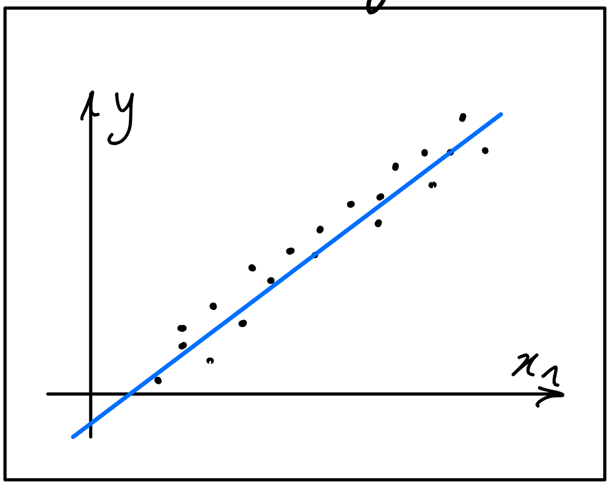
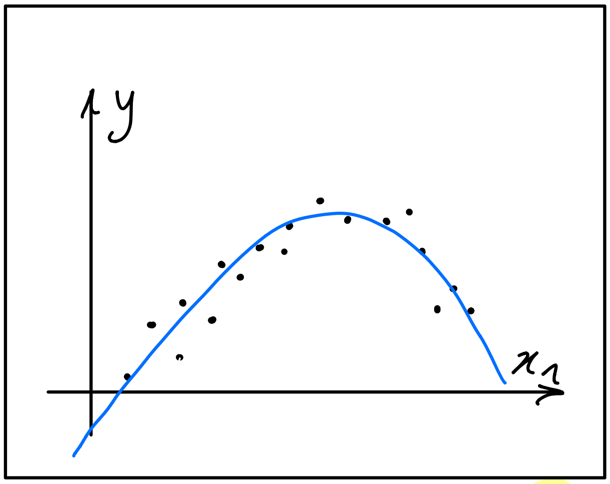
• NICHTLINEARE MODELLE

- Wie können wir nichtlineare Beziehungen erfassen?

- Nichtlineare Regression

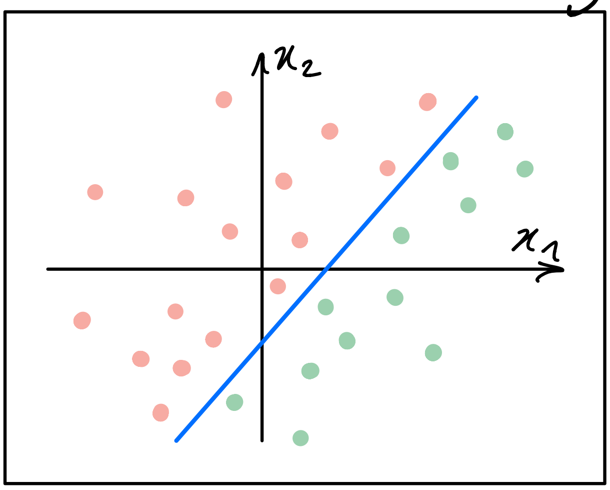


$h(x) = \omega_0 + \omega_1 x_1$ ✓

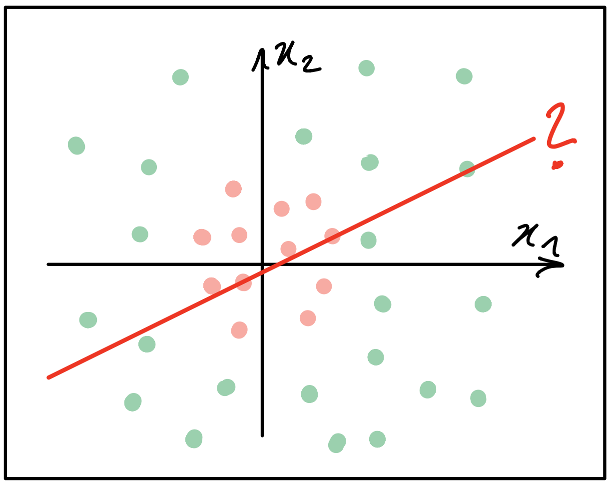


$h(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_1^2$

- Nichtlineare Klassifizierung



$h(x) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2)$ ✓



$h(x) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$

- **IDEE:** Kombiniere Merkmale x_1, x_2 auf nichtlineare Weise und führe neue Merkmale ein, eg.

