

Lineare Regression

Ein Beispiel besteht nun aus einem Vektor $X = (X_1, \dots, X_p)$ und einer reellen Zahl $Y \in \mathbb{R}$. Wir wollen lernen, Y aus X vorherzusagen, und zwar durch eine affine Funktion

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$$

(Spezialfall: Thresholdfunktion, falls $Y \in \{-1, 1\}$)

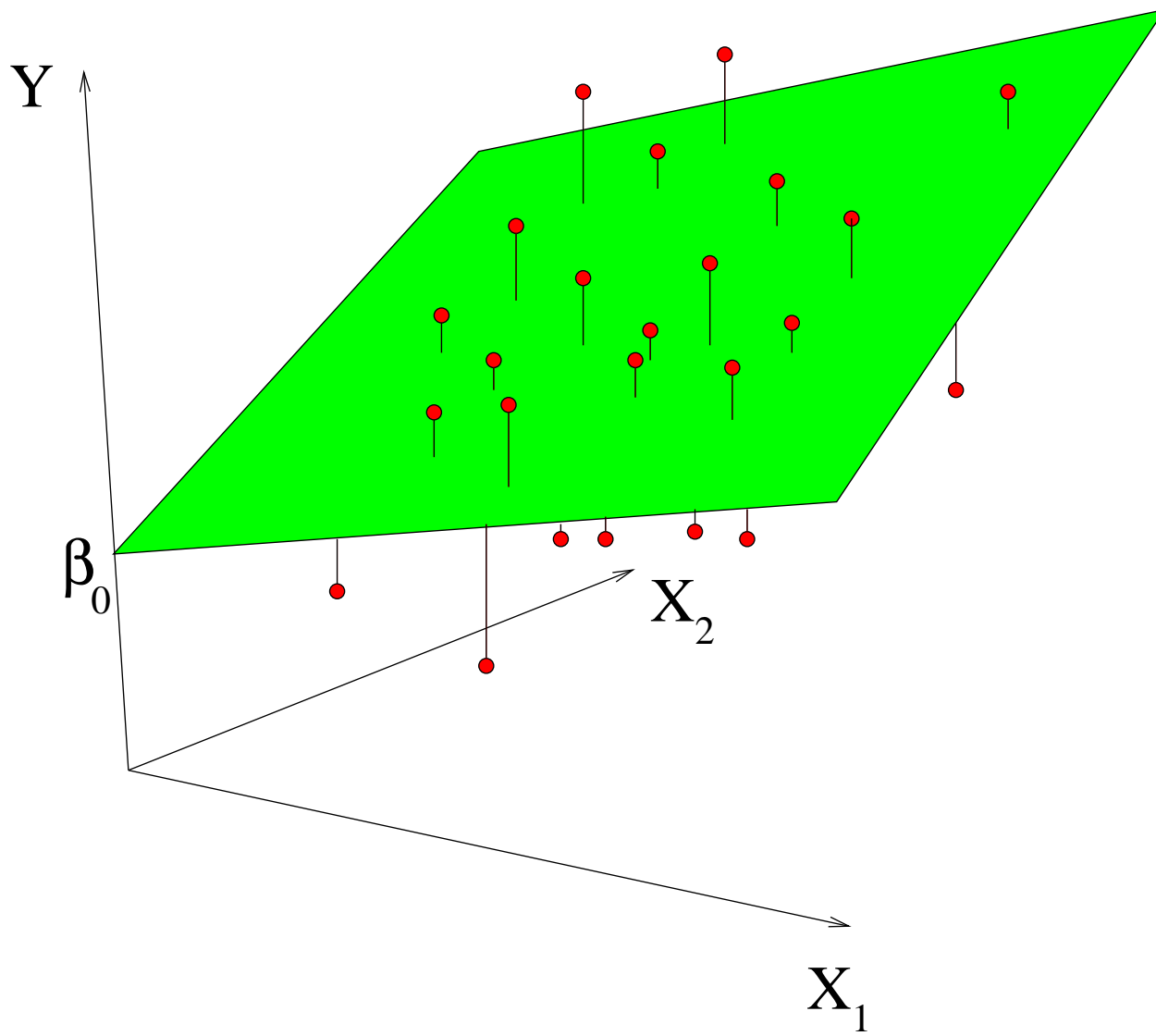
Es sind also $\beta_0, \beta_1, \dots, \beta_p$ aus Trainingsdaten $(x_1, y_1), \dots, (x_N, y_n)$ für (X, Y) zu schätzen.

