Vorlesung Lernalgorithmen

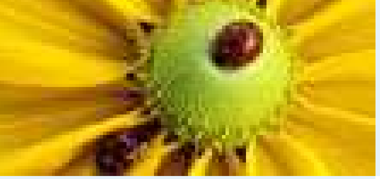
Support Vector Machines

Stefan Gulan

Lineare Separierbarkeit

- *Hyperebene* H im $\mathbb{R}^n$ definiert durch Normale $\mathbf{w}$ und Skalar b

$$\mathbf{w}\cdot(\mathbf{h} - \mathbf{x}) = \mathbf{w}\cdot\mathbf{h} + b = 0$$

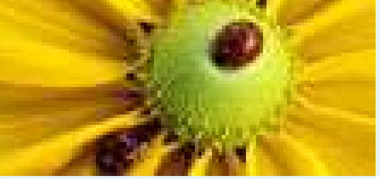


Lineare Separierbarkeit

- *Hyperebene* H im $\mathbb{R}^n$ definiert durch Normale $\mathbf{w}$ und Skalar b

$$\mathbf{w}\cdot(\mathbf{h}-\mathbf{x}) = \mathbf{w}\cdot\mathbf{h} + b = 0$$

- $X, Y \subset \mathbb{R}^n$ heissen *linear separierbar*, wenn sie in verschiedenen Halbräumen von H liegen:



Lineare Separierbarkeit

- *Hyperebene* H im $\mathbb{R}^n$ definiert durch Normale $\mathbf{w}$ und Skalar b

$$\mathbf{w}\cdot(\mathbf{h}-\mathbf{x}) = \mathbf{w}\cdot\mathbf{h} + b = 0$$

- $X, Y \subset \mathbb{R}^n$ heissen *linear separierbar*, wenn sie in verschiedenen Halbräumen von H liegen:

$$\forall x \in X : \mathbf{x}\cdot\mathbf{w} + b < 0, \quad \forall y \in Y : \mathbf{y}\cdot\mathbf{w} + b > 0$$

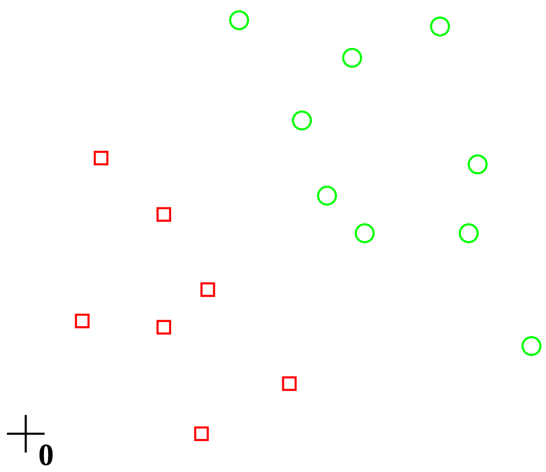
Lineare Separierbarkeit

- *Hyperebene* H im $\mathbb{R}^n$ definiert durch Normale $\mathbf{w}$ und Skalar b

$$\mathbf{w}\cdot(\mathbf{h}-\mathbf{x}) = \mathbf{w}\cdot\mathbf{h} + b = 0$$

- $X, Y \subset \mathbb{R}^n$ heissen *linear separierbar*, wenn sie in verschiedenen Halbräumen von H liegen:

$$\forall x \in X : \mathbf{x}\cdot\mathbf{w} + b < 0, \quad \forall y \in Y : \mathbf{y}\cdot\mathbf{w} + b > 0$$