$$\underbrace{x_1^2}_{x_3},\ \underbrace{x_1^3}_{x_4},\ \underbrace{x_1^4}_{x_5},\ \underbrace{x_2^2}_{x_6},\ \underbrace{x_1 \cdot x_2}_{x_7},\ \underbrace{x_1^2 x_2}_{x_8},\ \ldots$$

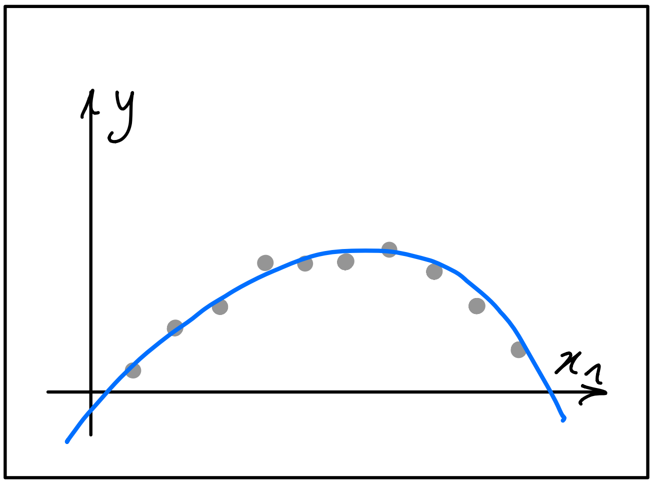
(Neue Merkmale!)

- **Bemerkung**: $w^T x$ ist weiterhin linear in w!

$$w^T x = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 \cdot x_1^3 + w_4 \cdot x_1^4 + \cdots$$

Zahlen

BSP. Polynomiale Regression (n=1)



- $D = \{(1, 0.5), (2, 1.5), (3, 2), (4, 3), (5, 3), (6, 3), (7, 3.2), (8, 2.9), (9, 2), (10, 1)\}$
- Nichtlineare Beziehung zwischen x_1 und y

- Kein gutes $h(x) = w_0 + w_1 x_1$ möglich

  $\longrightarrow$ Kosten gehen nicht herunter!

- Neues Merkmal $x_2 := x_1^2$ einführen: $x = (x_1, x_1^2) = (x_1, x_2)$

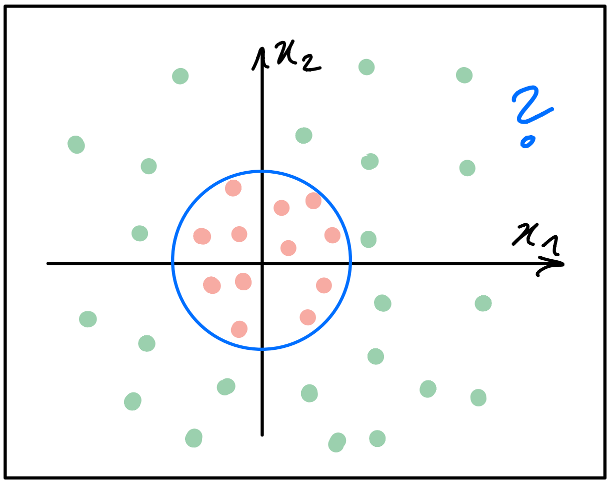
  $\longrightarrow$ $x^{(1)} = (1, 1)$, $x^{(2)} = (2, 4)$, $x^{(3)} = (3, 9)$, ...

- Hypothesenfunktion: $h(x) = w_0 + w_1 x_1 + w_2 x_1^2$

  $\longrightarrow$ $h(x) = -0.5 + 1.2\, x_1 - 0.1\, x_1^2$ besser

  $\longrightarrow$ $h(x)$ lineare Funktion in x_1 und x_2 (quadratisch in x_1)

BSP. # Polynomiale Klassifikation $(n=2)$



- $\mathcal{X} = \mathbb{R}^2 \qquad \mathcal{Y} = \{0,1\}$
- Daten nicht linear separierbar.
- Keine gute lineare Hypothese
  $h(x) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2)$

- Durch die Einführung der neuen Merkmale $x_3 := x_1^2$ und $x_4 := x_2^2$ erhalten wir die neue Hypothesenfunktion:

$$h_\omega(x) = \sigma\left(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 \cdot x_1^2 + \omega_4 \cdot x_2^2\right)$$

- Sei $\omega = (-4, 0, 0, 1, 1)$

- Entscheidungsgrenze : $\omega^T x = 0$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

$$-4 + 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_1^2 + 1 \cdot x_2^2 = 0$$