Matrixdarstellung:

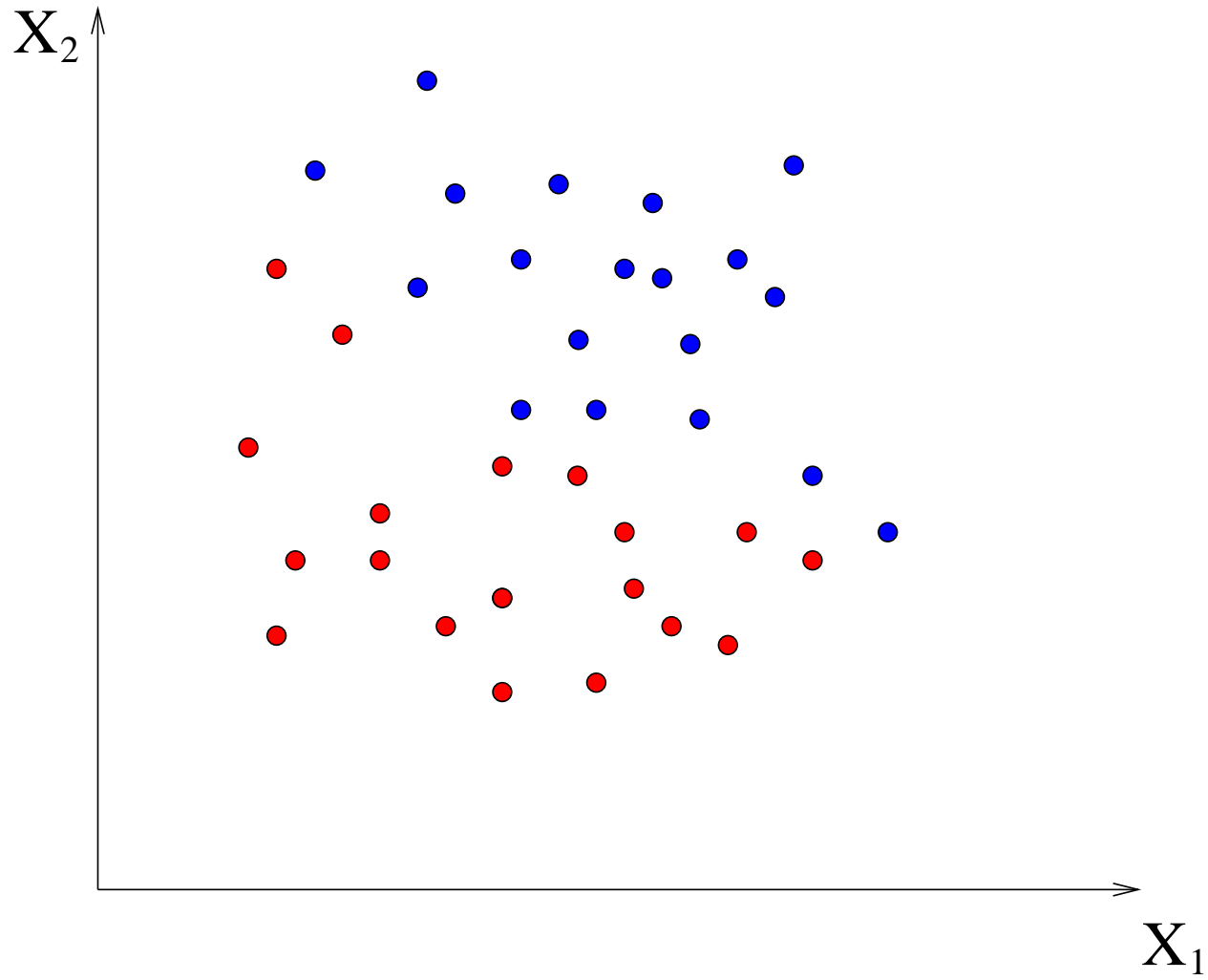
$$\mathbf{y} \approx f(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

mit

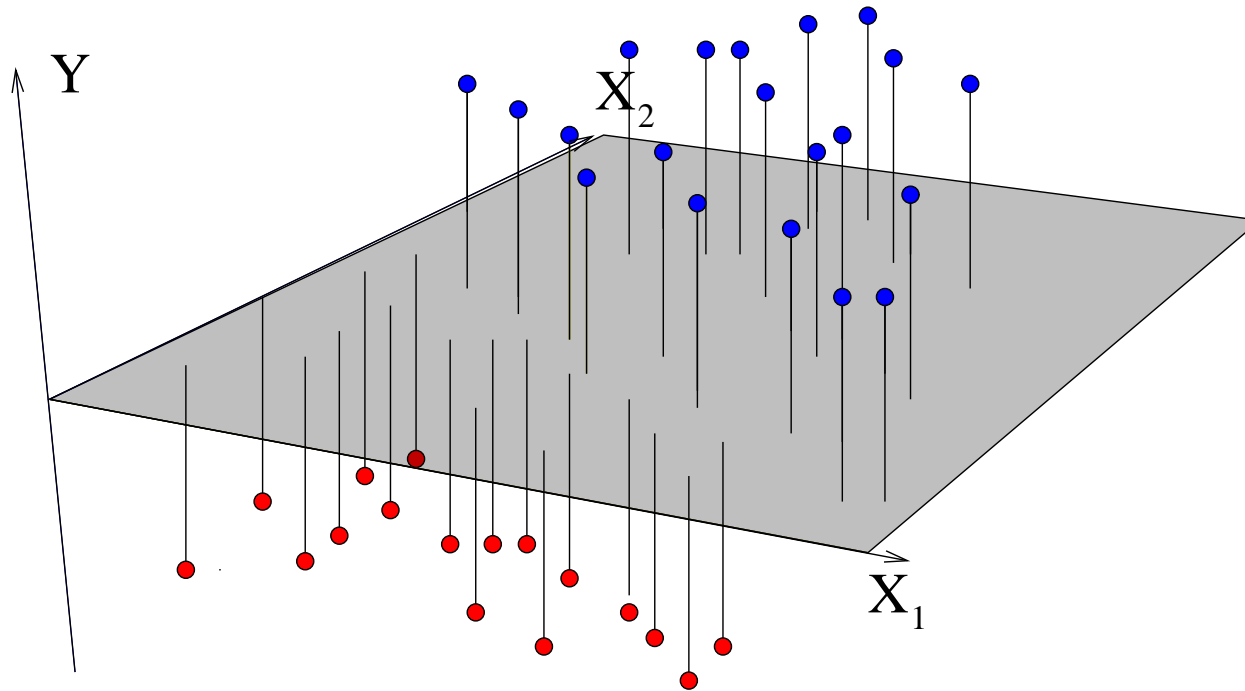
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix}$$



