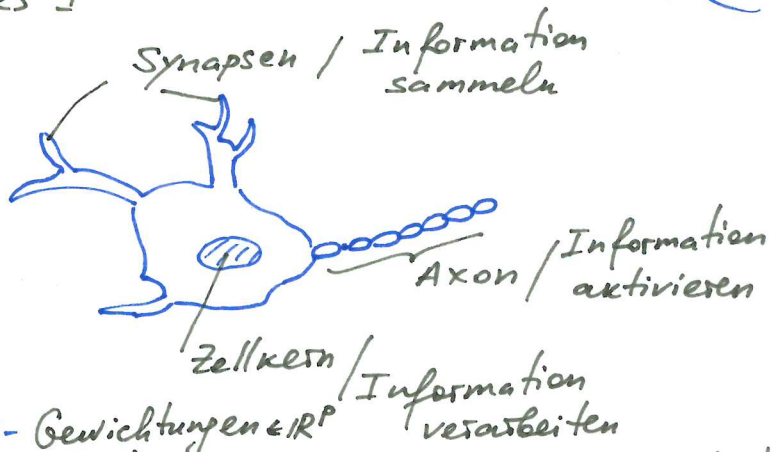


Neural Networks I

1

Neuron / Nervenzelle:

biologisch:
(natürlich)

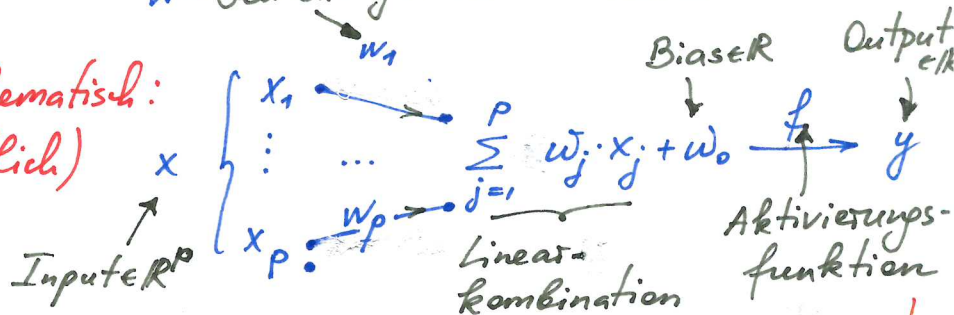


$$y = f(x^T \cdot w)$$

wobei

$$x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix}, w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{pmatrix}$$

mathematisch:
(künstlich)

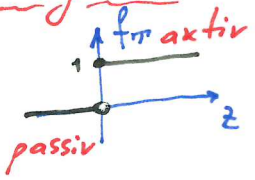


$(\varphi_j(x) = x_j, j=1, \dots, p)$
verallgemeinerte
Lineare Regression

Aktivierung: Schwelle

1) Threshold-Funktion:

$$f_T(z) = \begin{cases} 1, & \text{falls } z \geq 0 \\ 0, & \text{falls } z < 0 \end{cases}$$



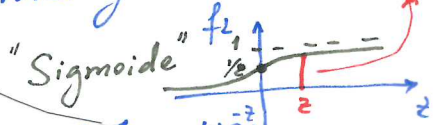
↑ deterministisch

McCulloch-Pitt-Modell

↓ stochastisch

2) Logistische Funktion:

$$f_L(z) = \frac{1}{1 + e^{-z}} - \text{Aktivierungswahrscheinlichkeit}$$



Weber-Fechner Law

$$\ln \frac{f_L(z)}{1 - f_L(z)} = \ln \left(\frac{1}{1 + e^{-z}} : \left(1 - \frac{1}{1 + e^{-z}} \right) \right) = \ln \left(\frac{1}{1 + e^{-z}} : \frac{e^{-z}}{1 + e^{-z}} \right) = \ln e^z = z$$

Psychophysik

logarithmischer
Zusammenhang

objektive Intensität des Inputs/Reizes

subjektiv empfundene
Stärke des Inputs/Reizes

Daten: $x^i \in \mathbb{R}^{p+1}, y^i \in \{0, 1\}, i=1, \dots, n$ ← Anzahl der Datensätze

"Classification!"

$$\text{Bayes-Ansatz: } L(w) = \prod_{i=1}^n f_L((x^i)^T w)^{y_i} \cdot (1 - f_L((x^i)^T w))^{1-y_i}$$

~ Bayes a-posteriori
mit gleichverteiltem
Prior w

← i.i.d. Messungen von Daten

$$p(y^i=1) = f((x^i)^T w)$$

$$p(y^i=0) = 1 - f((x^i)^T w)$$

~ Bernoulli ($f((x^i)^T w)$)

← Parameter p

Maximum Likelihood
max w

$$\ln L(w) = \sum_{i=1}^n y_i \cdot \ln f_L((x^i)^T w) + (1 - y_i) \cdot \ln (1 - f_L((x^i)^T w))$$

cross-entropy

Logistische Regression

$$\min_w \frac{1}{n} \sum_{i=1}^n \left(-y_i \cdot \ln f_z((x^i)^T w) - (1-y_i) \ln (1-f_z((x^i)^T w)) \right)$$

↑
Mittelung

= $F_i(w)$ - i-te Kreuzentropie

=: $F(w)$ - Kreuzentropie

- $F_i(w)$ ist konvex in w , denn:

$$\left[f'_z(z) = \left(\frac{1}{1+e^{-z}} \right)' = \frac{+e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{-z}} \cdot \frac{1}{e^z+1} = f_z(z) \cdot (1-f_z(z)) \right]$$

$$\nabla F_i(w) = -y^i \frac{1}{f_z((x^i)^T w)} \cdot f'_z((x^i)^T w) \cdot x^i - (1-y^i) \cdot \frac{1}{1-f_z((x^i)^T w)} \cdot (-f'_z((x^i)^T w)) \cdot x^i$$