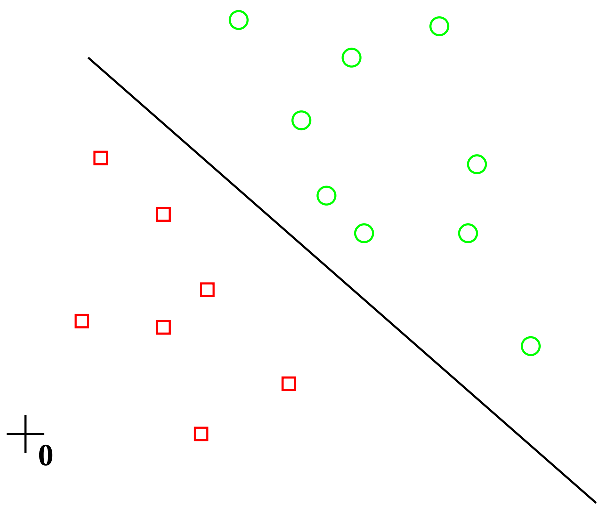
Lineare Separierbarkeit

- *Hyperebene* H im $\mathbb{R}^n$ definiert durch Normale $\mathbf{w}$ und Skalar b

$$\mathbf{w}\cdot(\mathbf{h}-\mathbf{x}) = \mathbf{w}\cdot\mathbf{h} + b = 0$$

- $X, Y \subset \mathbb{R}^n$ heissen *linear separierbar*, wenn sie in verschiedenen Halbräumen von H liegen:

$$\forall x \in X : \mathbf{x}\cdot\mathbf{w} + b < 0, \quad \forall y \in Y : \mathbf{y}\cdot\mathbf{w} + b > 0$$

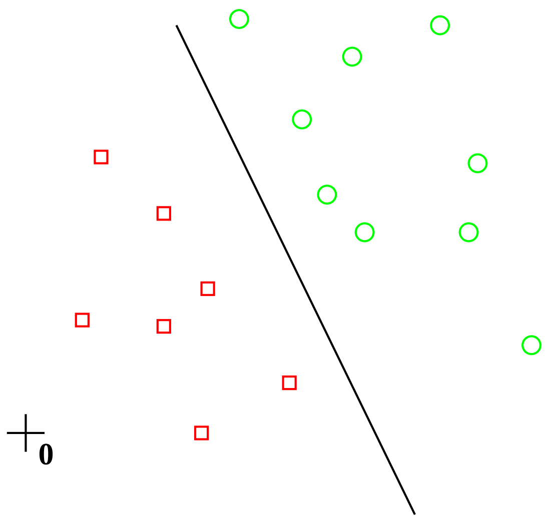


Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i

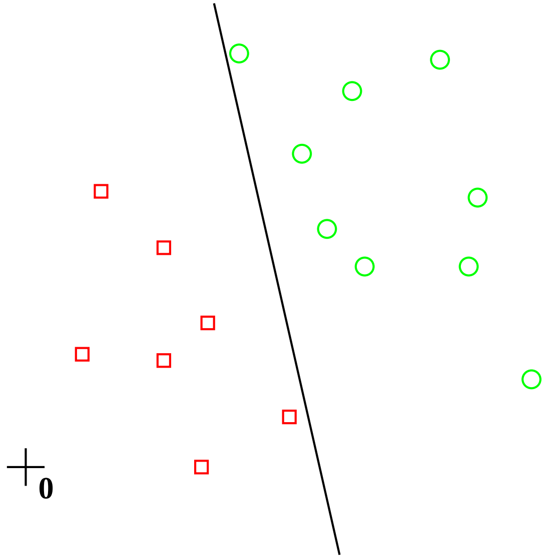
Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i



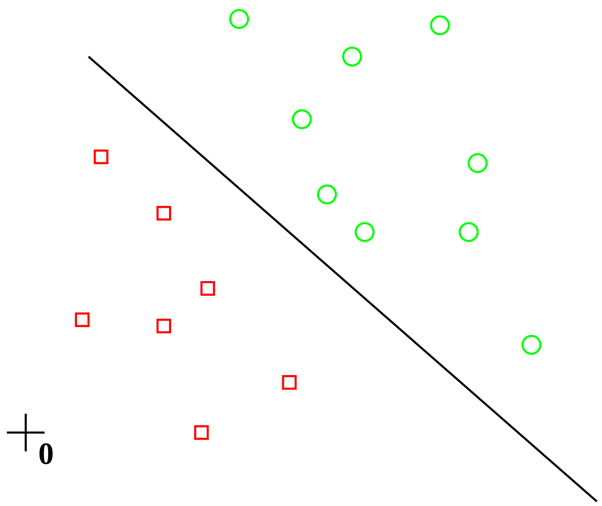
Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i