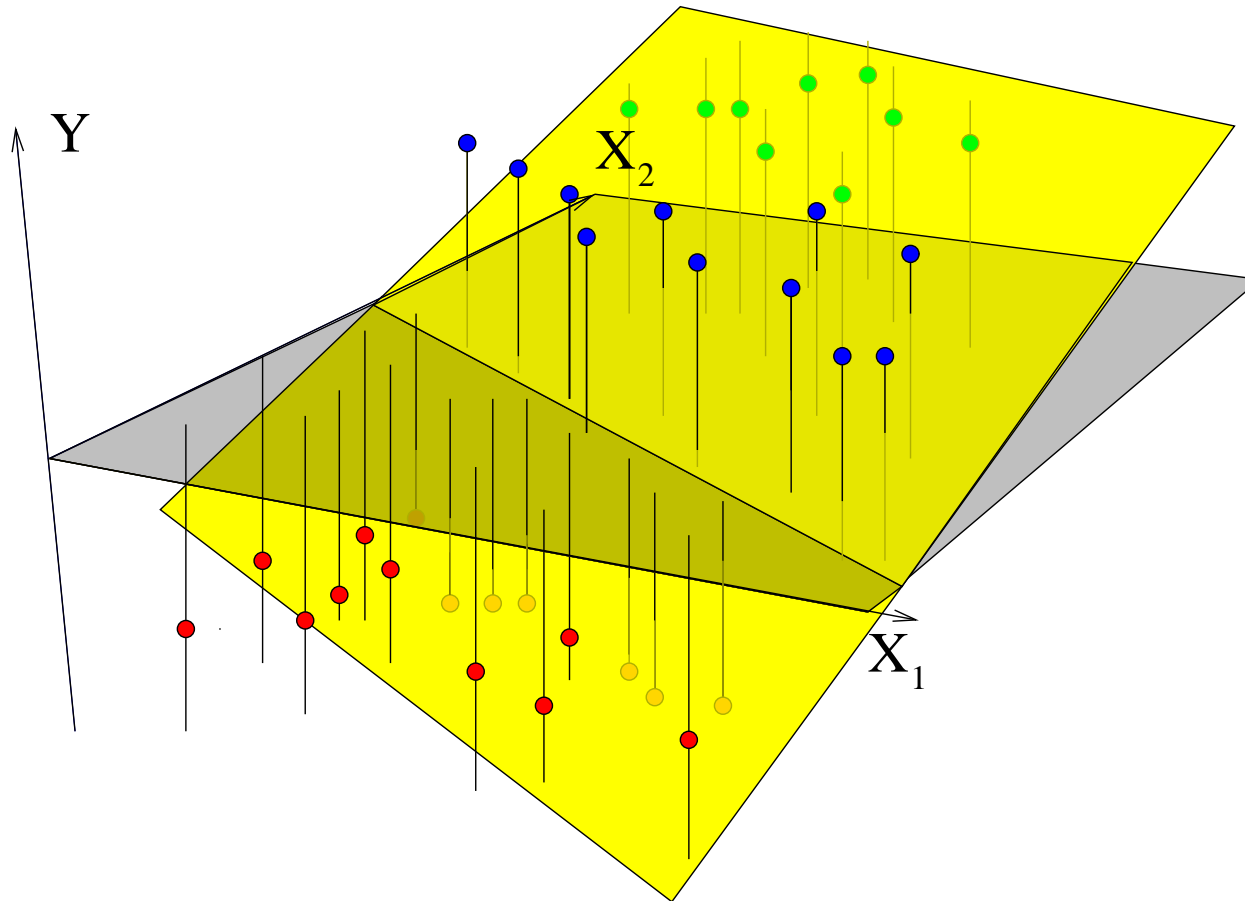
Diskriminanz mit linearer Regression



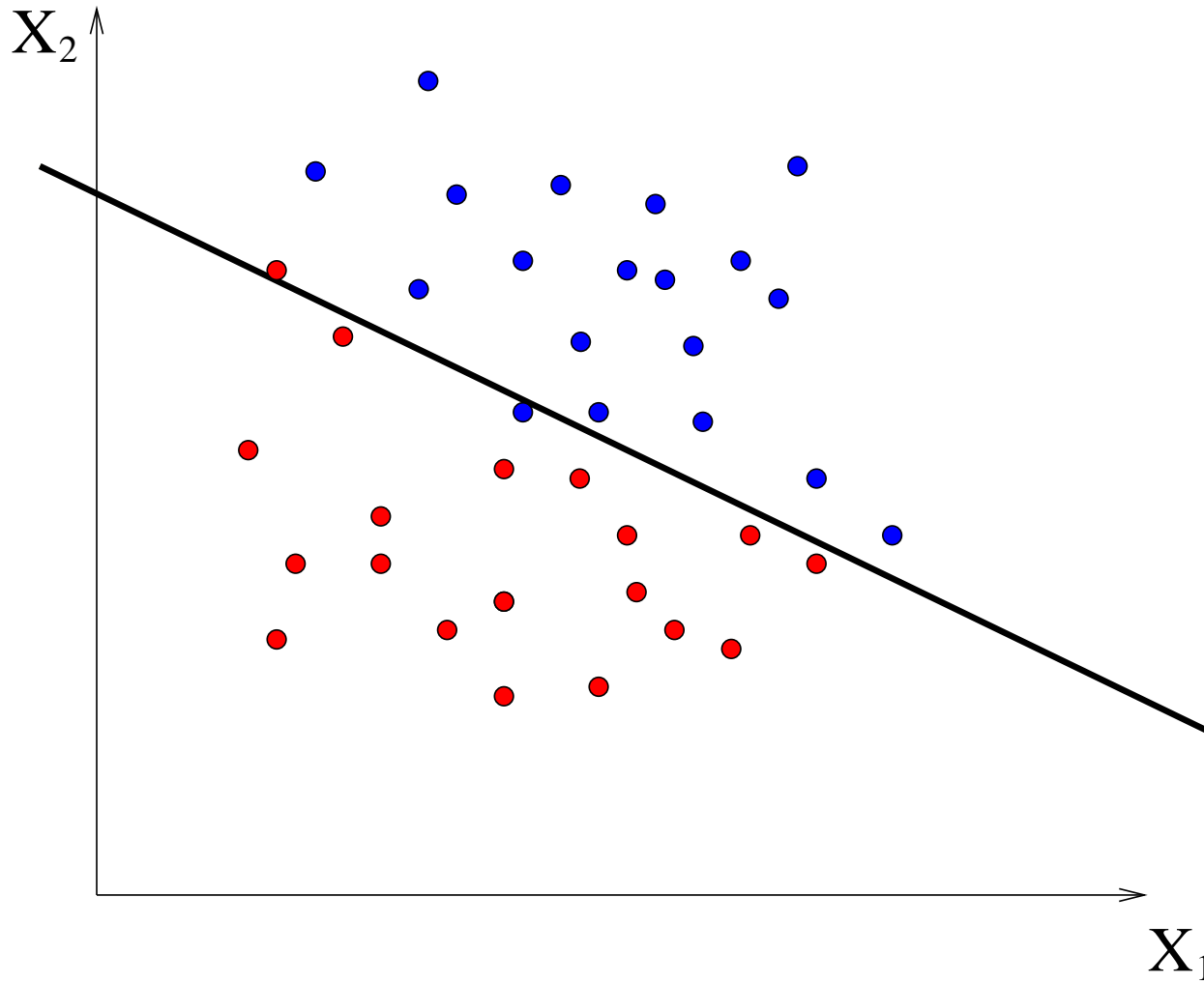
Diskriminanz mit linearer Regression



Diskriminanz mit linearer Regression



Diskriminanz mit linearer Regression



Farbe	X1	X2		Y	X1	X2
rot	0.23	0.41		-1	0.23	0.41
rot	0.13	0.72		-1	0.13	0.72
rot	0.52	0.35		-1	0.52	0.35
blau	0.16	0.46	\Rightarrow	1	0.16	0.46
blau	0.42	0.56		1	0.42	0.56
.
.
.

$$\mathbf{y} \approx \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 = \mathbf{X}\beta$$

Datensatz anorexia in der MASS-library in R mit $\text{Postwt} \sim \text{Prewt} * \text{Treat}$

	Treat	Prewt	Postwt
1	Cont	80.7	80.2
2	Cont	89.4	80.1
.	.	.	.
.	.	.	.
26	Cont	89.0	78.8
27	CBT	80.5	82.2
28	CBT	84.9	85.6
.	.	.	.
.	.	.	.
55	CBT	87.4	86.7
56	FT	83.8	95.2
57	FT	83.3	94.3
.	.	.	.
.	.	.	.

