Lernen ist induktiv!

überwacht

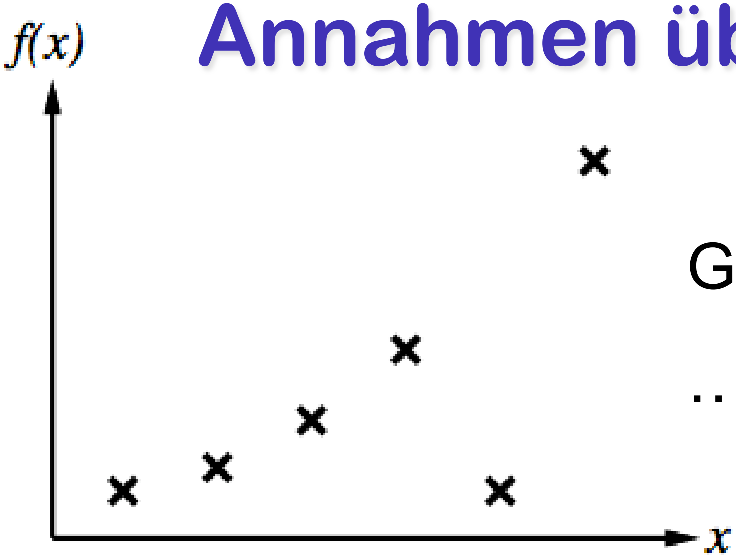
Aus Menge von **Lernbeispielen** $\left\langle \begin{array}{|c|c|c|} \hline O & O & X \\ \hline & X & \\ \hline & & \\ \hline \end{array}, (0,0) \right\rangle$

leite Fkt. h (Hypothese) ab, die Zielfunktion f approximiert!
↪ **Induktives** statt deduktivem Schließen

Vereinfachende Annahmen, Beispiele

- <u>Genau</u> diese Funktion ist zu lernen
- Lernbeispiele sind vorgegeben, fehlerfrei
- Kein Vorwissen ist zu berücksichtigen („*tabula rasa*“)
- Umgebung ist deterministisch und beobachtbar

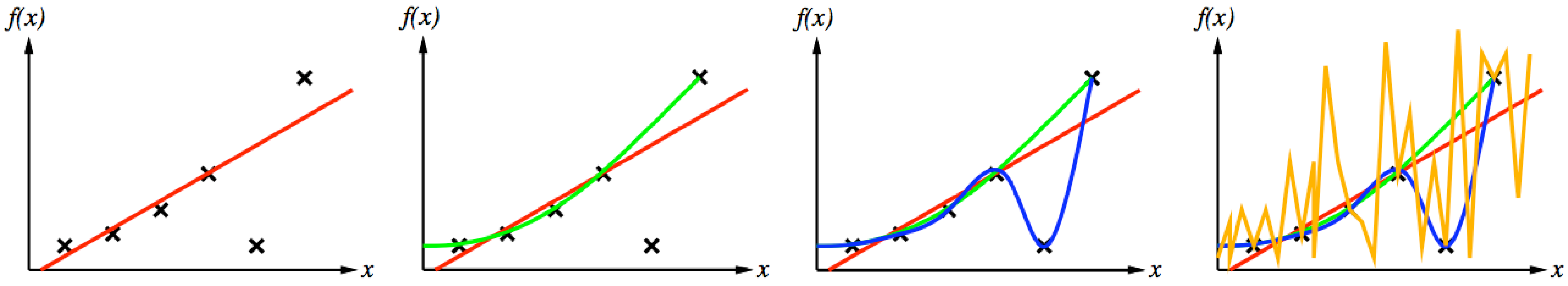
Annahmen über die Zielfunktion



Gegeben Lernbeispiele …

… ist die Hypothese …

… linear? … quadratisch? … kubisch? … ??? ?



Bekanntlich n Werte ausdrückbar durch Polynom $(n\text{-}1)$-ten Grades, doch führt das automatisch zur besten Zielfunktion?

