j=1}^m \xi_j y_j$ und $\forall i = 1, \dots, n : \beta_i^* = \sum_{j=1}^m \xi_j y_j x_{ji}$
3. $y_j (\langle x_j, \beta^* + \beta_0^* \rangle) \geq 1$

aus 1. und 2. folgt: $\beta_i^* = \sum_{\{j | y_j (\langle x_j, \beta^* \rangle - \beta_0^*) = 1\}} \xi_j y_j x_{ji}$

Satz 9 (Starker Dualitätssatz) Seien f, h_1, \dots, h_m konvex, L die verallgemeinerte Lagrange-Funktion

$$L(x, \xi) := f(x) + \sum_j \xi_j h_j(x)$$

und g die duale Lagrange-Funktion

$$g(\xi) := \inf_x L(x, \xi)$$

Dann gilt:

$$\max_{\{\xi | \forall j: \xi_j \geq 0\}} g(\xi) = \min_{\forall j: h_j(x) \leq 0} f(x)$$

sofern das Minimum existiert.

Beweis: Aus $\forall j : h_j(x) \leq 0 \wedge \xi_j \geq 0$ folgt

$$\forall x : g(\xi) = \inf_y f(y) + \sum_j \xi_j h_j(y) \leq f(x)$$

Ist x^* Minimalstelle von f und ξ^* zugehörige KKT-Multiplikatoren, so gilt

$$\sum_j \xi_j^* h_j(x^*) = 0$$

und damit x^* ist nach 2. Optimum vom $f(x) + \sum_j \xi_j^* h_j(x)$. Damit folgt $g(\xi^*) = f(x^*)$ □

in unserem Fall:

$$g(\xi) = \frac{1}{2} \|\beta\|^2 - \sum_{j=1}^m \xi_j y_j (\langle x_j, \beta \rangle + \beta_0) + \sum_{j=1}^m \xi_j$$

Für optimales ξ^* gilt nach vorherigen Ergebnissen:

$$\begin{aligned} g(\xi^*) &= \frac{1}{2} \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \\ &\quad - \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \\ &\quad - \beta_0 \sum_j \xi_j^* y_j + \sum_j \xi_j^* \\ &= \sum_j \xi_j^* - \frac{1}{2} \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \end{aligned}$$

Das maximierende ξ^* lässt sich effizient numerisch finden.

Aus den KKT-Bedingungen erhalten wir u.a. $\beta = \sum_{i=j}^m y_j \xi_j x_j \in V$.