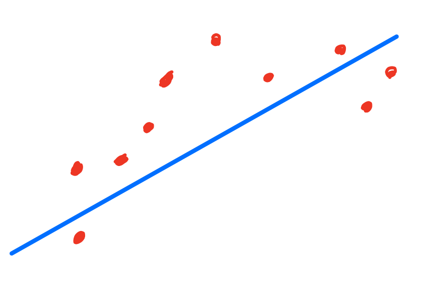
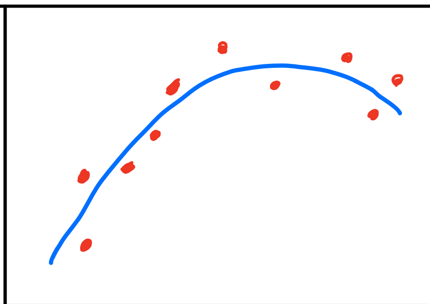
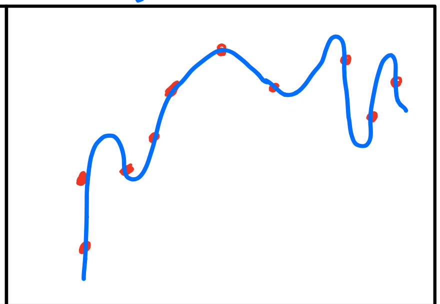
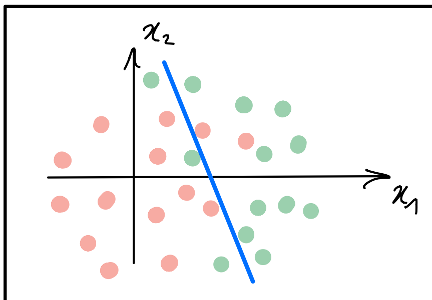
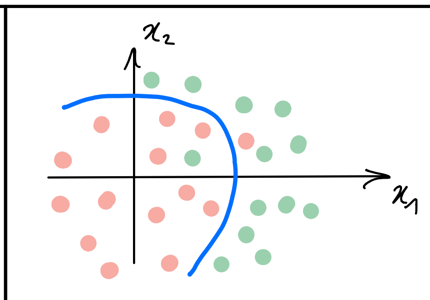
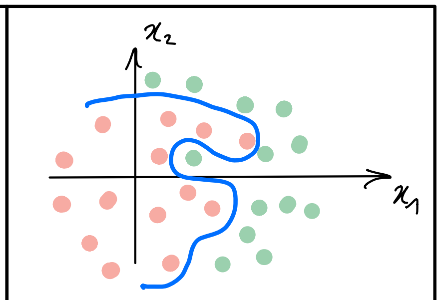
$$\underbrace{x_1^2 + x_2^2 = 4}_{\text{Entscheidungsgrenze}} \qquad (\text{Kreis mit Radius } 2.)$$

- $-4 + x_1^2 + x_2^2 < 0 \implies x_1^2 + x_2^2 < 4$

  $\implies$ Das Innere des Kreises wird negativ klassifiziert.

• OVERFITTING (Überanpassung)

z.B. Polynom mit Grad 10

Regression			
	• Unteranpassung Underfit • High Bias (unflexibel) • Modell zu einfach	• gute Anpassung "Right" Fit (flexibel genug)	• Überanpassung Overfit • High Variance (zu flexibel) • Modell zu komplex
Klassifikation			
	• $\sigma(\omega_0+\omega_1 x_1+\omega_2 x_2)$ (zu wenig Merkmale)	• $\sigma(\omega_0+\omega_1 x_1+\omega_2 x_2$ $+\omega_3 x_1^2+\omega_4 x_2^2+\omega_5 x_1 x_2)$	• $\sigma(\omega_0+\omega_1 x_1+\omega_2 x_2$ $+\omega_3 x_1^2+\omega_4 x_2^2+\omega_5 x_1 x_2$ $+\omega_6 x_1^3+\omega_7 x_1^2 x_2^2+\cdots$ (zu viele Merkmale)

• Die Komplexität des Modells ist zu hoch; höher als es die Zielfunktion erfordert. Das Modell nutzt die extra "Freiheitsgrade", um sich den Eigenartigkeiten und "Noise" in den Daten anzupassen.

- „Was im Bereich des maschinellen Lernens Professionelle von Amateuren unterscheidet, ist ihre Fähigkeit mit Überanpassung umzugehen." (Abu Mostafa, Learning from Data)

- Diagnose
  - Sehr sehr gute Performanz auf Trainingsdaten

    → Kosten $J(\theta) \approx 0$

    Sehr schlechte Performanz auf neuen unbekannten Daten

    → Schlechte Generalisierung / Vorhersagekraft !