$$\implies \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 80.7 & 80.7 & 0 \\ 1 & 1 & 0 & 89.4 & 89.4 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & 89.0 & 89.0 & 0 \\ 1 & 0 & 0 & 80.5 & 0 & 0 \\ 1 & 0 & 0 & 84.9 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 0 & 87.4 & 0 & 0 \\ 1 & 0 & 1 & 83.8 & 0 & 83.8 \\ 1 & 0 & 1 & 83.3 & 0 & 83.3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 80.2 \\ 80.1 \\ \cdot \\ \cdot \\ 78.8 \\ 82.2 \\ 85.6 \\ \cdot \\ \cdot \\ 86.7 \\ 95.2 \\ 94.3 \\ \cdot \\ \cdot \end{pmatrix}$$

Ein Ansatz zum Schätzen von β_i :

least squares (kleinste Quadrate)

minimiere die Summe der Quadrate der Residuen (*residual sum of squares*):

$$RSS(\beta) = \sum_{i=1}^N (y_i - f(x_i))^2 = \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

Sei dazu

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix}$$

und somit $RSS(\beta) = \langle \mathbf{y} - \mathbf{X}\beta, \mathbf{y} - \mathbf{X}\beta \rangle = \|\mathbf{y} - \mathbf{X}\beta\|^2$.

