Einführung in Deep Learning

Blatt 2 - Abgabe am Mi, 23.10.2024 um 12:00

Geben Sie Ihre Lösung als `blatt2.pdf` ab. Eine Möglichkeit ist das einscannen einer handschriftlichen Abgabe, eine andere Möglichkeit das Nutzen eines Jupyter Notebooks (oder LaTeX) mit PDF Export.

In diesem Blatt üben wir die Grundlagen, die notwendig sind, um das klassische Perzeptron (mit Updateregel) als Spezialfall des verallgemeinerten Perzeptron (mit allgemeiner Updateregel, die sich aus einer Loss-Funktion ergibt) aufzufassen. Da das 'Lernen' ein Gradientenabstieg ist, müssen wir Gradienten, d.h. Ableitungen, tatsächlich berechnen können.

Bemerkung. *Es gibt Funktionen $f\colon \mathbb{R} \to \mathbb{R}$, die nicht überall differenzierbar sind (d.h. die an manchen Stellen keine Ableitung haben). Ein klassisches Beispiel einer solchen Funktion ist der Absolutbetrag $f(x) := |x|$, denn der Funktionsgraph hat einen 'Knick' bei $x = 0$, sodass man an der Stelle keine 'beste' lineare Näherung an den Funktionsgraph anlegen kann (keine Tangente). In der ϵ-δ-Charakterisierung der Ableitung ist die Ableitung an der Stelle $x = 0$ als Grenzwert $\lim_{h\to 0} \frac{f(h)}{h}$ definiert, der nicht existiert. Es existieren die einseitigen Grenzwerte $\lim_{0>h\to 0} \frac{f(h)}{h} = -1$ und $\lim_{0<h\to 0} \frac{f(h)}{h} = +1$, die aber verschieden sind.*

Die Ableitung f' einer Funktion $f\colon \mathbb{R} \to \mathbb{R}$ hat die Eigenschaft, dass $\int_{-\infty}^{c} f'(x)dx = f(c)$ ist. So ein Zusammenhang lässt sich auch für die Betragsfunktion herstellen.

Definition. *Wir nennen eine Funktion $g\colon \mathbb{R} \to \mathbb{R}$ eine* schwache Ableitung *von f, wenn $\int_{-\infty}^{c} g(x)dx = f(c)$ gilt.*

Die Betragsfunktion hat die Vorzeichenfunktion $\operatorname{sign}\colon \mathbb{R} \to \mathbb{R}$ mit $\operatorname{sign}(0) = 0$ und $\operatorname{sign}(x) = \frac{x}{|x|}$ sonst als schwache Ableitung.

Aufgabe 1. *Bestimmen Sie eine schwache Ableitung $g\colon \mathbb{R} \to \mathbb{R}$ für die* Maximums-Funktion $m(x) := \max(0, x) = \begin{cases} 0, & x \leq 0 \\ x, & x \geq 0. \end{cases}$

Tipp: Überlegen Sie, dass die schwache Ableitung mit einer normalen Ableitung übereinstimmen kann, an den x, an denen $m(x)$ in einem Intervall um x differenzierbar ist. Sie müssen keine Integrale berechnen.

Aufgabe 2. *Sei $x \in \mathbb{R}^n$ und $s\colon \mathbb{R}^n \to \mathbb{R}$ gegeben als $s(w) := w \cdot x$ (Skalarprodukt). Bestimmen Sie die Ableitungen von s nach den einzelnen w_j, d.h. $\frac{d}{dw_j} s(w)$ für $j = 1, \ldots, n$.*

Tipp: Überlegen Sie sich erst die Fälle $n = 1$ und $n = 2$.

Aufgabe 3. *Seien $y \in \{0, 1\}$ und $x \in \mathbb{R}^n$ fest. Bestimmen Sie mit Hilfe der Kettenregel $\frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$ eine schwache Ableitung nach w_j für die Funktion $L_{x,y}\colon \mathbb{R}^n \to \mathbb{R}$ mit der Definition $w \mapsto L_{x,y}(w) := \max(0, (1 - 2y)(w \cdot x))$.*

Aufgabe 4. *Gegeben sei ein Datenpunkt $x \in \mathbb{R}^2$ mit Label $y = 1$ und ein Perzeptron mit Gewichtsvektor $w^{(0)} \in \mathbb{R}^2$ sodass $w^{(0)} \cdot x < 0$ ist, also x fälschlicherweise das Label $[w^{(0)} \cdot x > 0] = 0$ zugewiesen bekommt. Die Update-Regel definiert einen neuen Gewichtsvektor $w^{(k+1)} := w^{(k)} + r\left(y - [w^{(k)} \cdot x > 0]\right) x$ (in Abhängigkeit von einer Lernrate r, z.B. $r = 0.001$). Überlegen Sie, warum das gilt: wenn man oft genug ein Update durchführt, dann wird x irgendwann richtig klassifiziert (mit $w^{(k)}$ für großes k).*

Tipp: Berechnen Sie $w^{(1)} \cdot x$ mit Hilfe der Bilinearität des Skalarprodukts.