1. (2 Punkte) Ein Netz gibt die Wahrscheinlichkeit $\hat{y}$ für die Klasse 1 aus. Berechne jeweils die Wahrscheinlichkeit p, die das Netz der richtigen Klasse y gibt, und den Fehler $L = -\ln(p)$.
   (a) $\hat{y} = 0.95$, $y = 1$
   (b) $\hat{y} = 0.6$, $y = 0$
   (c) $\hat{y} = 0.02$, $y = 1$

Lösung:
(a) $p = 0.95$, $L = -\ln(0.95) = 0.0513$
(b) $p = 0.4$, $L = -\ln(0.4) = 0.9163$
(c) $p = 0.02$, $L = -\ln(0.02) = 3.912$
Im dritten Fall ist das Netz fast sicher, dass die Klasse 0 ist, und liegt falsch. Deshalb ist der Fehler so groß.

2. (4 Punkte) Zwei Netze A und B sollen vier Bilder klassifizieren. Die richtigen Klassen und die Ausgaben $\hat{y}$ der Netze sind

Bild	1	2	3	4
richtige Klasse y	1	0	1	0
$\hat{y}$ von Netz A	0.99	0.02	0.97	0.9
$\hat{y}$ von Netz B	0.8	0.25	0.75	0.3

   (a) Bestimme für jedes Bild die Wahrscheinlichkeit p, die das Netz der richtigen Klasse gibt.
   (b) Berechne für beide Netze den mittleren Abstand $1 - p$ und das Produkt der Wahrscheinlichkeiten p.
   (c) Berechne für beide Netze den mittleren Fehler $-\ln(p)$.
   (d) Welches Netz ist besser? Warum kommt der mittlere Abstand zu einem anderen Ergebnis?

Lösung:
(a) Netz A: 0.99, 0.98, 0.97, 0.1, Netz B: 0.8, 0.75, 0.75, 0.7
(b) Abstand_A 0.24 Abstand_B 0.25
Produkt_A 0.0941 Produkt_B 0.315
(c) L_A 0.5908 L_B 0.2888
(d) Netz B: Es hat das größere Produkt und den kleineren Fehler. Es ist sich nie sehr sicher, liegt aber immer richtig. Nach dem Abstand wäre Netz A etwas besser. Der Abstand ist aber höchstens 1 und bestraft deshalb kaum, dass Netz A bei Bild 4 fast sicher und falsch ist ($p = 0.1$, Fehler 2.3026).

3. (4 Punkte) Der Fehler eines Netzes für ein Beispiel ist $L = -\big(y \ln(\hat{y}) + (1-y)\ln(1-\hat{y})\big)$.
   (a) Vereinfache die Formel für $y = 1$ und für $y = 0$. Was gibt der Fehler in Worten an?
   (b) Bei mehreren Beispielen multipliziert man die Wahrscheinlichkeiten für die richtigen Klassen. Warum?
   (c) Warum nimmt man den Logarithmus und warum das Minus davor?
   (d) Warum ist $-\ln(p)$ als Fehler besser geeignet als der Abstand $1 - p$?

Lösung:

(a) $y = 1$: $L = -\ln(\hat{y})$. $y = 0$: $L = -\ln(1-\hat{y})$. In beiden Fällen ist L der negative Logarithmus der Wahrscheinlichkeit, die das Netz der richtigen Klasse gibt.

(b) Das Produkt ist die Wahrscheinlichkeit, dass das Netz bei allen Beispielen die richtige Klasse auswürfeln würde (Pfadregel). Je größer sie ist, desto besser ist das Netz.

(c) Der Logarithmus macht aus dem Produkt eine Summe, $\ln(p_1 p_2) = \ln(p_1) + \ln(p_2)$. Damit wird das Rechnen einfacher, und die Zahlen werden nicht winzig. Weil ln monoton steigt, bleibt die Reihenfolge der Netze gleich. Wegen $p \leq 1$ ist $\ln(p) \leq 0$. Das Minus macht den Fehler positiv, aus möglichst groß wird möglichst klein.