Gesucht ist also $\hat{\beta}$, so dass $\hat{y} = \mathbf{X}\hat{\beta}$ minimalen euklidischen Abstand zu y hat.

Wir minimieren nun RSS , indem wir die Nullstelle der Ableitung, also des Gradienten

$$\frac{\partial RSS(\beta)}{\partial \beta} := \left(\frac{\partial RSS(\beta)}{\partial \beta_0}, \dots, \frac{\partial RSS(\beta)}{\partial \beta_p} \right) = -2(\mathbf{y} - \mathbf{X}\beta)^T \mathbf{X}$$

suchen (T steht für 'transponiert').

Wenn wir annehmen dürfen, dass $\mathbf{X}^T \mathbf{X}$ invertierbar ist, dann hat $(\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{X} = (0, \dots, 0)$ die eindeutige Lösung

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

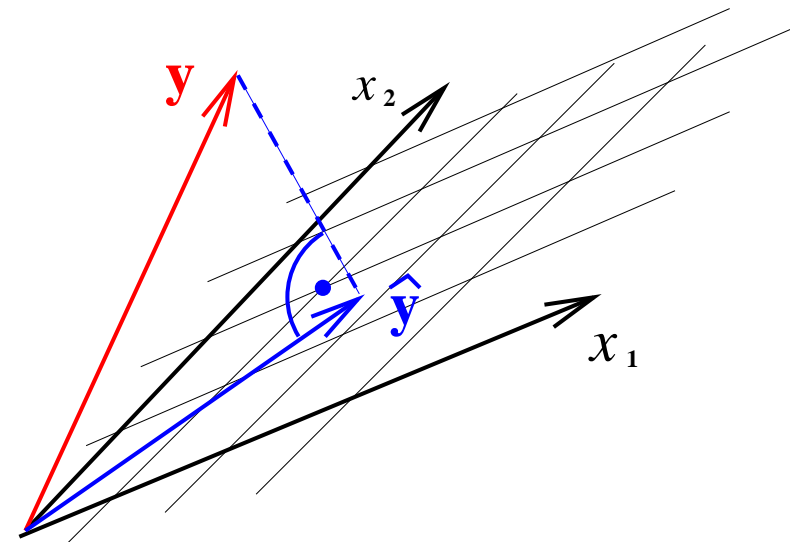
Geometrischer Lösungsweg: $RSS(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$ minimieren bedeutet, dass $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ die Projektion von \mathbf{y} auf den von den Vektoren x_0, x_1, \dots, x_N aufgespannten Raum sein soll (mit $x_0 = (1, \dots, 1)^T$). Also muss $\mathbf{y} - \mathbf{X}\hat{\beta}$ auf jedem x_i senkrecht stehen,

d.h.

$$\forall i : \langle \mathbf{y} - \mathbf{X}\hat{\beta}, x_i \rangle = 0,$$

d.h.

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{X} = (0, \dots, 0)$$



Wann und in welchem Sinne ist der kleinste-Quadrate-Schätzer optimal?

Präzisierung des Modells:

$$Y = \beta_0 + \sum_{j=1}^p X_j \beta_j + \varepsilon$$

Dabei ist $\mathbb{E}(\varepsilon) = 0$ und $\text{Var}(\varepsilon) = \sigma^2$ und damit $\mathbb{E}(Y) = \beta_0 + \sum_{j=1}^p X_j \beta_j$ und $\text{Var}(Y) = \sigma^2$.

Klassischerweise ist Normalverteilung vorausgesetzt: $\varepsilon \sim N(0, \sigma^2)$ und damit $Y \sim N(\beta_0 + \sum_{j=1}^p X_j \beta_j, \sigma^2)$.

Dies gilt für alle Komponenten von \mathbf{y} , die außerdem als stochastisch unabhängig vorausgesetzt sind.

Die Kovarianzmatrix $\text{Var}(V)$ eines Zufallsvektors $V = (V_1, \dots, V_n)$ enthält als Einträge $\text{Cov}(V_i, V_j) = \mathbb{E}((V_i - \mathbb{E}V_i)(V_j - \mathbb{E}V_j)) = \mathbb{E}(V_i V_j) - \mathbb{E}V_i \mathbb{E}V_j$. Z.B.:

$$\text{Var}(\mathbf{y}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{pmatrix} = \sigma^2 I$$

Ist Y ein Zufallsvektor, A eine Matrix, v ein Vektor und $Z = v + AY$, so gilt

$$\mathbb{E}(Z) = v + A \cdot \mathbb{E}(Y)$$

$$\text{Var}(Z) = A\text{Var}(Y)A^T$$

Ist Y (multivariat) normalverteilt, so ist auch Z normalverteilt. Eine multivariate Normalverteilung ist durch \mathbb{E} und Var eindeutig bestimmt.