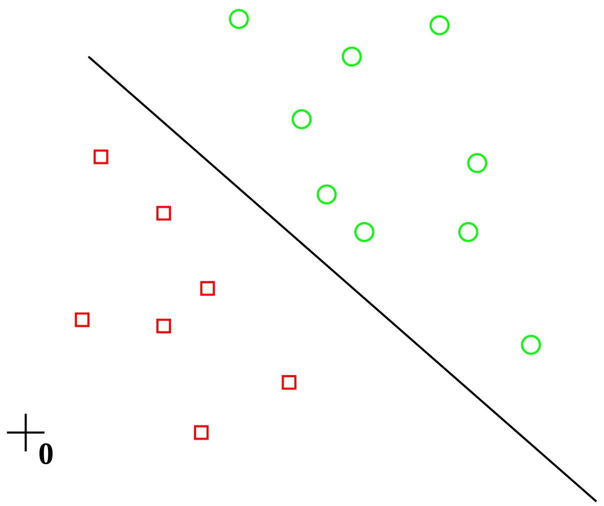
Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i



Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i

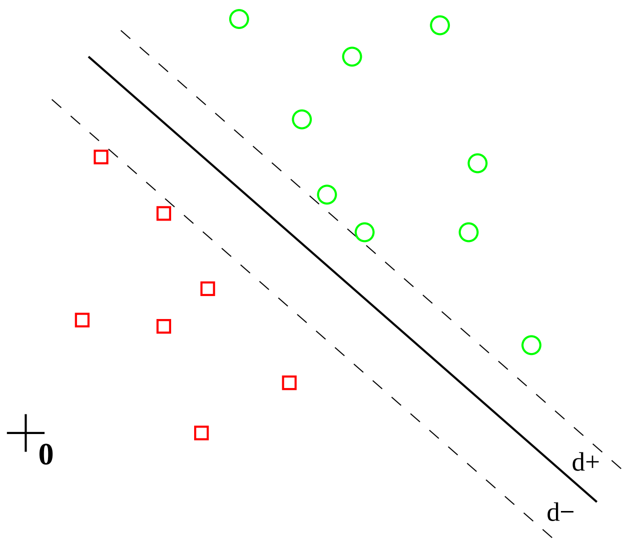


- Abstand v. H zu X, Y heisst *Randweite* (engl. *margin*)

$$d = d_+ + d_-$$

Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i

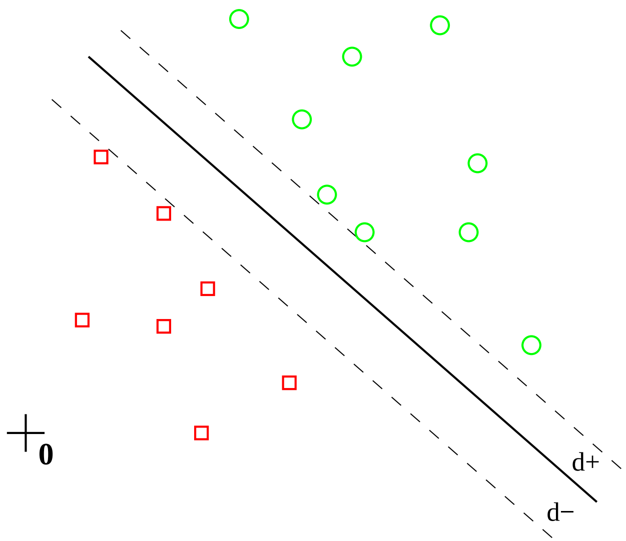


- Abstand v. H zu X, Y heisst *Randweite* (engl. *margin*)