Erinnerung: Definitionen aus der Statistik

Ausgangsmaterial:
N Datensätze, je n-dimensional

$$\mathbf{x}^1 = \langle x_1^1, x_2^1, \ldots, x_n^1 \rangle$$
$$\ldots$$
$$\mathbf{x}^N = \langle x_1^N, x_2^N, \ldots, x_n^N \rangle$$

Mittelwert der i-ten Dimension

$$\bar{x}_i = \frac{1}{N} \sum_{p=1}^{N} x_i^p$$

Standardabweichung

$$s_i = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \left(x_i^p - \bar{x}_i \right)^2}$$

Varianz

$$\sigma_i = s_i^2 = \frac{1}{N} \sum_{p=1}^{N} \left(x_i^p - \bar{x}_i \right)^2$$

Kovarianz

$$\sigma_{ij} = \frac{1}{N} \sum_{p=1}^{N} \left(x_i^p - \bar{x}_i \right) \left(x_j^p - \bar{x}_j \right)$$

Korrelationskoeffizient

$$K_{ij} = \frac{\sigma_{ij}}{s_i \cdot s_j}$$

Erinnerung: Abstandsmaße im $\mathcal{R}^n$

Für Punkte $\mathbf{x},\mathbf{y} \in \mathcal{R}^n$

Euklidischer Abstand $d(\mathbf{x},\mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Gewichteter Euklidischer Abstand

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$$

5.1 Überwachte Lernverfahren

Erinnerung Folie 228

- **Überwachtes Lernen**: Gegeben Paare $\langle \boldsymbol{In}, Erwartet\_Out \rangle$, leite Fkt. f ab, sodass für neue Eingaben gilt $f(\boldsymbol{In}')=Erwartet\_Out'$
- Ist Bildbereich von f endlich/diskret ➥ **Klassifikation**. Sonst **Regression** (kontinuierlich)

Hier behandelte Verfahren

- Perzeptron
- Nearest-Neighbor-Verfahren
- Entscheidungsbaum-Lernen
- Naive Bayes-Klassifikation

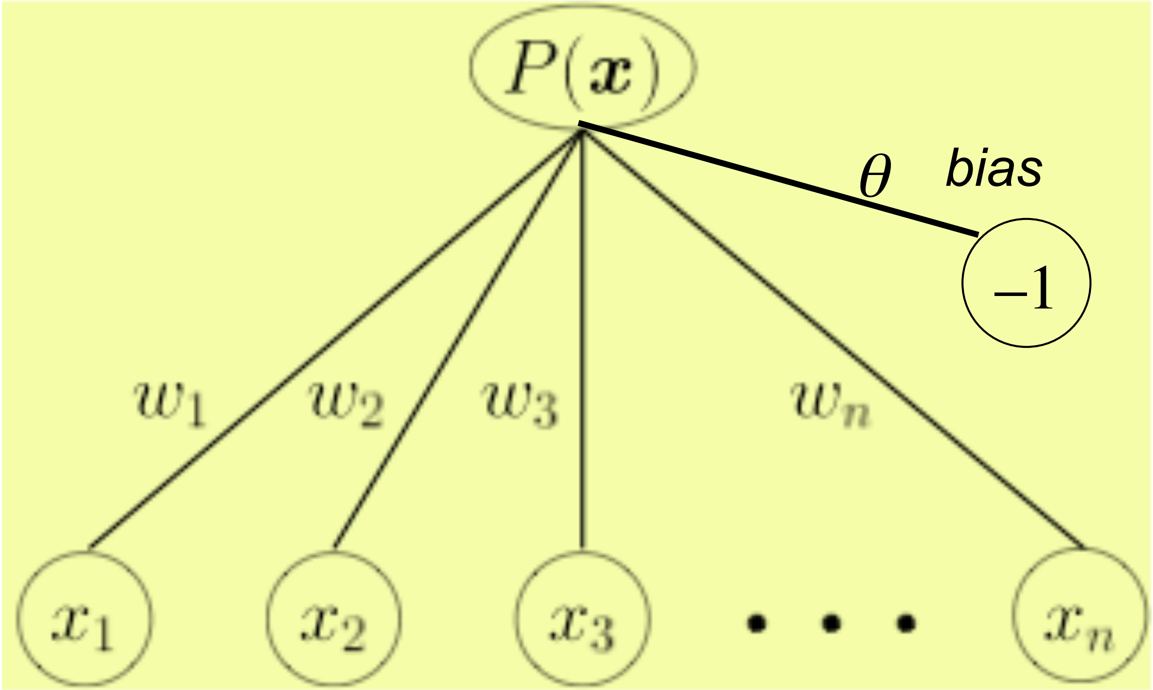
Perzeptron, Definition

Definition 8.8 *Sei* $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ *ein Gewichtsvektor und* $\boldsymbol{x} \in \mathbb{R}^n$ *ein Eingabevektor. Ein* **Perzeptron** *stellt eine Funktion* $P : \mathbb{R}^n \to \{0, 1\}$ *dar, die folgender Regel entspricht:*

$$P(\boldsymbol{x}) = \begin{cases} 1 \text{ falls} \quad \boldsymbol{w}\,\boldsymbol{x} = \sum_{i=1}^{n} w_i x_i > 0 \\ 0 \text{ sonst} \end{cases}$$