(siehe z.B. Rice: "Mathematical Statistics and Data Analysis", Duxbury Press, 1995)

⇒

$$\forall \beta \in \mathbb{R}^{p+1} : \mathbb{E}(\hat{\beta}) = \mathbb{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta$$

$\hat{\beta}$ ist also ein **erwartungstreuer** (*unbiased*) Schätzer für β .

Kovarianzmatrix von $\hat{\beta}$:

$$\begin{aligned} \text{Var}(\hat{\beta}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \cdot \text{Var}(\mathbf{y}) \cdot ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \end{aligned}$$

Für $\varepsilon \sim N(0, \sigma^2)$ ist $\hat{\beta}$ multivariat normalverteilt gemäß $N(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$.

Satz 7 (Gauß-Markov-Theorem) Ist $\tilde{\beta}$ ein erwartungstreuer linearer Schätzer für β so gilt $v \text{Var}(\hat{\beta})_{ij} v^T \leq v \text{Var}(\tilde{\beta})_{ij} v^T$ für jeden Zeilenvektor $v = (v_0, \dots, v_p)$, d.h. $\text{Var}(v\hat{\beta}) \leq \text{Var}(v\beta)$.

Beweis Sei $\tilde{\beta} = C\mathbf{y}$. Aus $\forall \beta \in \mathbb{R}^{p+1} : \beta = \mathbb{E}\tilde{\beta} = C\mathbb{E}\mathbf{y} = C\mathbf{X}\beta$ folgt $C\mathbf{X} = I$.

Sei $D := C - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Dann folgt wegen

$$D\mathbf{X} = C\mathbf{X} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = C\mathbf{X} - I = \mathbf{0}:$$

$$\begin{aligned} \text{Var}(\tilde{\beta}) &= C\text{Var}(\mathbf{y})C^T = \sigma^2 CC^T \\ &= \sigma^2 \left(DD^T + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T D^T + D\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \right) \\ &= \sigma^2 DD^T + \text{Var}(\hat{\beta}) \end{aligned}$$

Die Aussage folgt, da DD^T positiv-semidefinit sind.

□

3.2 LDA: Lineare Diskriminanz-Analyse

Angenommen es sind Klassen zu trennen, wobei die Punkte der k -ten Klasse aus einer p -dimensionalen Normalverteilung um μ_k mit Kovarianzmatrix C_k stammen, d.h. die Dichte ist

$$f_k(x) = \frac{e^{-\frac{1}{2}(x-\mu_k)^T C_k^{-1}(x-\mu_k)}}{\sqrt{(2\pi)^p \cdot \det C_k}}$$

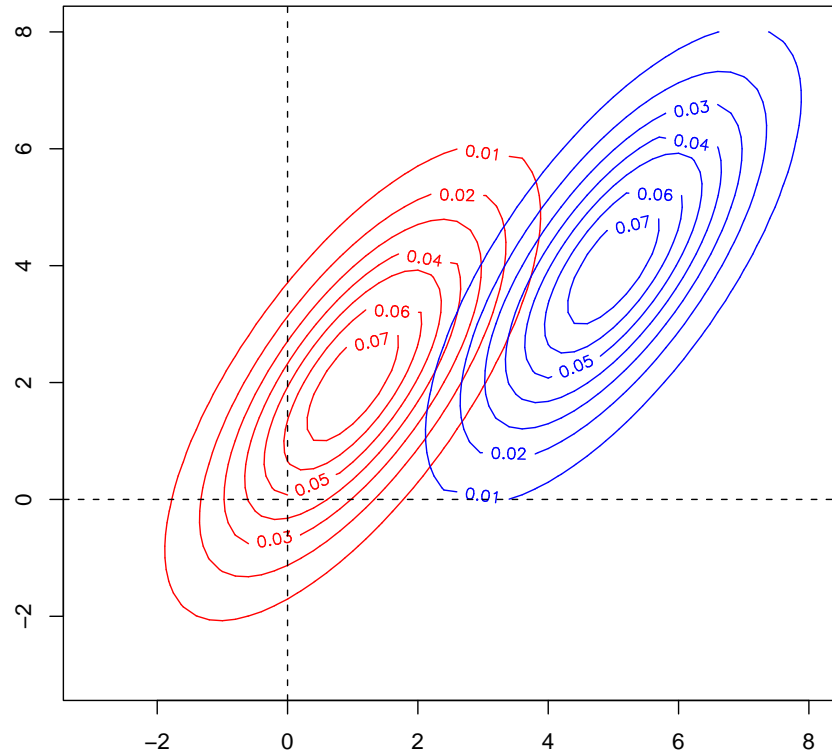
Wir gehen im folgenden zunächst von $C_k = C$ aus.

