

Verlustfunktion in Neuronales Netzwerk

SD62-1 Dr. Lingquan Zhao

29.05.2024

Wie die Verlustfunktion gestaltet wird?

Verlustfunktion - loss function $L(x, y, \hat{y})$

die allgemeinste Formulierung der Verlustfunktion:

$$L(x, y, \hat{y}) = \text{Utility(result of using } y \text{ given an input } x) - \text{Utility(result of using } \hat{y} \text{ given an input } x)$$

Oft wird eine vereinfachte Version verwendet:

$$L(y, \hat{y}) = \text{Utility(result of using } y \text{ given an input } x) - \text{Utility(result of using } \hat{y} \text{ given an input } x)$$



das Erkennen von E-Mail-Nachrichten als **Spam oder Nicht-Spam**

- Nicht-Spam als Spam einzustufen (und damit möglicherweise eine wichtige Nachricht zu übersehen): **0.5 % Fehlerquote**
- Spam als Nicht-Spam zu klassifizieren (und damit ein paar Sekunden Ärger zu erleiden): **1 % Fehlerquote**

Wie die Verlustfunktion gestaltet wird?

Potter Stewart (1915-1985)

US-amerikanischer Jurist und von 1958 bis Juli 1981
beisitzender Richter am Obersten Gerichtshof der
Vereinigten Staaten.

I shall not today attempt
further to define the
kinds of material
[pornography] . . . but I
know it when I see it.

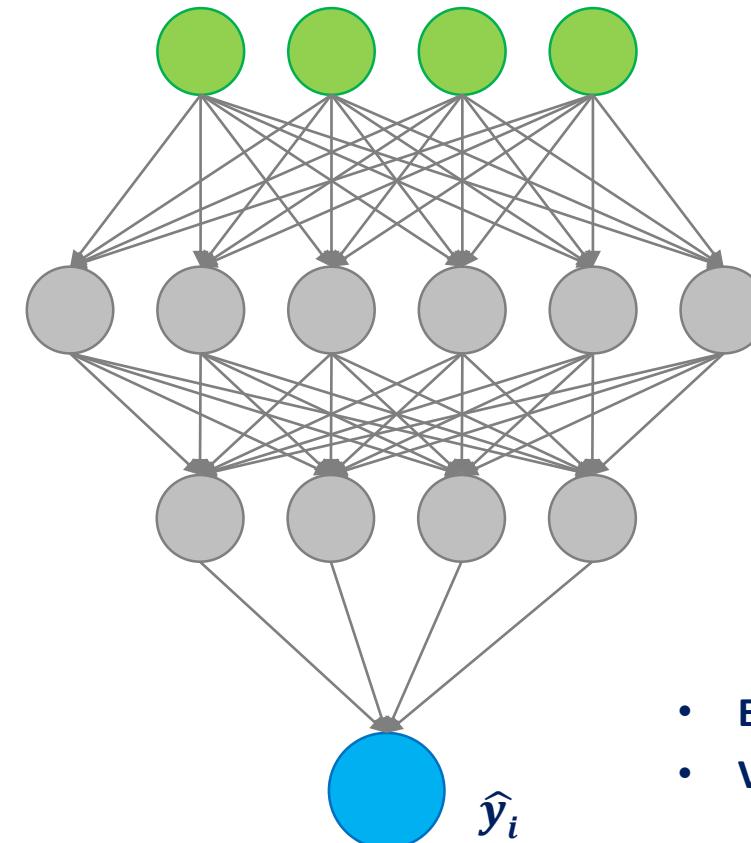


Für die breitere Öffentlichkeit mag Stewart durch ein Zitat, oder einem Fragment daraus, besonders bekannt geworden sein, das in dem Fall Jacobellis gegen Ohio (1964) gefallen ist. Stewart schrieb in seiner kurzen Einverständniserklärung, dass „hard-core pornography“ schwer zu definieren sei, aber (frei übersetzt) „ich erkenne sie, wenn ich sie sehe.“

Wall Street Journal, 2007

Wie die Verlustfunktion gestaltet wird?

nicht in der Realität existieren



Logistic Regression cost function

$$\hat{y}^{(i)} = \sigma(w^T x^{(i)} + b), \text{ where } \sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}} \quad z^{(i)} = w^T x^{(i)} + b$$

Given $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, want $\hat{y}^{(i)} \approx y^{(i)}$.