- Die Eingabevariablen x_i heißen **Merkmale** (*features*)
- genau die Punkte $\mathbf{x}$ oberhalb $(n\text{-}1)$-dimensionaler Hyperebene $\sum w_i x_i = 0$ werden positiv klassifiziert ($P(\mathbf{x})=1$)
- Meist füge künstlichen „*bias input*" $x_{n+1}=-1$ mit Gewicht (*bias*, Verschiebung) $\theta\in\mathcal{R}$ hinzu (**x** bezeichne weiter $x_1,\ldots,x_n$)
  - Perzeptron könnte $\mathbf{x}=\mathbf{0}$ ohne *bias* nicht frei klassifizieren!
  - *bias* wirkt wie variabler Schwellwert (statt >0 effektiv $>\theta$)

Das Perzeptron als Neuronales Netz



- Manche finden es hilfreich, sich das Perzeptron als „Neuronales Netz" vorzustellen
  (und so ist es historisch auch entstanden)
- Mehr dazu in 5.3
- Oft modelliert man mehrere Klassifikatoren $P_1(\mathbf{x}), P_2(\mathbf{x}), P_3(\mathbf{x})$ derselben Merkmale $\mathbf{x}$ im selben Netz (dann mit w_{ij})
- Es gibt auch mehrlagige Perzeptrons

Die Perzeptron-Lernregel

$\textsc{PerzeptronLernen}(M_+, M_-)$

w = beliebiger Vektor reeller Zahlen ungleich 0

Repeat

  For all $x \in M_+$

    If $w\,x \le 0$ **Then** $w = w + x$

  For all $x \in M_-$

    If $w\,x > 0$ **Then** $w = w - x$

Until alle $x \in M_+ \cup M_-$ werden korrekt klassifiziert

$\mathbf{x}$ und $\mathbf{w}$ sollen hier den *bias* einschließen!

