# Blatt 6

Luca Krüger, Jonas Otto, Jonas Merkle (Gruppe R)

17. Juni 2019

## 1 Cross Entropy

1. In diesem Beispiel werden beide Eingaben falsch klassifiziert, die jeweils falsche Klassifikation lässt aber keinen Rückschluss auf die tatsächliche Abweichung der Netzausgabe im Vergleich zur gewünschten Klassifikation zu. Die quadratische Fehlerfunktion ist zum Trainieren des Netzes nicht sinnvoll, da z.B. die Klassifikation von $x_2$ nicht unbedingt besser als die von $x_1$ ist. (Klasse 2 ist nicht „näher“an 1 als 2.)

3.

$$A(x_1, x_2, x_3, x_4) = \left(\tfrac{1}{8}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}\right)$$
$$B(x_1, x_2, x_3, x_4) = \left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\right)$$

a)

$$H_A(B) = \sum_{j=1}^{4} B(x_j) \log\left(\frac{1}{A(x_j)}\right) = 2.375 \text{ bits}$$

b)

$$H_B(A) = \sum_{j=1}^{4} A(x_j) \log\left(\frac{1}{B(x_j)}\right) = 2.25 \text{ bits}$$

c)

$$H_A(A) = H(A) = 1.75$$

d)

$$D_Q(P) = H_Q(P) - H(P)$$

$$\Rightarrow D_A(B) = 0.625$$
$$\Rightarrow D_B(A) = 0.5$$

4. zu I):

$$\begin{aligned} &\sum_x p(X) = \sum_x q(x) = 1 \\ \overset{\text{Gibb's inequality}}{\iff} &\sum_x p(x) \log_2(\frac{1}{p(x)}) \leq \sum_x q(x) \log_2(\frac{1}{p(x)}) \\ \iff &0 \leq \sum_x q(x) \log_2(\frac{1}{p(x)}) - \sum_x p(x) \log_2(\frac{1}{p(x)}) \\ \iff &0 \leq H_Q(P) - H(P) = d_Q(P) \quad \square \end{aligned}$$

zu II):

$$d_Q(P) = d_P(Q) \iff H_Q(P) - H(P) = H_P(Q) - H(Q)$$

$\Rightarrow$ Widerspruch mit dem obigen Beispiel:
$H(P) = H(Q)$, aber $H_Q(P) \neq H_P(Q)$

zu III):

$$d_Q(Q) = H_Q(Q) - H(Q) = H(Q) - H(Q) = 0 \quad \square$$

5. a) Ein Minimum von $D_C(B)$ wird bei $C = B$ erwartet, da dann $D_C(B) = 0$ gilt. Für $C = B$ muss $t = \frac{2}{3}$ gelten.
   b) (siehe Jupyter Notebook)

# 2 Cross Entropy als Kostenfunktion

1. a)

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{e^{cu_i}}{\sum_{j=1}^{n} e^{cu_j}} = \frac{\sum_{i=1}^{n} e^{cu_i}}{\sum_{j=1}^{n} e^{cu_j}} = 1$$

$$e^x > 0 \; \forall x \in \mathbb{R} \implies y_i \geq 0 \; \forall i$$

   b)

$$y_1 = \frac{e^{cu_1}}{e^{cu_1} + e^{cu_2} + e^{cu_3}} = \frac{1}{1 + e^{c(u_2-u_1)} + e^{c(u_3-u_1)}}$$

   c) Grenzwertbetrachtung $c \to \infty$:

| Fall I: $u_1 > u_2 > u_3$ | Fall II: $u_2 > u_1 > u_3$ | Fall III: $u_2 > u_3 > u_1$ |
|---|---|---|
| $\lim_{c\to\infty} y_1 = 1$ | $\lim_{c\to\infty} y_1 = 0$ | $\lim_{c\to\infty} y_1 = 0$ |

   d) Dämpfung der dendritischen Potentiale $u_q, \ldots, u_n$ in Abhängigkeit von $c$:

| | |
|---|---|
| $c > 0$: | Verteilung über alle $y_i$ in Abhängigkeit von $u_i$ |
| $c = 0$: | $y_1 = \cdots = y_k$ (gleichverteilt) |
| $c < 0$: | Verteilung über alle $y_i$ in negativer Abhängigkeit von $u_i$ |

2. a) Ableitungen der Fehlerfunktion nach der Netzwerkausgabe $y_i$:

$$\frac{\partial E}{\partial y_1} = -\frac{t_1}{y_1}$$
$$\frac{\partial E}{\partial y_2} = -\frac{t_2}{y_2}$$

b) Ableitungen der Netzwerkausgabe nach dem dendritischen Potential $u_2$:

$$y_1 = \frac{e^{u_1}}{e^{u_1} + e^{u_2}} \qquad y_2 = \frac{e^{u_2}}{e^{u_1} + e^{u_2}}$$

$$\begin{aligned}
\frac{\partial y_1}{\partial u_2} &= -\frac{e^{u_2}}{(e^{u_1} + e^{u_2})^2} e^{u_2} = -\frac{e^{u_1}}{e^{u_1} + e^{u_2}} \cdot \frac{e^{u_2}}{e^{u_1} + e^{u_2}} = -y_1 y_2 \\
\frac{\partial y_2}{\partial u_2} &= \frac{e^{u_2}}{e^{u_1} + e^{u_2}} - \frac{e^{u_2}}{(e^{u_1} + e^{u_2})^2} \cdot e^{u_2} = y_2 \cdot (1 - \frac{e^{u_2}}{e^{u_1} + e^{u_2}}) \\
&= y_2 \cdot (1 - y_2)
\end{aligned}$$

c) Ableitung des dentritischen Potentials nach dem Gewicht $w_2$

$$u_2 = w_2 x + b_2 \Rightarrow \frac{\partial u_2}{w_2} = x$$

d) Ableitung der Fehlerfunktion nach dem Gewicht $w_2$:

$$\begin{aligned}
\frac{\partial E}{\partial w_2} &= \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial u_2} \frac{\partial u_2}{\partial w_2} + \frac{\partial E}{\partial y_2} \frac{\partial y_2}{\partial u_2} \frac{\partial u_2}{\partial w_2} \\
&= (-\frac{t_1}{y_1}) \cdot (-y_1 y_2) x + (-\frac{t_2}{y_2}) \cdot (y_2 \cdot (1 - y_2)) x \\
&= t_1 y_2 x - t_2 x (1 - y_2) \\
&= t_1 y_2 x - t_2 x + t_2 x y_2 \\
&= (y_2 (t_1 + t_2) - t_2) x \\
&= (y_2 - t_2) x
\end{aligned}$$

e) Der Ableitungsterm bei quadratischer Fehlerfunktion enthält einen weiteren Faktor $f'(y)$, abhängig von der Übertragungsfunktion $f$.