Loss (error) function: $L(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

$$L(\hat{y}, y) = -[y \log \hat{y} + (1-y) \log(1-\hat{y})] \leftarrow$$

If $y=1$: $L(\hat{y}, y) = -\log \hat{y} \leftarrow$ want $\log \hat{y}$ large, want \hat{y} large.
If $y=0$: $L(\hat{y}, y) = -\log(1-\hat{y}) \leftarrow$ want $\log(1-\hat{y})$ large ... want \hat{y} small

- Echter Wert:** $y_i \in \{0, 1\}$
- Vorhersagewert:** $\hat{y}_i \sim (0, 1)$

Methode der kleinsten Quadrate

Logistic Regression cost function

$$\hat{y}^{(i)} = \sigma(w^T x^{(i)} + b), \text{ where } \sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}} \quad z^{(i)} = w^T x^{(i)} + b$$

Given $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, want $\hat{y}^{(i)} \approx y^{(i)}$.

Loss (error) function: $L(\hat{y}, y) = \frac{1}{2}(\hat{y}-y)^2$

$$L(\hat{y}, y) = -(\underbrace{y \log \hat{y}}_{\text{If } y=1} + \underbrace{(1-y) \log(1-\hat{y})}_{\text{If } y=0}) \leftarrow$$

If $y=1$: $L(\hat{y}, y) = -\log \hat{y} \leftarrow$ Want $\log \hat{y}$ large, want \hat{y} large.

If $y=0$: $L(\hat{y}, y) = -\log(1-\hat{y}) \leftarrow$ Want $\log(1-\hat{y})$ large ... want \hat{y} small

$$L(\hat{y}, y) = \frac{1}{2}(\hat{y}-y)^2$$

$$L(\hat{y}, y) = -(\underbrace{y \log \hat{y}}_{\text{If } y=1} + \underbrace{(1-y) \log(1-\hat{y})}_{\text{If } y=0}) \leftarrow$$

Prof. Andrew Ng, 2018

Methode der kleinsten Quadrate

Logistic Regression cost function

$\hat{y}^{(i)} = \sigma(w^T \underline{x}^{(i)} + b)$, where $\sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}}$ $z^{(i)} = w^T \underline{x}^{(i)} + b$
 Given $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, want $\hat{y}^{(i)} \approx y^{(i)}$.

Loss (error) function: $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$

$L(\hat{y}, y) = -[y \log \hat{y} + (1-y) \log(1-\hat{y})] \leftarrow$
 If $y=1$: $L(\hat{y}, y) = -\log \hat{y} \leftarrow$ Want $\log \hat{y}$ large, want \hat{y} large.
 If $y=0$: $L(\hat{y}, y) = -\log(1-\hat{y}) \leftarrow$ Want $\log(1-\hat{y})$ large ... want \hat{y} small

$$L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

Prof. Andrew Ng, 2018

$$L_i(y, \hat{y}) = \min \sum_{i=1}^n |y - \hat{y}|$$

$$L(y, \hat{y}) = \min \sum_{i=1}^n \frac{1}{2}(y - \hat{y})^2$$

$$L_{ii}(y, \hat{y}) = \min \sum_{i=1}^n (y - \hat{y})^2$$

Maximum-Likelihood-Schätzung

In der **Statistik** ist die **Maximum-Likelihood-Schätzung** (MLE) eine Methode zur Schätzung der Parameter einer angenommenen Wahrscheinlichkeitsverteilung **bei bestimmten beobachteten Daten**. Dies wird durch die Maximierung einer Likelihood-Funktion erreicht, so dass die beobachteten Daten unter dem angenommenen statistischen Modell **am wahrscheinlichsten** sind.