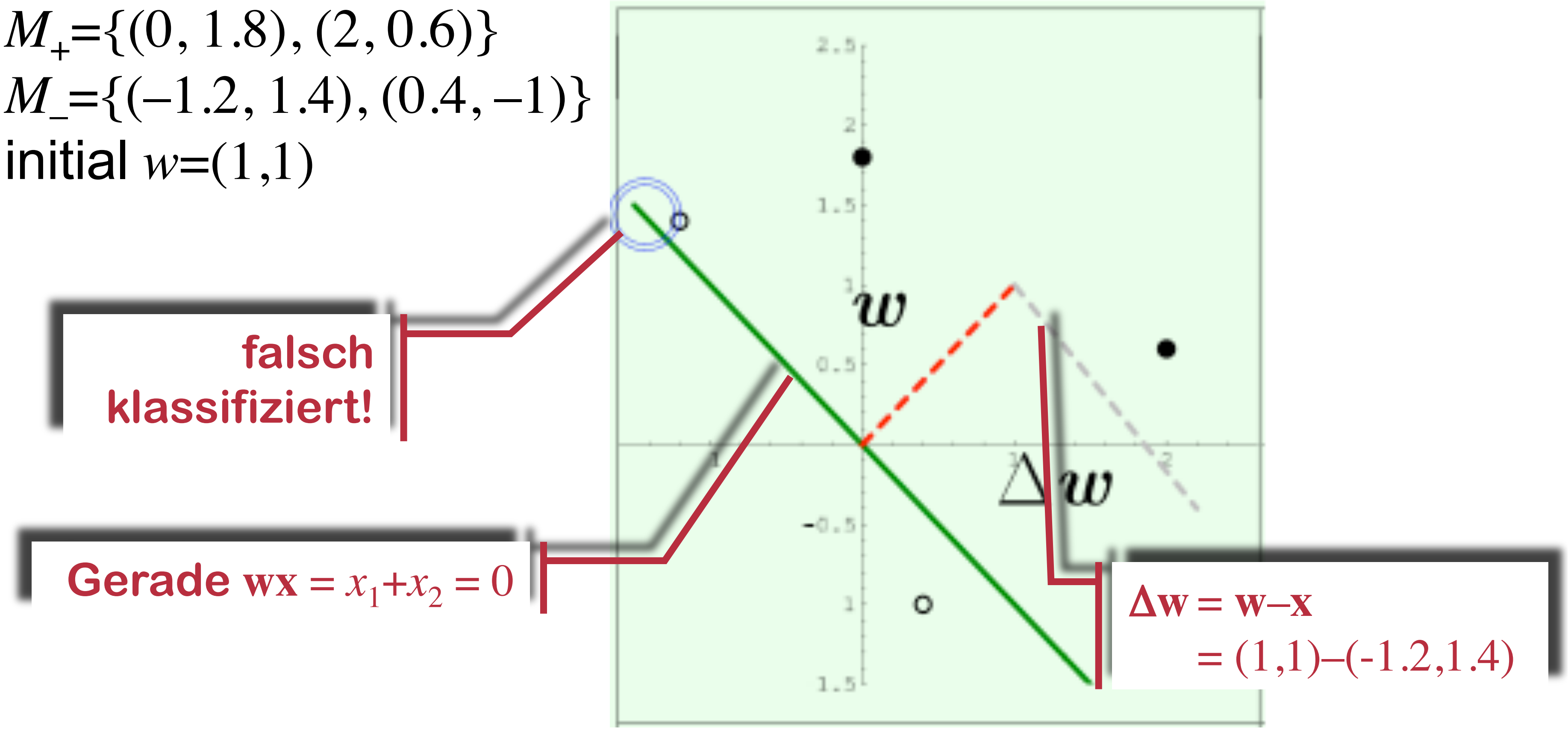
- Spezialfall für Perzeptron mit genau 1 Klassifikator
- Üblicherweise update durch $w \pm \alpha x$ für **Lernrate** $0<\alpha<1$
- Lernt das beliebige Zielfunktionen?
- Ist Terminierung gewährleistet?

Beispiel aus Ertel 1/2

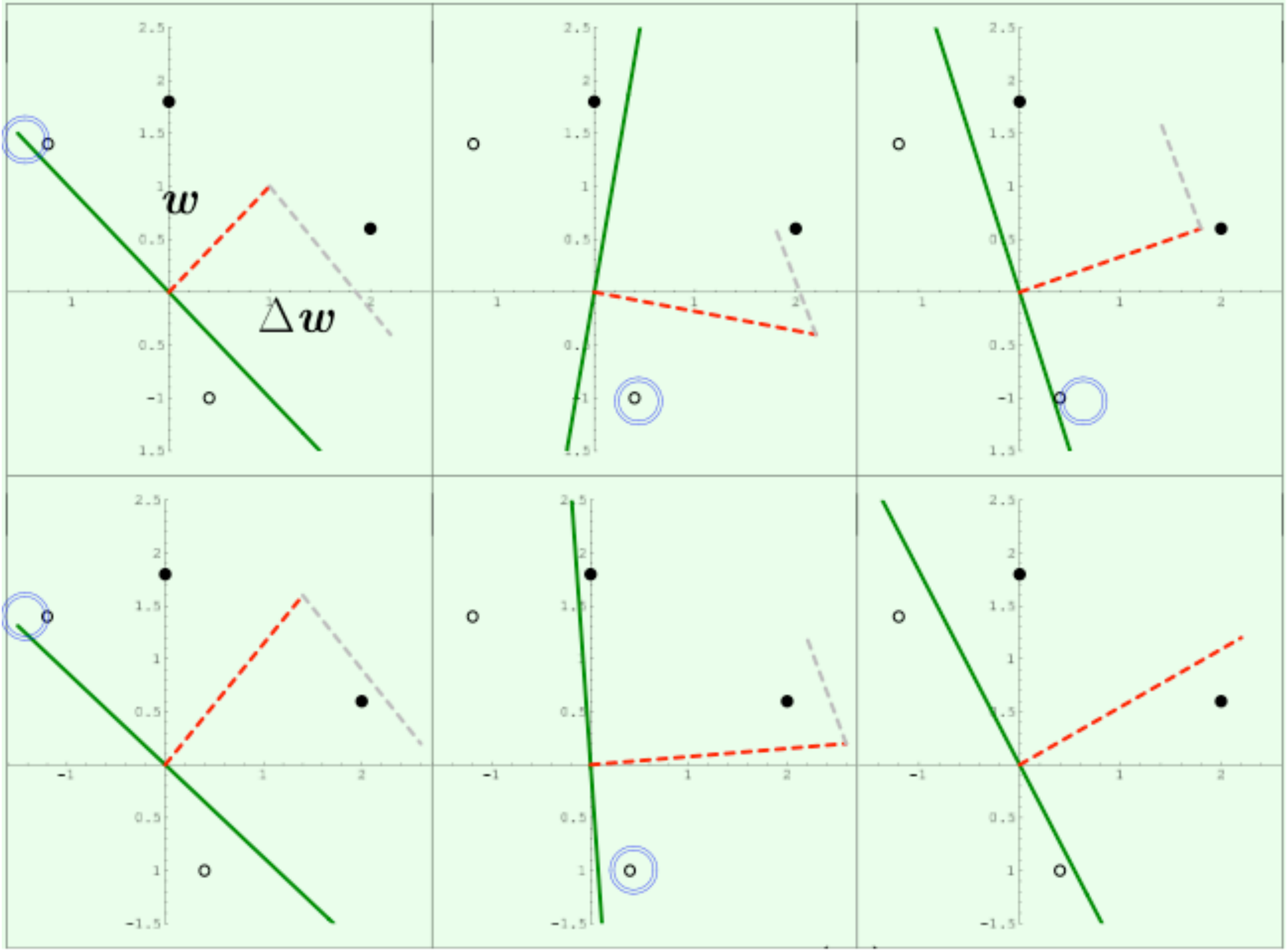
Hier ein Perzeptron mit 2 Merkmalen ohne *bias*!
➥ Hyperebene (Gerade) geht durch Nullpunkt!

