Universität Salzburg		Florian Graf
	Machine Learning	
Übungsblatt **8**		22 Punkte

Aufgabe 1. *XOR-Problem* 22 P.

Gegeben seien zwei-dimensionale binäre Daten $\{(0,0)^\top, (0,1)^\top, (1,0)^\top, (1,1)^\top\} \subset \{0,1\} \times \{0,1\}$ deren Klassen durch die XOR-Funktion bestimmt sind, also $y = \mathrm{XOR}(\mathbf{x})$, wobei

$$\begin{aligned} \mathrm{XOR} : \{0,1\} \times \{0,1\} &\to \{0,1\} \\ (a,b) &\mapsto \begin{cases} 1 & \text{falls } a \neq b \\ 0 & \text{falls } a = b \end{cases} \quad . \end{aligned}$$

Wir trainieren auf den Daten Regressionsmodelle f_θ, deren Parameter θ wir durch Minimierung des MSE (als Lossfunktion) bestimmen, d.h., $\mathcal{L}(\theta) = \frac{1}{4}\sum_{i=1}^4 \left(f_\theta(\mathbf{x}_i) - XOR(\mathbf{x}_i)\right)^2$

(a) Wir wählen zunächst ein lineares Modell $f_\theta(\mathbf{x}) = \mathbf{x}^\top\mathbf{w} + b$, wobei $\mathbf{w} \in \mathbb{R}^2$ und $b \in \mathbb{R}$. Zeigen Sie, dass $\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b) > 0$, und somit, dass das lineare Modell die Trainingsdaten nicht fitten kann.
Hinweis: Dies ist ein lineares Regressionsproblem und Sie dürfen bekannte Resultate zur Berechnung von $(\hat{\mathbf{w}}, \hat{b}) = \operatorname{argmin} \mathcal{L}(\mathbf{w}, b)$ verwenden.

(b) Wir wählen nun ein MLP mit zwei Layern und einer ReLU Aktivierungsfunktion, sodass

$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \operatorname{ReLU}\left(\mathbf{W}^\top\mathbf{x} + \mathbf{c}\right) + b \quad .$$

Eine mögliche Lösung des XOR-Problems ist gegeben durch

$$\begin{aligned} \mathbf{W} &= \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad , & \mathbf{c} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \mathbf{w}^\top &= \begin{pmatrix} 1 & 1 \end{pmatrix} \quad , & b &= 0 \quad . \end{aligned}$$

Verfolgen Sie, wie die Daten durch das Netzwerk transformiert werden, indem Sie folgende Zwischenergebnisse berechnen.

- Die Datenmatrix $\mathbf{X} \in \mathbb{R}^{4\times 2}$.
- Den Output des ersten Layers $\mathbf{XW}$.
- Den Input des zweiten Layers $\operatorname{ReLU}(\mathbf{XW})$.
- Den Output des Modells.

Zeichnen Sie außerdem handschriftlich in jedem Schritt (außer dem letzten) die Zeilenvektoren des jeweiligen Zwischenergebnisses in ein Koordinatensystem ein und kennzeichnen Sie die zugehörigen Klassen. Zeichnen Sie außerdem in die Zeichnung zu $\operatorname{ReLU}(\mathbf{XW})$ den Vektor $\mathbf{w}$ und die 'Entscheidungsgrenze' $\{\mathbf{v} \in \mathbb{R}^2 : \mathbf{w}^\top\mathbf{v} = 1/2\}$ ein.

(c) Wir ersetzen nun die ReLU-Aktivierungsfunktion durch eine tanh-Aktivierungsfunktion. Ist das Modell in der Lage die Daten zu fitten,

(i) …falls $\mathbf{W}, \mathbf{c}, \mathbf{w}, b$ wie in Aufgabenteil (b) sind?
(ii) …falls $\mathbf{W}, \mathbf{c}$ wie in Aufgabenteil (b) sind, und $\mathbf{w}, b$ gelernt werden?
(iii) …falls alle Modellparameter $\mathbf{W}, \mathbf{c}, \mathbf{w}, b$ gelernt werden?
(iv) …falls alle Modellparameter $\mathbf{W}, \mathbf{c}, \mathbf{w}, b$ gelernt werden und die Aktivierungsfunktion entfernt, bzw. durch die Identität ersetzt wird.

Begründen Sie Ihre Antworten.