Um ein neues x zu klassifizieren, berechnen wir das Vorzeichen von

$$\langle x, \beta \rangle + \beta_0$$

Das ist aber

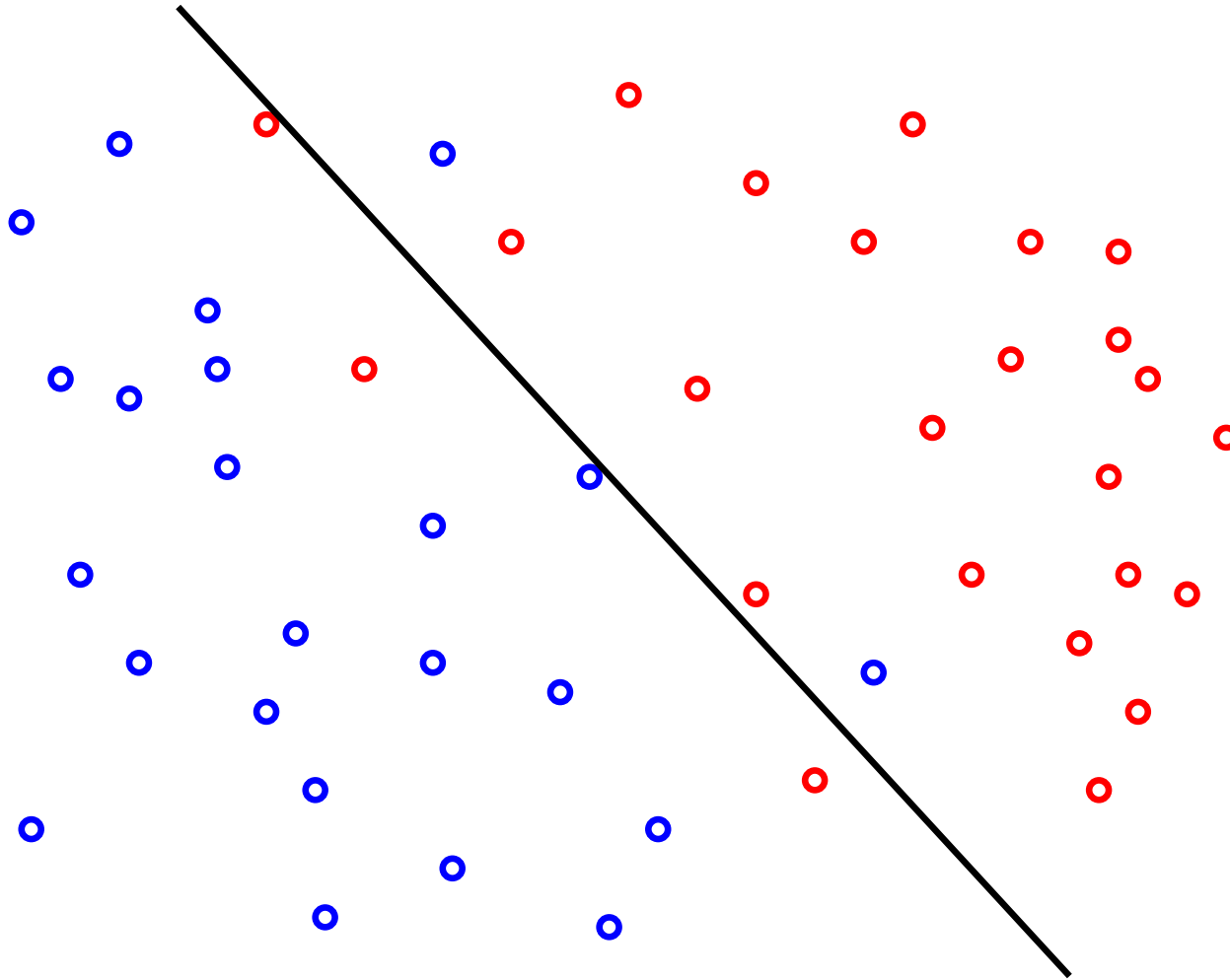
$$\langle x, \sum_{i=j}^m y_j \xi_j x_j \rangle + \beta_0 = \sum_{i=j}^m y_j \xi_j \langle x, x_j \rangle + \beta_0 = \sum_{i=j}^m y_j \xi_j \langle x, x_j \rangle + \beta_0$$

Dazu kann man also statt im \mathbb{R}^m auch im \mathbb{R}^n rechnen.