Seien

- $M_+=\{(0, 1.8), (2, 0.6)\}$
- $M_-=\{(-1.2, 1.4), (0.4, -1)\}$
- initial $w=(1,1)$



Beispiel aus Ertel 2/2



Lineare Separierbarkeit

… führt zur Charakterisierung der Leistung des Perzeptrons!

Eine $(n\text{-}1)$-dimensionale Hyperebene im $\mathcal{R}^n$ ist für reelles θ gegeben durch

$$\sum_{i=1}^{n} a_i x_i = \theta$$

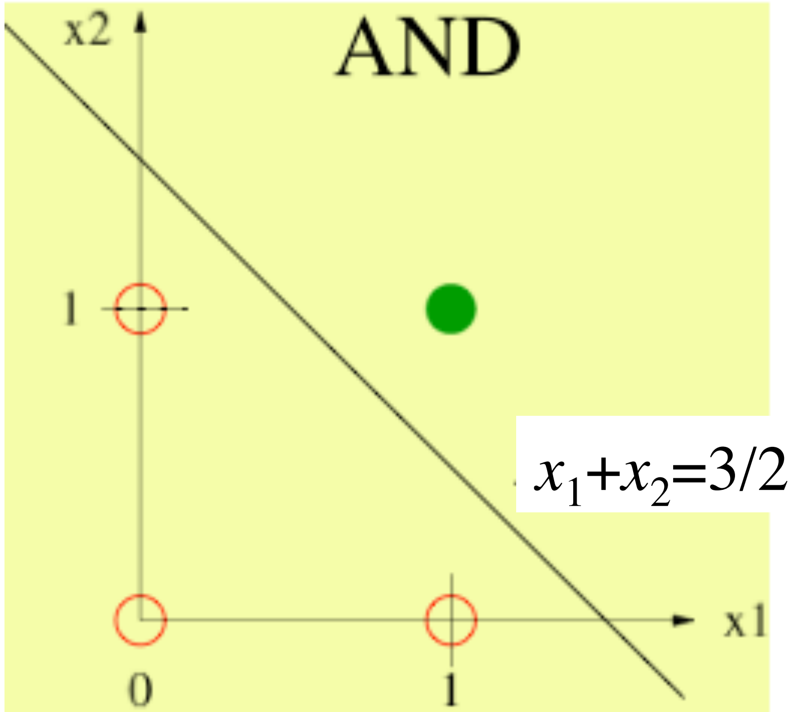
Definition 8.2 *Zwei Mengen $M_1 \subset \mathbb{R}^n$ und $M_2 \subset \mathbb{R}^n$ heißen linear separabel, wenn reelle Zahlen $a_1, \ldots, a_n, \theta$ existieren mit*