$$d = d_+ + d_-$$

Lineare Separierbarkeit

- Lineare Separierbarkeit von X, Y vermöge beliebig vieler H_i



- Abstand v. H zu X, Y heisst *Randweite* (engl. *margin*)

$$d = d_+ + d_-$$

- Trennung intuitiv besser, je grösser d ist

Formalisierung

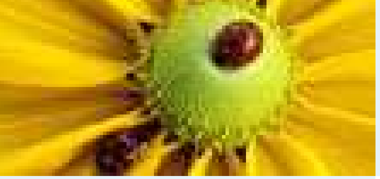
- Gegeben: Beispiele $(s_i, l_i) \in (X \times \{+1\}) \cup (Y \times \{-1\})$
  Gesucht: trennende Hyperebene $(\mathbf{w}, b)$ mit maximalem Rand

Formalisierung

- Gegeben: Beispiele $(s_i, l_i) \in (X \times \{+1\}) \cup (Y \times \{-1\})$
  Gesucht: trennende Hyperebene $(\mathbf{w}, b)$ mit maximalem Rand
- Skalierung ermöglicht

$$\mathbf{s}_i \cdot \mathbf{w} + b \geq +1 \quad \text{falls} \quad l_i = 1$$
$$\mathbf{s}_i \cdot \mathbf{w} + b \leq -1 \quad \text{falls} \quad l_i = -1$$

zusammengefasst: $l_i(\mathbf{s}_i \cdot \mathbf{w} + b) \geq 1, \forall i$

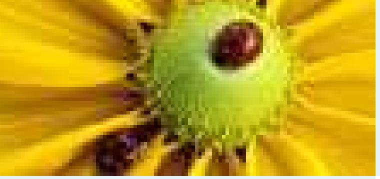


Formalisierung

- Gegeben: Beispiele $(s_i, l_i) \in (X \times \{+1\}) \cup (Y \times \{-1\})$
  Gesucht: trennende Hyperebene $(\mathbf{w}, b)$ mit maximalem Rand
- Skalierung ermöglicht

$$\mathbf{s}_i \cdot \mathbf{w} + b \geq +1 \quad \text{falls} \quad l_i = 1$$
$$\mathbf{s}_i \cdot \mathbf{w} + b \leq -1 \quad \text{falls} \quad l_i = -1$$

  zusammengefasst: $l_i(\mathbf{s}_i \cdot \mathbf{w} + b) \geq 1, \forall i$
- Abstand von H_+, H_- zu H: $d_+ = d_- = \frac{1}{\|\mathbf{w}\|}$
  $\Rightarrow$ Minimierung von $\|\mathbf{w}\|$ maximiert Randweite
  $\Rightarrow$ äquivalent: minimiere $\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w} \cdot \mathbf{w}$



Formalisierung

- Gegeben: Beispiele $(s_i, l_i) \in (X \times \{+1\}) \cup (Y \times \{-1\})$
  Gesucht: trennende Hyperebene $(\mathbf{w}, b)$ mit maximalem Rand
- Skalierung ermöglicht

$$\begin{aligned} \mathbf{s}_i \cdot \mathbf{w} + b &\geq +1 \quad \text{falls} \quad l_i = 1 \\ \mathbf{s}_i \cdot \mathbf{w} + b &\leq -1 \quad \text{falls} \quad l_i = -1 \end{aligned}$$

  zusammengefasst: $l_i(\mathbf{s}_i \cdot \mathbf{w} + b) \geq 1, \forall i$
- Abstand von H_+, H_- zu H: $d_+ = d_- = \frac{1}{\|\mathbf{w}\|}$
  $\Rightarrow$ Minimierung von $\|\mathbf{w}\|$ maximiert Randweite
  $\Rightarrow$ äquivalent: minimiere $\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w} \cdot \mathbf{w}$
- Quadratische Optimierung unter Nebenbenbedingungen

$$\begin{aligned} \text{minimiere} \quad & \frac{1}{2}\|\mathbf{w}\| \\ \text{unter NB} \quad & l_i(\mathbf{s}_i \cdot \mathbf{w} + b) \geq 1 \end{aligned}$$