Likelihood-Schätzung: eine umgekehrte Anwendung der Wahrscheinlichkeitsrechnung



Maximum-Likelihood-Schätzung

Die **Maximum-Likelihood-Schätzung** wird häufig verwendet, wenn wir **das Ergebnis bereits kennen** und auf der Grundlage des Ergebnisses auf das probabilistische Modell zurückgreifen, das dieses Verhalten hervorgebracht hat.

Likelihood-Schätzung: eine umgekehrte Anwendung der Wahrscheinlichkeitsrechnung



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,5	0,5



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,5	0,5



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,5	0,5



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,5	0,5



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,5	0,5



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	x	x



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p	0,7	0,3



Die Welt der Realität



Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p_1	0,7	0,3
p_2	0,2	0,8
p_3	0,1	0,9



Die Welt der Realität



Maximum-Likelihood-Schätzung

Bedingte Wahrscheinlichkeit:

$$P(C_1, C_2, C_3, \dots, C_{10} | \theta) = \prod_{i=1}^{10} P(C_i | \theta)$$

unter der Bedingung θ :

x	Adler	Fußball
p_1	0,7	0,3
p_2	0,2	0,8
p_3	0,1	0,9

$C_1, C_2, C_3, \dots, C_{10}$ sind unabhängige Ereignisse

Das Produktzeichen \prod (Pi) kennzeichnet die Multiplikation von mehreren mathematischen Objekten.

Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p ₁	0,7	0,3
p ₂	0,8	0,2
unter der Bedingung θ3:	p₃	0,1
		0,9



Die Welt der Realität



Bedingte Wahrscheinlichkeit:

$$P(C_1, C_2, C_3, \dots, C_{10} | \theta) = 0,1^7 * 0,9^3 = 0,0000000729$$

Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p ₁	0,7	0,3
p ₂	0,8	0,2
p ₃	0,1	0,9

unter der Bedingung θ2:



Die Welt der Realität



Bedingte Wahrscheinlichkeit:

$$P(C_1, C_2, C_3, \dots, C_{10} | \theta) = 0,8^7 * 0,2^3 = 0,0016777216$$

Maximum-Likelihood-Schätzung

Die Welt der Gedanken

x	Adler	Fußball
p_1	0,7	0,3
p_2	0,8	0,2
p_3	0,1	0,9

unter der Bedingung θ_1 :



Die Welt der Realität



Bedingte Wahrscheinlichkeit:

$$P(C_1, C_2, C_3, \dots, C_{10} | \theta) = 0,7^7 * 0,3^3 = 0,0022235661$$

Maximum-Likelihood-Schätzung

Die Welt der Gedanken

unter der Bedingung θ_1 :

x	Adler	Fußball
p_1	0,7	0,3
p_2	0,8	0,2
p_3	0,1	0,9



Die Welt der Realität



Bedingte Wahrscheinlichkeit:

$$P(C_1, C_2, C_3, \dots, C_{10} | \theta) = 0,7^7 * 0,3^3 = 0,0022235661$$

Maximum-Likelihood-Schätzung

Bedingte Wahrscheinlichkeit P_1, P_2, P_3 : Wert der Likelihood-Schätzung

Die Welt der Gedanken: Probabilistisches Model

Die Welt der Gedanken

	x	Adler	Fußball
unter der Bedingung θ1:	p_1	0,7	0,3
unter der Bedingung θ2:	p_2	0,8	0,2
unter der Bedingung θ3:	p_3	0,1	0,9

Bedingte Wahrscheinlichkeit:
 $P(C_1, C_2, C_3, \dots, C_{10} | \theta)$

$$P_1 = 0,0022235661$$

$$P_2 = 0,0016777216$$

$$P_3 = 0,0000000729$$



Die Welt der Realität



Wir können das theoretische Wahrscheinlichkeitsmodell, das das Ergebnis hervorbringt, nicht einfach bestimmen, sondern nur eine unendliche Annäherung an das Modell ermitteln.