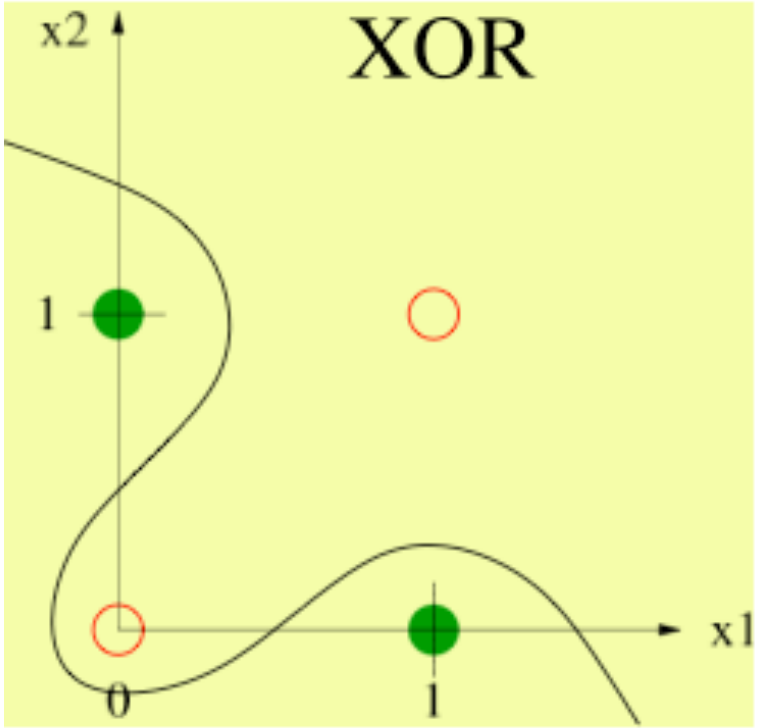
$$\sum_{i=1}^{n} a_i x_i > \theta \quad \text{für alle } \boldsymbol{x} \in M_1 \qquad \text{und} \qquad \sum_{i=1}^{n} a_i x_i \leq \theta \quad \text{für alle } \boldsymbol{x} \in M_2.$$

Der Wert θ wird als Schwelle bezeichnet.

Beispiel Lineare Separierbarkeit

Seien im $\mathcal{R}^2$ die Punkte ● in M_+ und die ○ in M_-





AND ist linear separierbar, XOR nicht!

Konvergenz der Perzeptron-Lernregel

Satz 8.1 *Es seien die Klassen* M_+ *und* M_- *linear separabel durch eine Hyperebene* $\boldsymbol{w}\,\boldsymbol{x} = 0$. *Dann konvergiert die* PERZEPTRONLERNEN *für jede Initialisierung* ($\neq 0$) *des Vektors* w. *Das Perzeptron* P *mit dem so berechneten Gewichtsvektor trennt die Klassen* M_+ *und* M_-, *d.h.*

$$P(\boldsymbol{x}) = 1 \quad \Leftrightarrow \quad \boldsymbol{x} \in M_+$$

und

$$P(\boldsymbol{x}) = 0 \quad \Leftrightarrow \quad \boldsymbol{x} \in M_-.$$

Beweisidee (Rosenblatt 1958)

Zeige, dass die Lernregel den Betrag des Fehlers im Mittel in jedem Schritt um einen Mindestbetrag reduziert. (Gradientenabstieg)

Das Perzeptron-Theorem

Satz 8.2 *Eine Funktion* $f : \mathbb{R}^n \rightarrow \{0, 1\}$ *kann von einem Perzeptron genau dann dargestellt werden, wenn die beiden Mengen der positiven und negativen Eingabevektoren linear separabel sind.*

Beweis: Nach Konstruktion der Perzeptron-Definition mit bias

- Dass das Perzeptron „nur“ linear separieren kann, hat das Gebiet Neuronale Netze für ca 15 Jahre fast stillgelegt!
- Andere, ausdrucksmächtigere Neuronale Netze und Lernregeln s. 5.3