(d) $-\ln(p)$ wächst gegen unendlich, wenn p gegen 0 geht. Eine Antwort, die sicher und falsch ist, wird stark bestraft. Der Abstand $1 - p$ ist höchstens 1, ein sicherer Fehler fällt bei vielen Beispielen kaum auf.

4. (3 Punkte) Ein Neuron mit einem Input hat $z = w \cdot x + b$ und $\hat{y} = \sigma(z)$. Für das Trainingsbeispiel $x = 1$, $y = 1$ und $b = 0$ ist der Fehler $L(w) = -\ln(\sigma(w))$. Am Anfang ist $w = 0$.
   (a) Berechne $L(0)$, $L(0.01)$ und $L(-0.01)$.
   (b) Berechne die numerische Ableitung $\dfrac{L(0.01) - L(-0.01)}{0.02}$. Soll w größer oder kleiner werden?
   (c) Mache einen Schritt des Gradientenabstiegs mit $lr = 1$ und berechne den neuen Fehler.

Lösung:

(a) $L(0)$ 0.6931 $L(0.01)$ 0.6882 $L(-0.01)$ 0.6982

(b) Ableitung -0.5 Sie ist negativ, also soll w größer werden.

(c) w^{neu} 0.5 $L(w^{neu})$ 0.4741

Der Fehler ist von 0.6931 auf 0.4741 gesunken.

5. (4 Punkte) Die Funktion $f(x) = (x-3)^2 + 1$ hat die Ableitung $f'(x) = 2(x-3)$. Der Gradientenabstieg startet bei $x = 0$ und rechnet $x_{neu} = x - lr \cdot f'(x)$.
   (a) Berechne drei Schritte mit $lr = 0.25$.
   (b) Berechne zwei Schritte mit $lr = 1$ und zwei Schritte mit $lr = 1.25$.
   (c) Beschreibe, was in den drei Fällen passiert.

Lösung:

(a) x_1 1.5 x_2 2.25 x_3 2.625

(b) $lr = 1$: x_1 6 x_2 0 $lr = 1.25$: x_1 7.5 x_2 −3.75

(c) Mit $lr = 0.25$ halbiert sich der Abstand zum Minimum $x = 3$ in jedem Schritt. Mit $lr = 1$ springt x zwischen 0 und 6 hin und her und kommt nie an. Mit $lr = 1.25$ werden die Sprünge immer größer, der Abstand zum Minimum wächst.

6. (3 Punkte) Gegeben ist $f(x_1, x_2) = x_1^2 + x_2^2$ an der Stelle $(2, 1)$.
   (a) Berechne die partiellen Ableitungen und den Gradienten an dieser Stelle. Welcher Input hat dort den größeren Einfluss auf f?
   (b) Gehe mit $lr = 0.25$ einen Schritt gegen den Gradienten. Wie groß ist f vorher und nachher?

Lösung:

(a) $\frac{\partial f}{\partial x_1} = 2x_1 = 4$, $\frac{\partial f}{\partial x_2} = 2x_2 = 2$, Gradient $(4, 2)$. Eine kleine Änderung von x_1 ändert f doppelt so stark wie dieselbe Änderung von x_2.

(b) x_1^{neu} 1 x_2^{neu} 0.5 f^{vorher} 5 $f^{nachher}$ 1.25

7. (3 Punkte) Gegeben ist $f(a, b) = a^2 + 2b^2$. Der Gradientenabstieg startet bei $(a, b) = (2, 1)$ mit $lr = 0.1$.
   (a) Bestimme die partiellen Ableitungen $\frac{\partial f}{\partial a}$ und $\frac{\partial f}{\partial b}$.
   (b) Berechne zwei Schritte und jeweils den Funktionswert.