Verallgemeinerung (1)

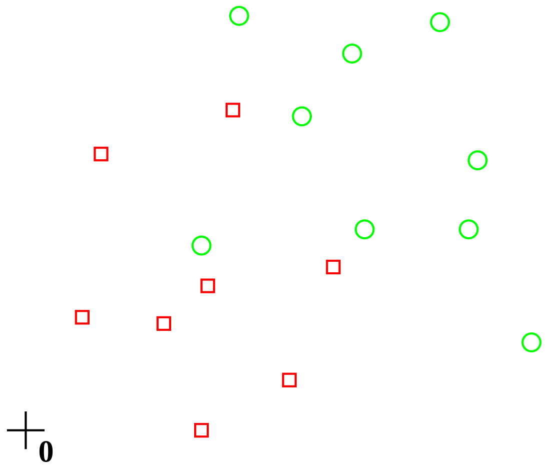
- Lineare Separierbarkeit eher unrealistische Annahme

Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern

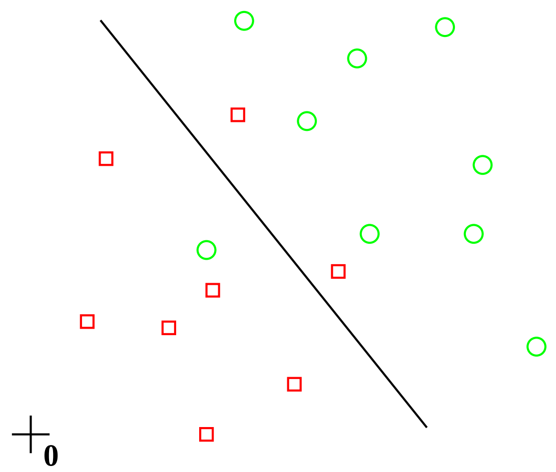
Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern



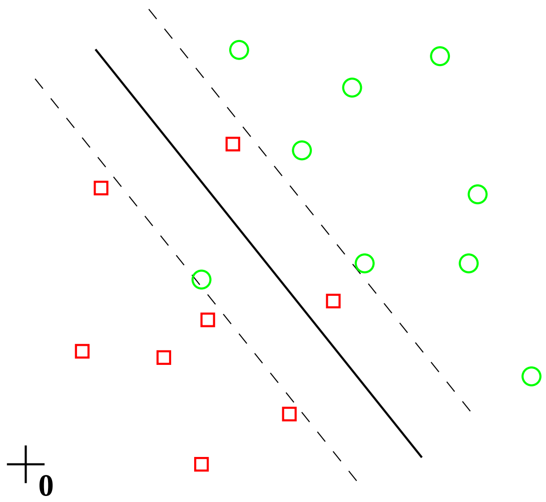
Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern



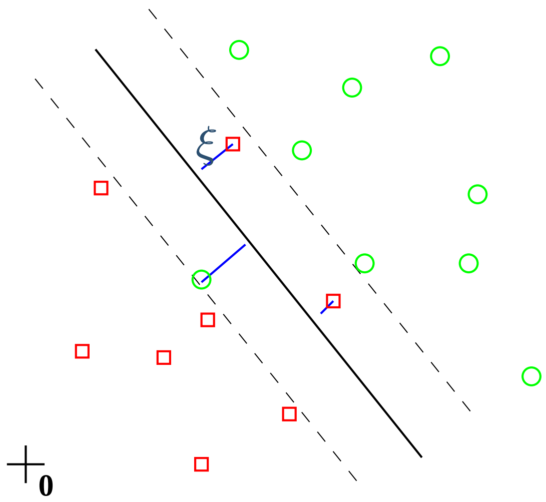
Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern



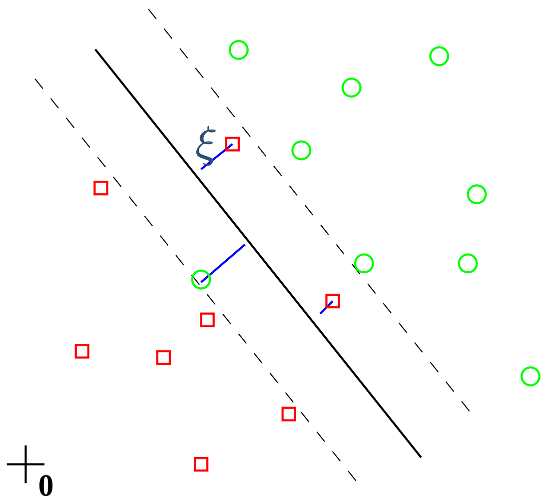
Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern



Verallgemeinerung (1)

- Lineare Separierbarkeit eher unrealistische Annahme
- Zulassen von Fehlern



- modifiziertes Problem

$$\text{minimiere} \quad \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_i \xi_i \qquad , C < \infty$$

$$\text{unter NB} \quad l_i(\mathbf{s}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i \qquad , l_i \in \{-1, +1\}$$

Lagrange-Dualität

- Minimierung unter Nebenbenbedingungen:

$$\begin{array}{rc} \text{minimiere} & f(\mathbf{x}) \\ \text{unter NB} & g_i(\mathbf{x}) \leq 0 \end{array}$$

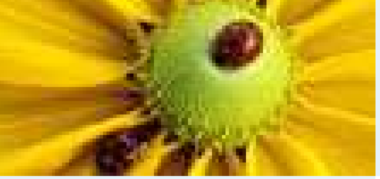
Lagrange-Dualität

- Minimierung unter Nebenbenbedingungen:

$$\begin{array}{rl} \text{minimiere} & f(\mathbf{x}) \\ \text{unter NB} & g_i(\mathbf{x}) \leq 0 \end{array}$$

ist äquivalent zu Maximierung des *Lagrange-Dualen*

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_i \lambda_i g_i(\mathbf{x})$$



Lagrange-Dualität

- Minimierung unter Nebenbenbedingungen:

$$\begin{array}{rl} \text{minimiere} & f(\mathbf{x}) \\ \text{unter NB} & g_i(\mathbf{x}) \leq 0 \end{array}$$

ist äquivalent zu Maximierung des *Lagrange-Dualen*

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_i \lambda_i g_i(\mathbf{x})$$

- hier ist
$f(\mathbf{x}, \xi) = \frac{1}{2}\mathbf{x} \cdot \mathbf{x} + C \sum_i \xi_i$, $g_i(\mathbf{x}, \xi_i) = l_i(\mathbf{s}_i \cdot \mathbf{x} + b) + \xi_i - 1$
also

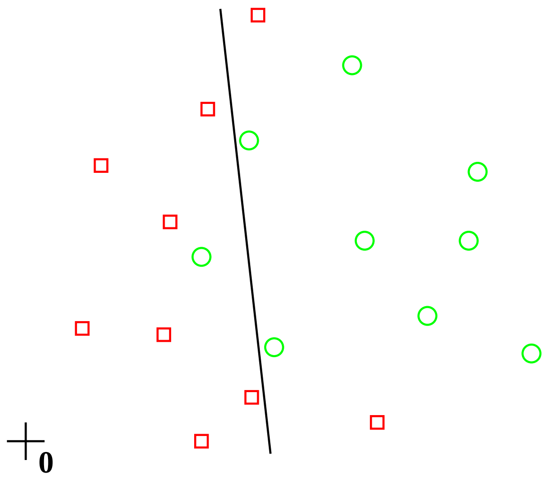
$$\begin{aligned} L(\mathbf{x}, \xi, \lambda) &= \frac{1}{2}\mathbf{x} \cdot \mathbf{x} + C \sum_i \xi_i \\ &\quad - \sum_i \lambda_i (l_i(\mathbf{s}_i \cdot \mathbf{x} + b) + \xi_i - 1) - \sum \beta_i \xi_i \end{aligned}$$