  - Hypothesenfunktion lernt "gesehene" Daten auswendig, kann nicht gesehene Beispiele nicht gut vorhersagen.

- Maßnahmen
  - Reduziere #Merkmale (reduziere Komplexität des Modells)

    → Welche Merkmale weglassen?

  - Regularisierung: behalte alle Merkmale

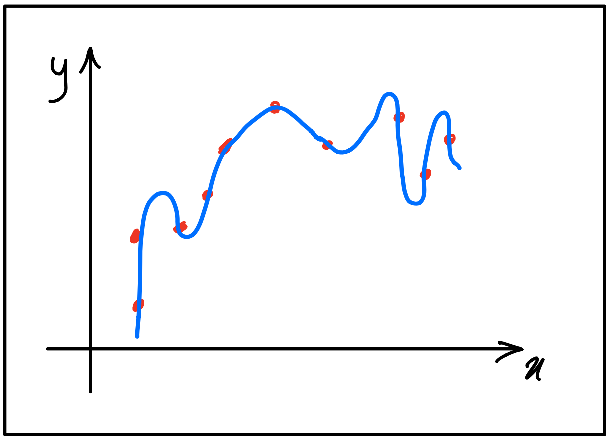
    reduziere Absolutwerte der Gewichte w_j

    → führt zu einfacheren Modellen.

canan.yildiz@tau.edu.tr

• Regularisierung – Intuition

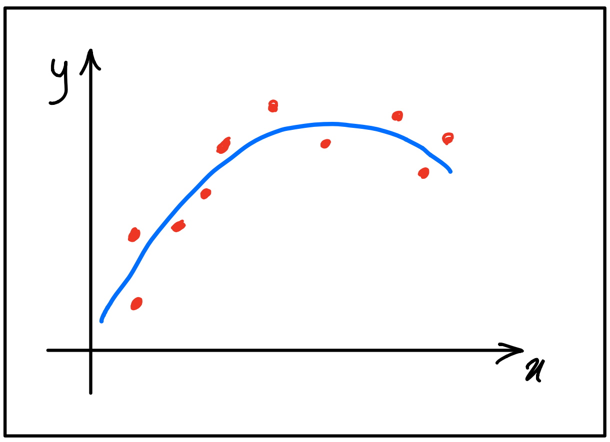
- $h(x) = w^T x = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6$



Overfit

- Lasse die Merkmale x^3, x^4, x^5, x^6 weg $(w_3 = w_4 = w_5 = w_6 = 0)$
  (Terme höherer Ordnung weggelassen)

$$h(x) = w^T x = w_0 + w_1 x + w_2 x^2$$



Right Fit

- Reduziere Absolutwerte von w_1, w_2, w_3, w_4, w_5 und w_6 mit Hilfe der Kostenfunktion:

$$\min_{w} \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)} \right)^2 + 1000 \cdot \left(\underbrace{w_1^2}_{=0} + \underbrace{w_2^2}_{=0} + \cdots + \underbrace{w_6^2}_{\approx 0} \right) \right]$$

- Regularisierte Lineare Regression

Kostenfunktion

Regularisierungsparameter

$$\boxed{J(\omega) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h(x^{(i)})-y^{(i)}\right)^2 + \lambda \cdot \sum_{j=1}^{n} \omega_j^2\right]}$$

ω_0 nicht regularisieren!

- Regularisierte Logistische Regression

Kostenfunktion

$$\boxed{J(\omega) = \frac{1}{m}\left[\sum_{i=1}^{m}\left(-y^{(i)}\cdot\log(h(x^{(i)})) - (1-y^{(i)})\cdot\log(1-h(x^{(i)}))\right) + \underbrace{\frac{\lambda}{2}\cdot\sum_{j=1}^{n}\omega_j^2}_{\partial/\partial\omega_j}\right]}$$

$$\frac{\partial}{\partial \omega_j}\left(\frac{\lambda}{2}\left(\omega_1^2+\omega_2^2+\dots+\omega_j^2+\dots+\omega_n^2\right)\right) = \frac{\lambda}{2}\cdot 2\omega_j = \lambda\cdot\omega_j$$

- Gewichtsaktualisierung (für $j \geq 1$)

$$\boxed{\omega_j := \omega_j - \frac{\alpha}{m}\cdot\left[\sum_{i=1}^{m}\left(\left(h(x^{(i)})-y^{(i)}\right)\cdot x_j^{(i)}\right) + \lambda\cdot\omega_j\right]}$$