Beispiel:

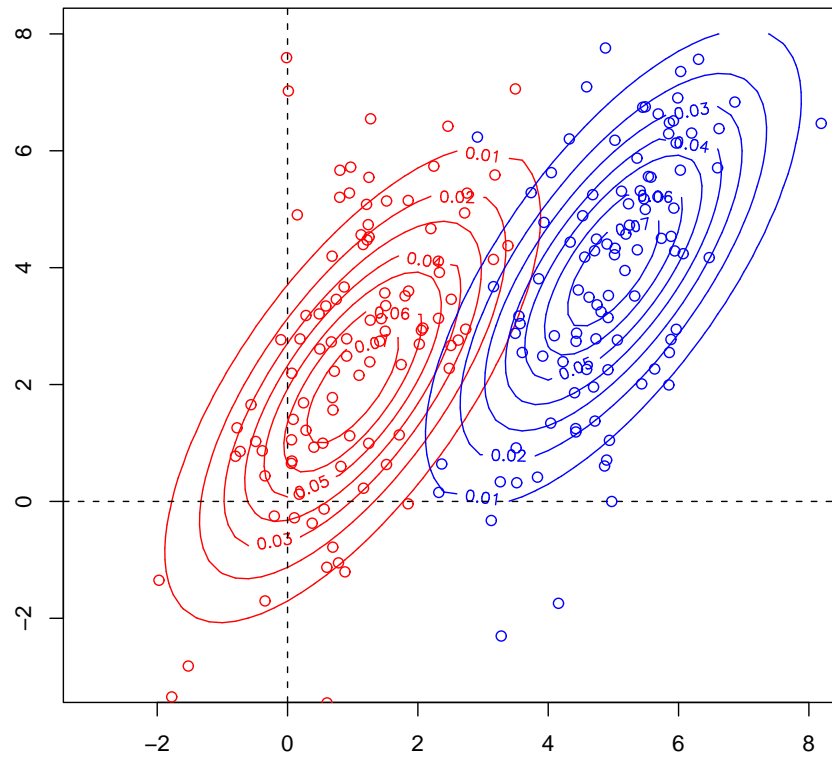
$$\mu_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}^T = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$$

$$\Rightarrow C^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$



$$f_k(x) = \frac{e^{-\frac{1}{2}(x-\mu_k)^T C_k^{-1}(x-\mu_k)}}{\sqrt{(2\pi)^p \cdot \det C_k}}$$



$$\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \end{pmatrix}$$

mit Z_i unabhängig Standard-
Normalverteilt

Seien K Klassen $1, \dots, K$ zu trennen. Die Klasse k habe Häufigkeit π_k .

Bayes-Ansatz: Ist G die Klassenzugehörigkeit eines Punktes mit Position X , so gilt

$$\text{Ws}(G = k \mid X = x) = \frac{f_k(x) \cdot \pi_k}{\sum_{\ell=1}^K f_{\ell}(x) \pi_{\ell}}$$

mit $f_k(x)$: Dichte der Klasse k an der Stelle x .

Mit unseren Annahmen betrachten wir zur Trennung der Klassen 1 und 2:

$$\begin{aligned}\ln \frac{\text{Ws}(G = 1 \mid X = x)}{\text{Ws}(G = 2 \mid X = x)} &= \ln \frac{f_1(x)}{f_2(x)} + \ln \frac{\pi_1}{\pi_2} \\ &= \ln \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^T C^{-1}(\mu_1 - \mu_2) + x^T C^{-1}(\mu_1 - \mu_2)\end{aligned}$$

Das ist **linear** in x und genau dann positiv wenn

$$\text{Ws}(G = 1 \mid X = x) > \text{Ws}(G = 2 \mid X = x).$$

LDA-Entscheidungsregel:

Schätze

$$\hat{\pi}_k = \frac{N_k}{N}$$

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$$

$$\hat{C} = \sum_{k=1}^K \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$

und entscheide für Klasse 1 falls

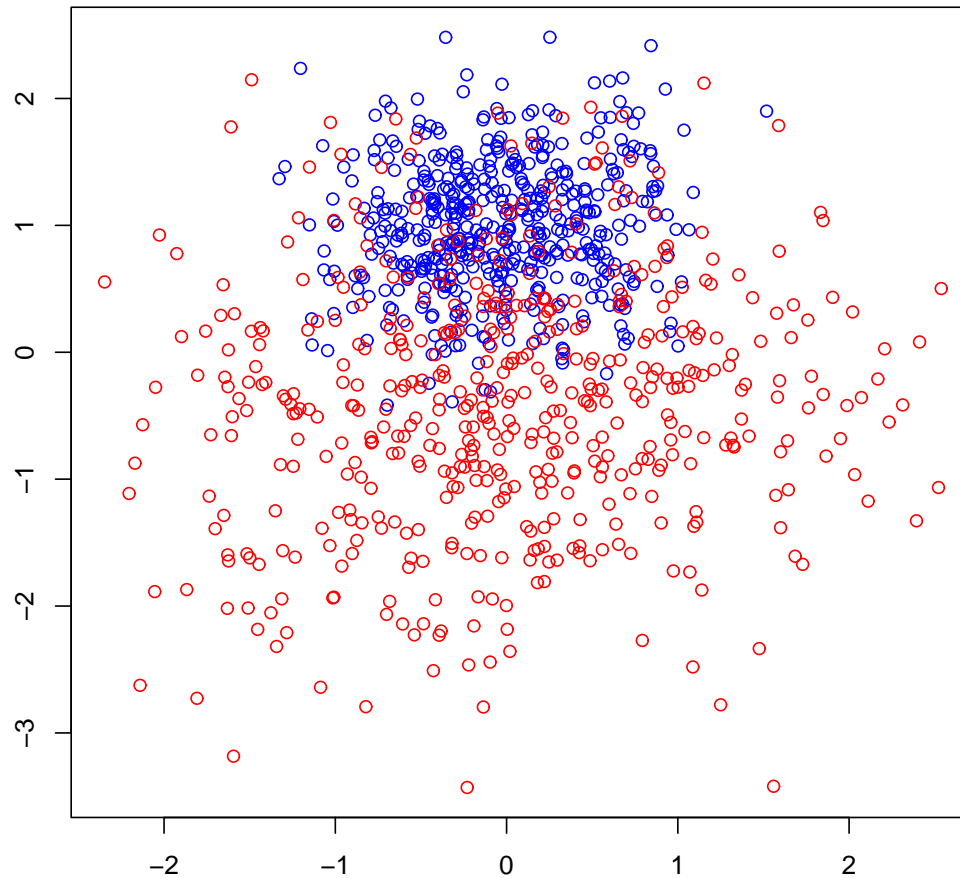
$$x^T C^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \leq \frac{1}{2} \hat{\mu}_2^T C^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^T C^{-1} \hat{\mu}_1 + \ln \frac{N_1}{N} - \ln \frac{N_2}{N}$$