↑
Gradient

$$z = (x^i)^T w$$

$$= -y^i \frac{1}{f_z(z)} \cdot f_z(z) \cdot (1-f_z(z)) \cdot x^i + (1-y^i) \frac{1}{1-f_z(z)} \cdot f_z(z) \cdot (1-f_z(z)) \cdot x^i$$

$$= [-y^i + f_z(z)] \cdot x^i = [f_z((x^i)^T w) - y^i] \cdot x^i$$

Hesse-Matrix $\in \{0,1\}$ w -keit der Aktivierung für $z = (x^i)^T w$

$$H_{F_i}(w) = \nabla [f_z((x^i)^T w) - y^i] \cdot (x^i)^T = f'_z((x^i)^T w) \cdot x^i \cdot (x^i)^T = \underbrace{f_z(z)}_{>0} \cdot \underbrace{(1-f_z(z))}_{>0} \cdot \underbrace{x^i \cdot (x^i)^T}_{\substack{1 \times p \\ p \times 1}}$$

⇒ konvexe Funktion $F_i(w)$: 

Rank 1

positive
semidefinit

- Gradienten von $F_i(w)$ sind beschränkt:

$$\xrightarrow{\text{Euklidische Norm}} \|\nabla F_i(w)\| = \|[f_z((x^i)^T w) - y^i] \cdot x^i\| = \underbrace{|f_z((x^i)^T w) - y^i|}_{\in [0,1]} \cdot \underbrace{\|x^i\|}_{\in \{0,1\}} \leq 2 \|x^i\|$$

$$\leq 2 \cdot \max_{1 \leq i \leq n} \|x^i\| =: R \quad \forall i=1, \dots, n.$$

Stochastisches Gradientenverfahren:

1. Wähle $i \in \{1, \dots, n\}$ mit Wahrscheinlichkeit $\frac{1}{n}$.

2. $w(t+1) := w(t) - \alpha \cdot \nabla F_i(w(t))$
 Aktivierungsfehler zum i-ten Input
 ↓ Schrittweite
 i-te Funktion
 Input i

$$\Leftrightarrow w(t+1) = w(t) - \alpha \cdot [f_z((x^i)^T w(t)) - y^i] \cdot x^i$$

LEARNING

↑ w -keit der Aktivierung beim i-ten Input

Konvergenzanalyse

(3)

$$\|z\|^2 = z^T z$$

$$\|w(t+1) - w^*\|^2 = \|w(t) - w^* - \alpha \cdot \nabla F_i(w(t))\|^2 =$$

Lösung

$$= \|w(t) - w^*\|^2 - 2\alpha \nabla^T F_i(w(t)) \cdot (w(t) - w^*) + \alpha^2 \underbrace{\|\nabla F_i(w(t))\|^2}_{\leq R^2}$$

→ Erwartungswert bzgl. $i \in \{1, \dots, n\}$

gleichverteilte Zufallsvariable

$$\mathbb{E} \|w(t+1) - w^*\|^2 \leq \mathbb{E} \|w(t) - w^*\|^2 - 2\alpha \mathbb{E}(\nabla^T F_i(w(t))) \cdot (w(t) - w^*) + \alpha^2 R^2$$

$$= \frac{1}{n} \sum_{i=1}^n \nabla^T F_i(w(t)) = \nabla^T F(w(t))$$

Funktion

Konvexität von F :

$$F(w^*) \geq F(w(t)) + \nabla^T F(w(t)) \cdot (w^* - w(t))$$

Tangente

$$\mathbb{E} \|w(t+1) - w^*\|^2 \leq \mathbb{E} \|w(t) - w^*\|^2 + 2\alpha \cdot (F(w^*) - F(w(t))) + \alpha^2 R^2$$

→ Summiere über $t=1, 2, \dots, T$

$$0 \leq \mathbb{E} \|w(T+1) - w^*\|^2 \leq \mathbb{E} \|w(1) - w^*\|^2 + 2\alpha \cdot \sum_{t=1}^T [F(w^*) - F(w(t))] + \alpha^2 R^2 T$$

$$= T \cdot F(w^*)$$

$$\sum_{t=1}^T F(w(t)) - T F(w^*) \leq \frac{\mathbb{E} \|w(1) - w^*\|^2 + \alpha^2 R^2 T}{2\alpha}$$

$$\frac{1}{T} \sum_{t=1}^T F(w(t)) - F(w^*) \leq \frac{\mathbb{E} \|w(1) - w^*\|^2}{2\alpha \cdot T} + \frac{\alpha \cdot R^2}{2}$$

$$\geq F\left(\frac{1}{T} \sum_{t=1}^T w(t)\right)$$

Konvex

$\bar{w}(t)$ - Mittelung von Gewichten

$$0 \leq F(\bar{w}(t)) - F(w^*) \leq \frac{\mathbb{E} \|w(1) - w^*\|^2}{2\alpha \cdot T} + \frac{\alpha \cdot R^2}{2} \approx O\left(\frac{1}{\sqrt{T}}\right)$$

immer

Konvergenzrate

Wie ist die Schrittweite zu wählen?

→ min α

$$\nabla_{\alpha} : -\frac{\mathbb{E} \|w_1 - w^*\|^2}{2\alpha^2 \cdot T} + \frac{R^2}{2} = 0 \Leftrightarrow \alpha = \underbrace{\frac{\sqrt{\mathbb{E} \|w_1 - w^*\|^2}}{R}}_{\text{Konstante}} \cdot \frac{1}{\sqrt{T}}$$