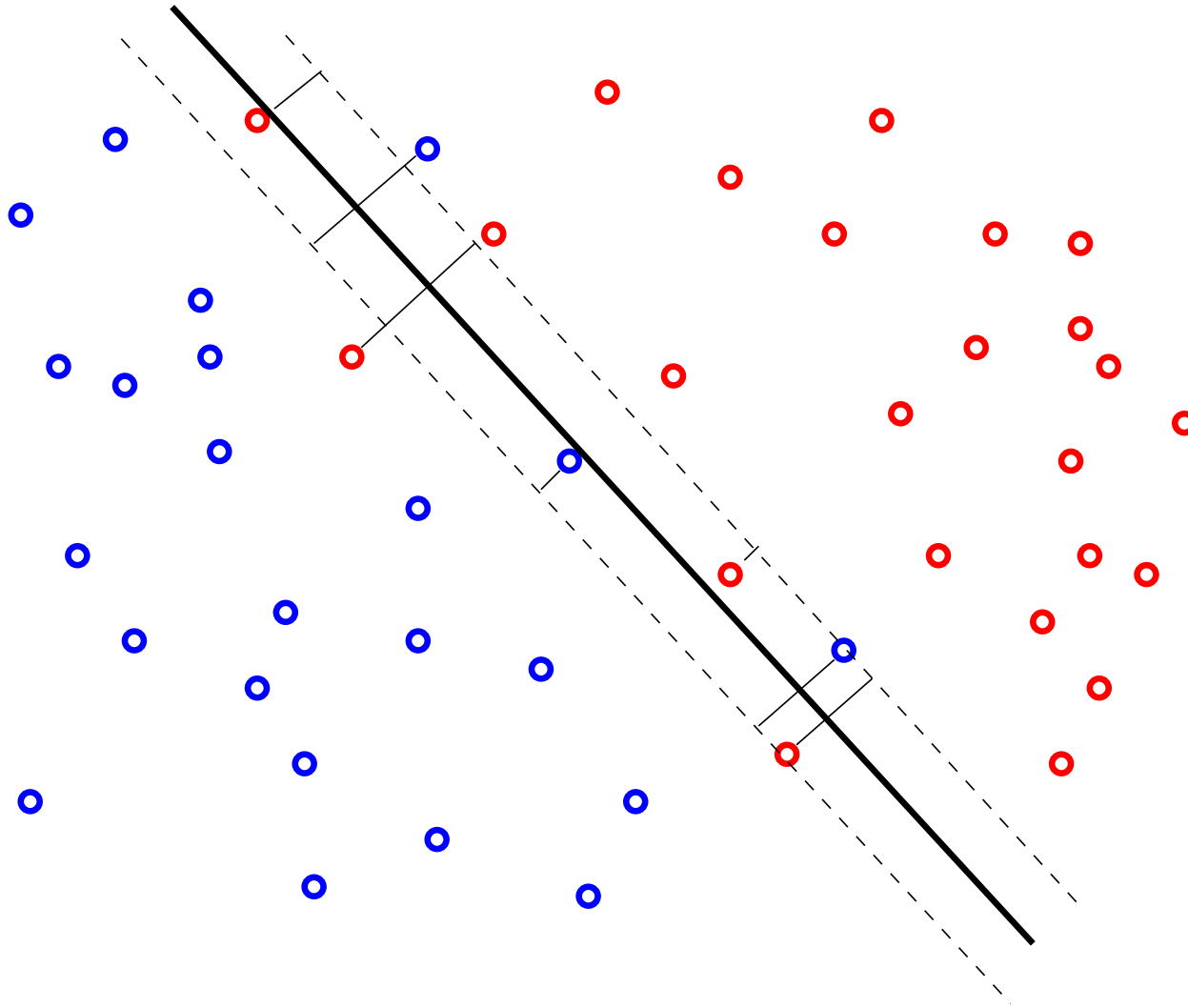
Die Dimension von ξ ist m , die von x und x_j ist n .

Soft Margin

Was ist wenn man nur bis auf Ausnahmen linear trennen kann?



Was ist wenn man nur bis auf Ausnahmen linear trennen kann?



Kriterium: Suche größtes C mit

$$y_i(\langle x_i, \beta \rangle + \beta_0) \geq C \cdot (1 - \delta_i)$$

wobei $\|\beta\| = 1$, $\delta_i \geq 0$ und $\sum_i \delta_i \leq \text{const}$.

Bzw. suche für gegebenes γ

$$\arg \min_{\beta, \beta_0} \left\{ \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^m \delta_i \mid \forall i : y_i (\langle x_i, \beta \rangle + \beta_0) \geq 1 - \delta_i \text{ und } \delta_i \geq 0 \right\}$$

Zu minimieren ist also die konvexe Funktion

$$f(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) = \frac{1}{2} \|\beta\|^2 + \gamma \sum_i \delta_i$$

auf der Menge $\{(\beta_0, \beta, \delta) \mid \forall i : h_i(\beta_0, \beta, \delta) \leq 0\}$, wobei für $j = 1, \dots, m$ gilt:

$$\begin{aligned} h_j(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) &= -y_j (\langle x_j, \beta \rangle + \beta_0) + 1 - \delta_j \\ h_{j+m}(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) &= -\delta_j \end{aligned}$$