Maximum-Likelihood-Schätzung

eine Münze werfen: $P(C_1, C_2, C_3, \dots, C_{10} | \theta)$

Objekte erkennen: $P(y_1, y_2, y_3, \dots, y_n | W, b)$

$$= \prod_{i=1}^n P(y_i | W, b)$$

$$= \prod_{i=1}^n P(y_i | \hat{y}_i)$$

$$y_i \in \{0, 1\}, \hat{y}_i \sim (0, 1) = P$$

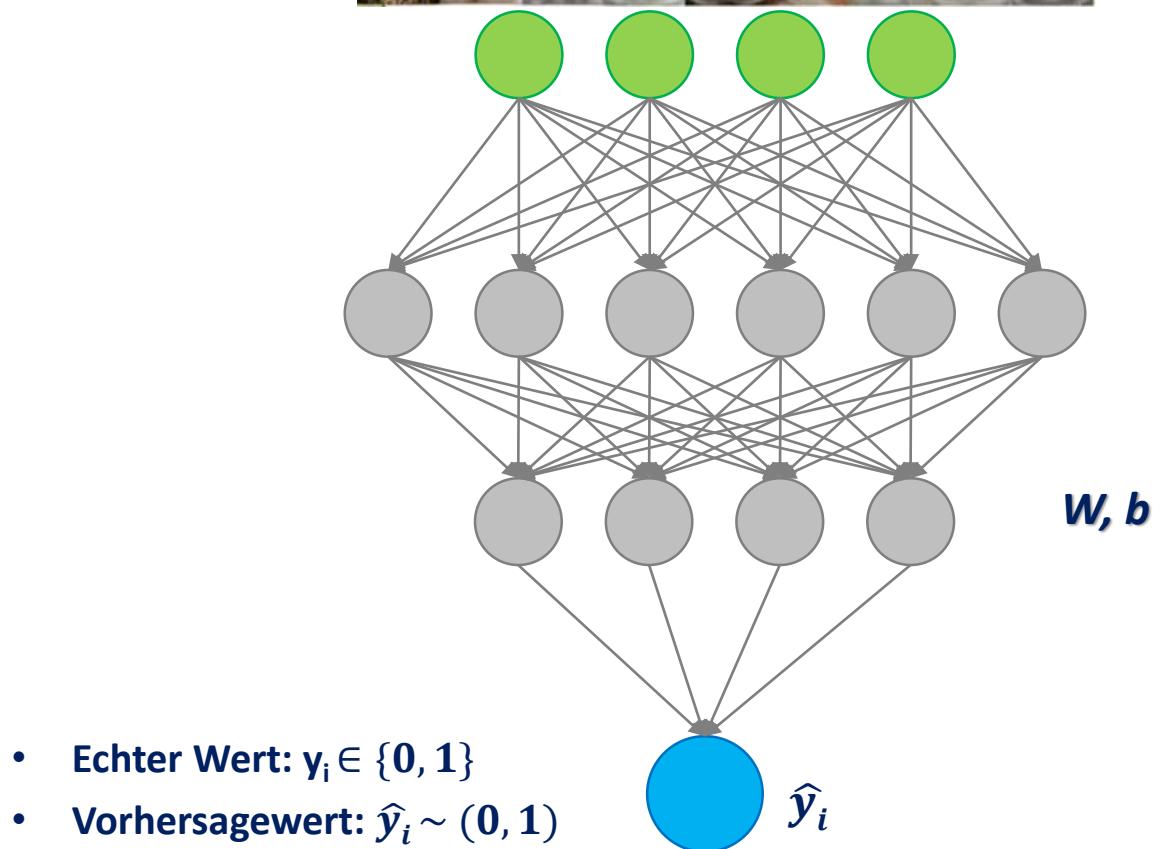
Bernoulli-Verteilung

$$f(x) = p^x (1-p)^{1-x} = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

...