Ist LDA und Diskriminanz mit Linearer Regression dasselbe?

Im Fall $N_1 = N_2$ ist es dasselbe, sonst liegen die trennenden Hyperebenen lediglich parallel, sind aber unterschiedlich weit vom Nullpunkt entfernt.

QDA: quadratische Diskriminanz-Analyse

Wenn die Kovarianzmatrizen C_k der einzelnen Gruppen verschieden sind, ist lineare Trennung nicht optimal:



Der selbe Ansatz wie bei LDA liefert für ungleiche C_k folgendes Ergebnis: Setze

$$\delta_k(x) = -\frac{1}{2} \ln(\det C_k) - \frac{1}{2} (x - \mu_k)^T C_k^{-1} (x - \mu_k) + \ln \pi_k$$

Die optimale Trennfläche zwischen den Klassen k und ℓ ist dann $\{x \mid \delta_k(x) = \delta_\ell(x)\}$.

Alternative zu QDA:

Wende LDA nicht auf (X_1, \dots, X_p) an, sondern auf $(X_1, \dots, X_p, X_1 \cdot X_1, X_1 \cdot X_2, \dots, X_p \cdot X_p)$.

Das führt zu ähnlichen Ergebnissen wie QDA.

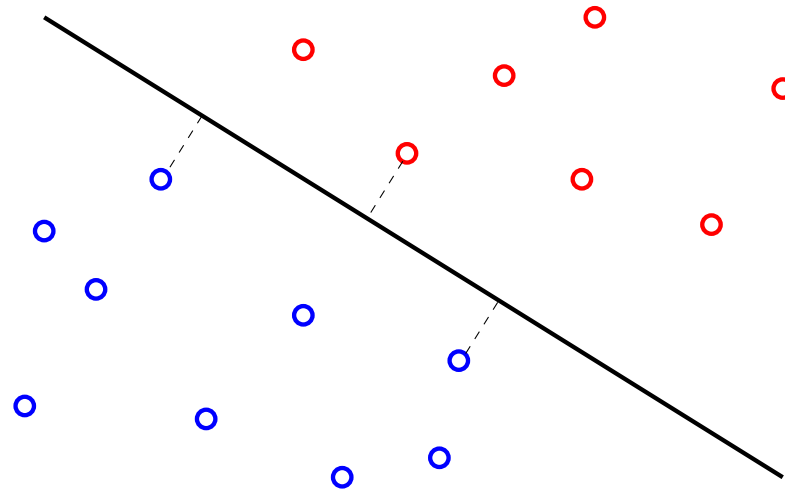
(Analog auch mit linearer Regression möglich.)

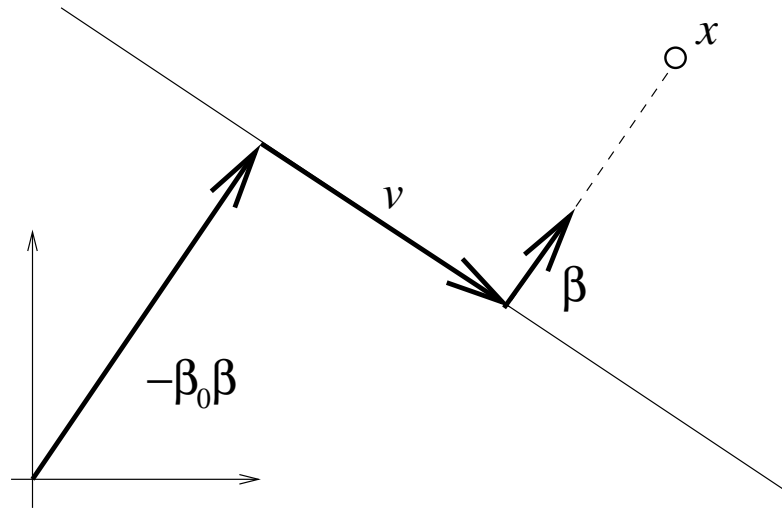
3.3 Lineare Support Vector Klassifikation

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.