Die Gradienten sind:

$$\begin{aligned} df(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) &= (0, \beta_1, \dots, \beta_n, \gamma, \dots, \gamma) \\ dh_j(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) &= (-y_j, -y_j x_{j1}, \dots, -y_j x_{jn}, 0, \dots, 0, -1, 0, \dots, 0) \\ dh_{j+m}(\beta_0, \beta_1, \dots, \beta_n, \delta_1, \dots, \delta_m) &= (0, \dots, 0, -1, 0, \dots, 0) \end{aligned}$$

Satz 8 (Karush-Kuhn-Tucker) Seien f, h_1, \dots, h_m konvexe Funktionen auf \mathbb{R}^n . x^* ist genau dann ein Minimum von f unter der Nebenbedingung $\forall_j h_j(x^*) \leq 0$, wenn es ein $\xi = (\xi_1, \dots, \xi_m) \in \mathbb{R}_{\geq 0}^m$ gibt, so dass gilt:

1. $\forall j : \xi_j \cdot h_j(x^*) = 0$

2. $df(x^*) = - \sum_{j=1}^m \xi_j \cdot dh_j(x^*)$

3. $\forall j : h_j(x^*) \leq 0$

(df ist der Gradient von f , d.h.: $df(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right)$)

Aus den ersten beiden Karush-Kuhn-Tucker-Bedingungen für $\xi = (\xi_1, \dots, \xi_{2m})$ erhalten wir für $j = 1, \dots, m$ und $i = 1, \dots, n$:

$$\xi_j \cdot (y_j(\langle x_j, \beta \rangle + \beta_0) - 1 + \delta_j) = 0$$

$$\xi_{j+m} \delta_j = 0$$

$$\sum_{j=1}^m \xi_j y_j = 0$$

$$\sum_{j=1}^m y_j x_{ji} \xi_j = \beta_i$$

$$\gamma = \xi_j + \xi_{j+m}$$

Satz 9 (Starker Dualitätssatz) Seien f, h_1, \dots, h_m konvex, L die verallgemeinerte Lagrange-Funktion

$$L(x, \xi) := f(x) + \sum_j \xi_j h_j(x)$$

und g die duale Lagrange-Funktion

$$g(\xi) := \inf_x L(x, \xi)$$

Dann gilt:

$$\max_{\{\xi | \forall j: \xi_j \geq 0\}} g(\xi) = \min_{\forall j: h_j(x) \leq 0} f(x)$$

sofern das Minimum existiert.

in unserem Fall:

$$g(\xi) = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{j=1}^m \delta_j - \sum_{j=1}^m \xi_j y_j (\langle x_j, \beta \rangle + \beta_0) + \sum_{j=1}^m \xi_j - \sum_{j=1}^m \xi_j \delta_j - \sum_{j=1}^m \xi_{j+1} \delta_j$$

Für optimales ξ^* gilt nach vorherigen Ergebnissen:

$$\begin{aligned} g(\xi^*) &= \frac{1}{2} \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \\ &\quad - \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \\ &\quad - \beta_0 \sum_j \xi_j^* y_j + \sum_j \xi_j^* \\ &\quad + \sum_{j=1}^m \delta_j (\gamma - \xi_j - \xi_{j+m}) \\ &= \sum_j \xi_j^* - \frac{1}{2} \sum_{j,k} \xi_j^* \xi_k^* y_j y_k \langle x_j, x_k \rangle \end{aligned}$$

Das ist die selbe Funktion g wie im Hard-Margin-Fall!

alternativer Ansatz: suche

$$\arg \min_{\beta, \beta_0} \left\{ \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^m \delta_i^2 \mid \forall i : y_i(\langle x_i, \beta \rangle + \beta_0) \geq 1 - \delta_i \text{ und } \delta_i \geq 0 \right\}$$

Hier ist das duale Problem: maximiere

$$g(\xi) = \sum_{j=1}^m \xi_j - \frac{1}{2} \sum_{j,k} \xi_j \xi_k y_j y_k \left(\langle x_j, x_k \rangle + \frac{\delta_{ij}}{\gamma} \right)$$

mit Nebenbedingung $\xi \geq 0$, $\sum_{j=1}^m \xi_j y_j = 0$, sowie $\delta_{ii} = 1$ und $\delta_{ij} = 0$ für $i \neq j$.
(Beweis als Übung)