Verallgemeinerung (2)

- Linearität immer noch unrealistisch

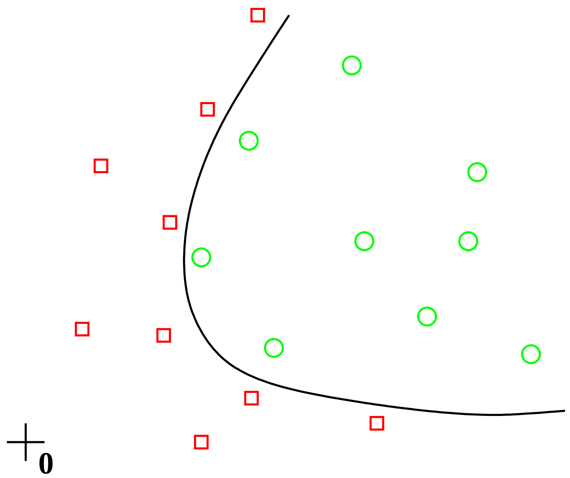
Verallgemeinerung (2)

- Linearität immer noch unrealistisch



Verallgemeinerung (2)

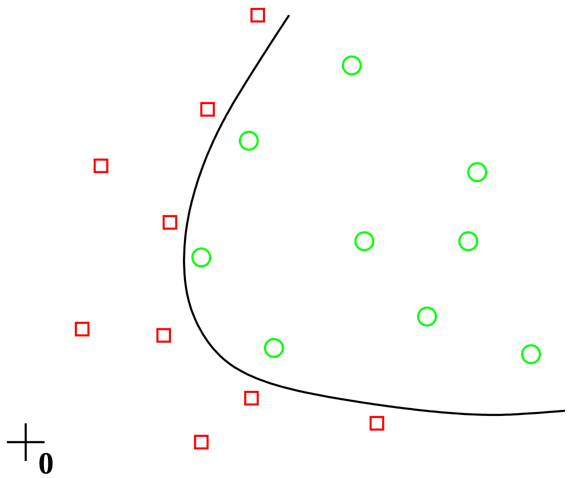
- Linearität immer noch unrealistisch



- Trennung durch nichtlineare Hyperkurve häufig natürlicher

Verallgemeinerung (2)

- Linearität immer noch unrealistisch

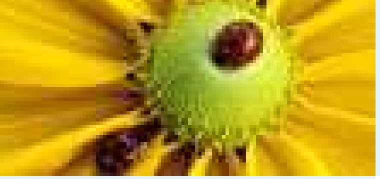


- Trennung durch nichtlineare Hyperkurve häufig natürlicher
- Idee: suche trennende Hyperebene in Raum höherer Dimension (*feature space*)
  zunächst $\Phi : \mathbb{R}^n \to \mathbb{R}^{n+k}$, dann lineare Lsg. in $\mathbb{R}^{n+k}$

Formalisierung

- Primaldarstellung:

$$\begin{array}{rl} \text{minimiere} & \frac{1}{2}\Phi(\mathbf{w})\cdot\Phi(\mathbf{w}) + C\sum_i \xi_i \\ \text{unter NB} & l_i(\Phi(s_i)\cdot\Phi(\mathbf{w})) \leq 1 - \xi_i \end{array}$$