$$= \prod_{i=1}^n \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}$$

$$\log (\prod_{i=1}^n \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i})$$



Maximum-Likelihood-Schätzung

Objekte erkennen: $P(y_1, y_2, y_3, \dots, y_n | W, b)$

$$= \prod_{i=1}^n P(y_i | W, b)$$

$$= \prod_{i=1}^n P(y_i | \hat{y}_i)$$

$$y_i \in \{0, 1\}, \hat{y}_i \sim (0, 1) = P$$

Bernoulli-Verteilung

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

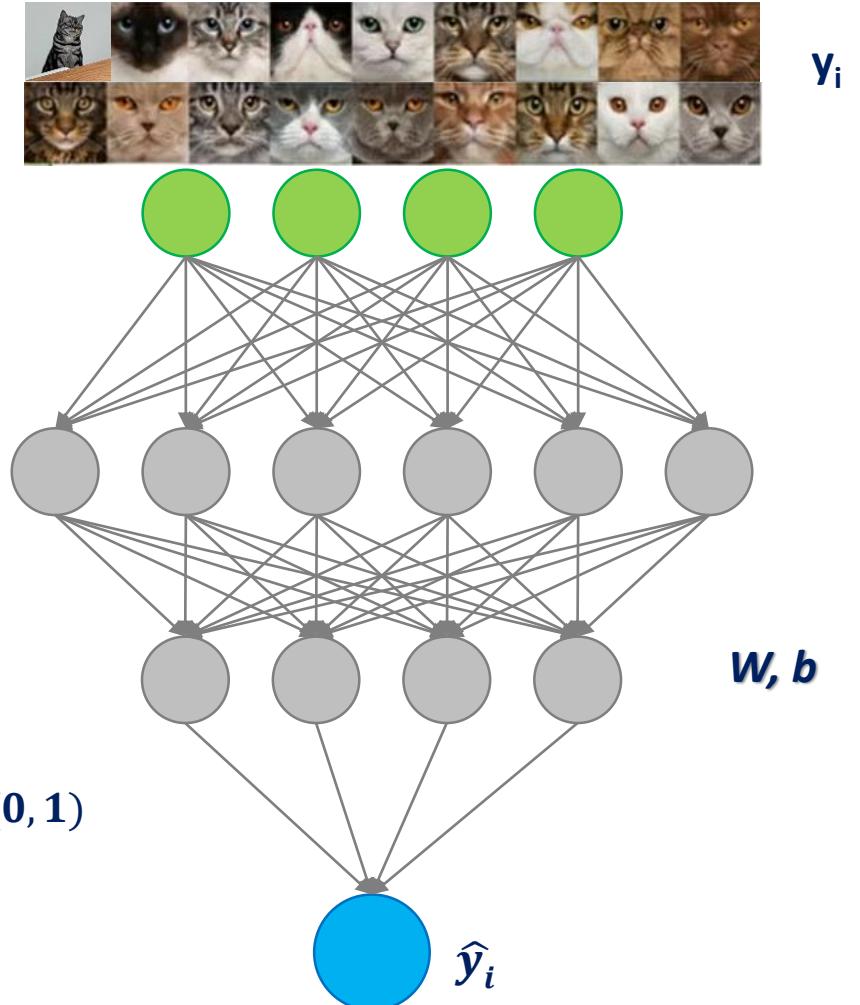
...

$$= \prod_{i=1}^n \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}$$

$$\log (\prod_{i=1}^n \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i})$$

$$= \sum_{i=1}^n \log(\hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}) = \sum_{i=1}^n (y_i \log \hat{y}_i + (1-y_i) \log(1 - \hat{y}_i))$$

$$\begin{aligned} \text{Max } & (\sum_{i=1}^n (y_i \log \hat{y}_i + (1-y_i) \log(1 - \hat{y}_i))) \\ \text{Min } & (-\sum_{i=1}^n (y_i \log \hat{y}_i + (1-y_i) \log(1 - \hat{y}_i))) \end{aligned}$$



$$\mathcal{L}(\hat{y}, y) = - (y \log \hat{y} + (1-y) \log(1-\hat{y})) \leftarrow$$