Lösung:

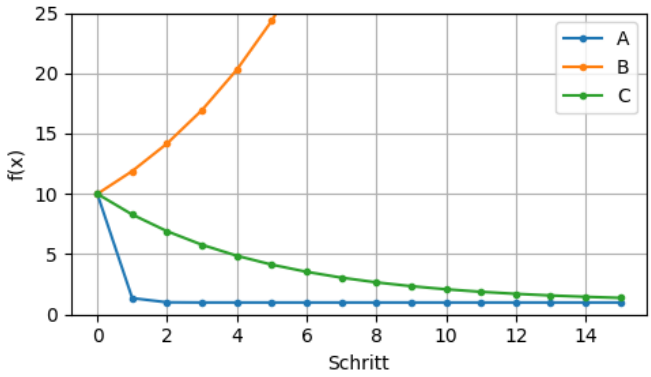
(a) $\frac{\partial f}{\partial a} = 2a$, $\frac{\partial f}{\partial b} = 4b$

(b) Gradient bei $(2, 1)$: $(4, 4)$.

a_1	1.6	a_2	1.28
b_1	0.6	b_2	0.36
f_1	3.28	f_2	1.8976

Der Funktionswert sinkt von 6 über 3.28 auf 1.8976.

8. (3 Punkte) Für $f(x) = (x-3)^2 + 1$ wurde der Gradientenabstieg von $x = 0$ aus mit drei Lernraten durchgeführt: 0.05, 0.4 und 1.05. Das Bild zeigt den Funktionswert nach jedem Schritt.



(a) Ordne die Kurven A, B und C den Lernraten zu und begründe.

(b) Was ist beim Training eines neuronalen Netzes die Folge einer zu kleinen und einer zu großen Lernrate?

Lösung:

(a) A: 0.4, der Wert sinkt schnell auf das Minimum 1. B: 1.05, jeder Schritt springt über das Minimum hinaus, und zwar immer weiter, der Wert steigt. C: 0.05, kleine Schritte, der Wert sinkt langsam.

(b) Zu klein: Das Training dauert sehr lange. Zu groß: Der Fehler springt hin und her oder wächst, das Netz lernt nicht oder nur schlecht.

9. (4 Punkte) (a) Was gibt die Ableitung des Fehlers nach einem Parameter an?

(b) Was ist der Gradient und in welche Richtung zeigt er?

(c) Was ist eine Epoche?

(d) Was ist ein lokales Minimum und warum ist es ein Problem für den Gradientenabstieg?

Lösung:

(a) Wie stark und in welche Richtung sich der Fehler ändert, wenn man den Parameter ein wenig vergrößert. Ist sie negativ, wird der Fehler kleiner, wenn der Parameter größer wird.

(b) Die Liste aller partiellen Ableitungen der Funktion nach ihren Parametern. Er zeigt in die Richtung des steilsten Anstiegs, der Gradientenabstieg geht in die Gegenrichtung.

(c) Ein Durchgang durch alle Trainingsbeispiele mit anschließendem Update der Parameter.

(d) Eine Stelle, an der die Funktion kleiner ist als in der Umgebung, aber nicht am kleinsten überhaupt. Der Gradientenabstieg geht immer bergab und bleibt dort hängen, weil die Ableitung 0 ist.

10. (4 Punkte) Schreibe eine Funktion `abstieg(f, x, lr, anzahl)`, die für eine beliebige Funktion `f` von der Stelle `x` aus `anzahl` Schritte des Gradientenabstiegs macht und das Ende zurückgibt. Die Ableitung soll numerisch berechnet werden. Teste sie mit $g(x) = x^4 - 3x^2 + x$, $lr = 0.01$ und 200 Schritten, einmal ab $x = 2$ und einmal ab $x = -2$. Was fällt auf?

Lösung:

```
def abstieg(f, x, lr, anzahl):
    for i in range(anzahl):
        ableitung = (f(x + 0.0001) - f(x - 0.0001)) / 0.0002
        x = x - lr * ableitung
    return x

def g(x):
    return x ** 4 - 3 * x ** 2 + x

for start in [2, -2]:
    x = abstieg(g, start, 0.01, 200)
    print(start, round(x, 4), round(g(x), 4))
```

Ergebnis: Ab 2 endet der Abstieg bei $x = 1.1309$ mit $g = -1.0702$, ab -2 bei $x = -1.3008$ mit $g = -3.5139$. Je nach Start landet man in einem anderen Tal. Das rechte ist nur ein lokales Minimum.