Formalisierung

- Primaldarstellung:

$$\begin{array}{rl} \text{minimiere} & \frac{1}{2}\Phi(\mathbf{w}) \cdot \Phi(\mathbf{w}) + C\sum_i \xi_i \\ \text{unter NB} & l_i(\Phi(s_i) \cdot \Phi(\mathbf{w})) \leq 1 - \xi_i \end{array}$$

- Dualdarstellung

$$\begin{aligned} \text{maximiere} \quad L = \quad & \frac{1}{2}\Phi(\mathbf{w}) \cdot \Phi(\mathbf{w}) + C\sum_i \xi_i \\ & - \sum_i \lambda_i(l_i(\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{w}) + b) + \xi_i - 1) - \sum \beta_i \xi_i \end{aligned}$$

Formalisierung

- Primaldarstellung:

$$\begin{array}{rl} \text{minimiere} & \frac{1}{2}\Phi(\mathbf{w}) \cdot \Phi(\mathbf{w}) + C \sum_i \xi_i \\ \text{unter NB} & l_i(\Phi(s_i) \cdot \Phi(\mathbf{w})) \leq 1 - \xi_i \end{array}$$

- Dualdarstellung

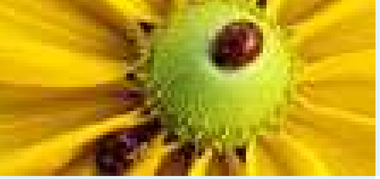
$$\begin{aligned} \text{maximiere} \quad L = \quad & \frac{1}{2}\Phi(\mathbf{w}) \cdot \Phi(\mathbf{w}) + C \sum_i \xi_i \\ & - \sum_i \lambda_i (l_i(\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{w}) + b) + \xi_i - 1) - \sum \beta_i \xi_i \end{aligned}$$

- Substituition partieller Ableitungen ($\frac{\partial L}{\partial[\cdot]} = 0$ f. Maximum) und KKT-Bedingungen

$$\text{maximiere} \;\; L_d(\lambda) = \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} l_i l_j \lambda_i \lambda_j \Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{s}_j)$$

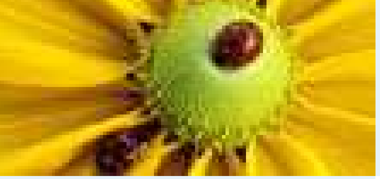
Kernelfunktionen

- Eingangsdaten treten in $\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{s}_j)$ auf



Kernelfunktionen

- Eingangsdaten treten in $\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{s}_j)$ auf
- Φ kann evtl. nach $\mathbb{R}^{\mathbb{N}}$ abbilden ... benötigt wird jedoch „nur“ Skalarprodukt von Bildern von Φ



Kernelfunktionen

- Eingangsdaten treten in $\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{s}_j)$ auf
- Φ kann evtl. nach $\mathbb{R}^{\mathbb{N}}$ abbilden ... benötigt wird jedoch „nur“ Skalarprodukt von Bildern von Φ
- Gewünscht ist *Kernelfunktion*

$$k : (X \cup Y)^2 \to \mathbb{N}, \quad (s_i, s_j) \mapsto \Phi(s_i) \cdot \Phi(s_j)$$

die explizite Berechnung von Φ umgeht

Kernelfunktionen

- Eingangsdaten treten in $\Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{s}_j)$ auf
- Φ kann evtl. nach $\mathbb{R}^{\mathbb{N}}$ abbilden ... benötigt wird jedoch „nur“ Skalarprodukt von Bildern von Φ
- Gewünscht ist *Kernelfunktion*

  $$k : (X \cup Y)^2 \to \mathbb{N}, \quad (s_i, s_j) \mapsto \Phi(s_i) \cdot \Phi(s_j)$$

  die explizite Berechnung von Φ umgeht
- Oft auch umgekehrte Problemstellung: Gegeben $f : M^2 \to \mathbb{N}$, ist f Kernelfunktion? — d.h. existiert Φ, so dass $f(a, b) = \Phi(a) \cdot \Phi(b)$?