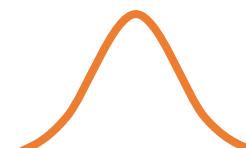
Prof. Andrew Ng, 2018

Maximum-Likelihood-Schätzung

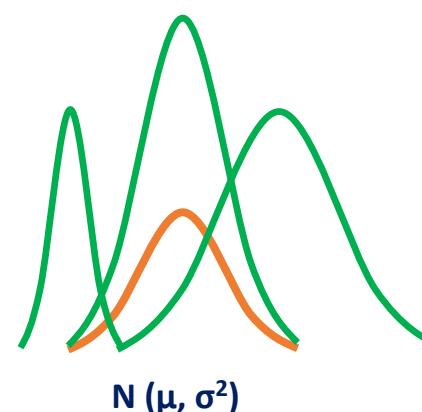
nicht in der Realität existieren



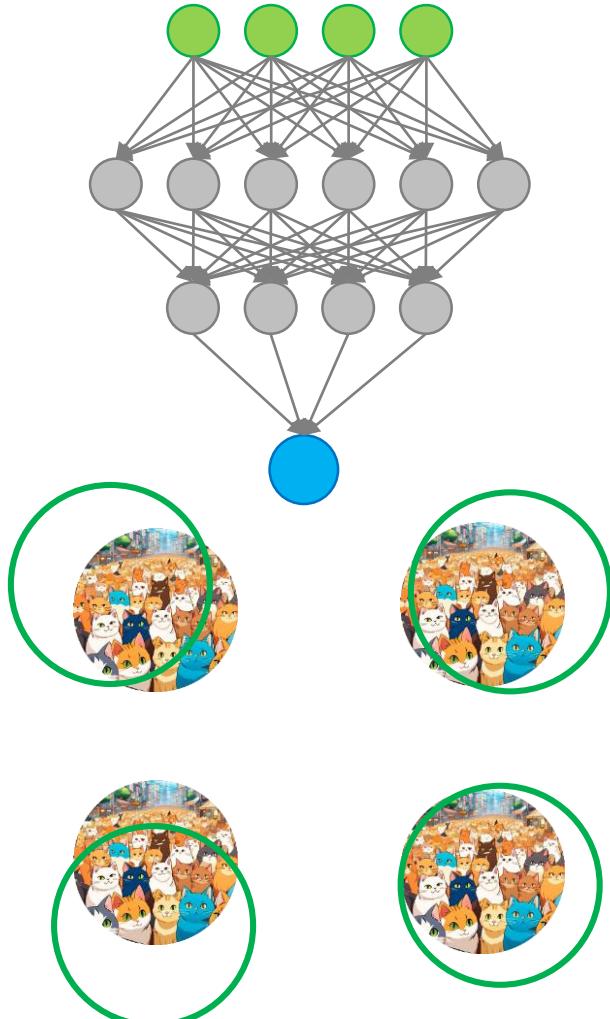
$$N(\mu, \sigma^2)$$



Annäherungsweise Darstellung des Modells



Neuronales Netzwerk



Take Home Messages

- Eine Verlustfunktion misst, wie gut ein neuronales Netzmodell eine bestimmte Aufgabe erfüllt, in den meisten Fällen eine Regression oder Klassifizierung.
- Wir müssen den Wert der Verlustfunktion während des Backpropagation-Schrittes minimieren, um das neuronale Netz zu verbessern.
- Verlustfunktion erstellen: Methode der kleinsten Quadrate, Maximum-Likelihood-Schätzung und Kreuzentropie (*nächste Folge*).

Nächste Schritte:

Kreuzentropie für Verlustfunktion

Vielen herzlichen Dank für eure Aufmerksamkeit!

