Übungsblatt 4: Stochastik und Neuronale Netzwerke
BA-INF 153: Einführung in Deep Learning für Visual Computing

Deadline:	29.05.2024 - 14:00 via eCampus	
Tutoren:	Alina Pollehn	s6aapoll@uni-bonn.de
	Johannes van de Locht	s6jovand@uni-bonn.de
Übungsgruppenleitung:	Jan Müller	muellerj@cs.uni-bonn.de

Einleitung

In der ersten Vorlesung "Einführung in künstliche Neuronale Netze" haben wir eine der einfachsten Formen eines Neuronalen Netzwerks (NN) kennen gelernt; das Multilayer Perceptron (MLP). Das MLP wurde als Implementierung eines biologischen Neuronalen Netzwerks motiviert, in dem jede Schicht ein Aktivierungspotenzial $z_i^{(l)}$ als gewichtet Summe berechnet und eine Aktivierungsfunktion ϕ verwendet um den Threshold-Mechanismus eines Neurons zu modellieren d.h.

$$z_i^{(l)} = \phi\left(\sum_{j=1}^{n_l} W_{i,j}^{(l)} z_j^{l-1} + b^{(l)}\right).$$

Auf diesem Übungsblatt wollen wir aber die biologische Motivation und die genauen Details eines MLPs erst einmal außer Acht lassen. Stattdessen betrachten wir ein NN als eine **parametrisierte Funktion** $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$ **die einen Vektor auf einen anderen Vektor abbildet** und dessen Verhalten durch die Parameter $\theta = \{W_{ij}^l\}_{ijl} \cup \{b^l\}_l$ kontrolliert wird.

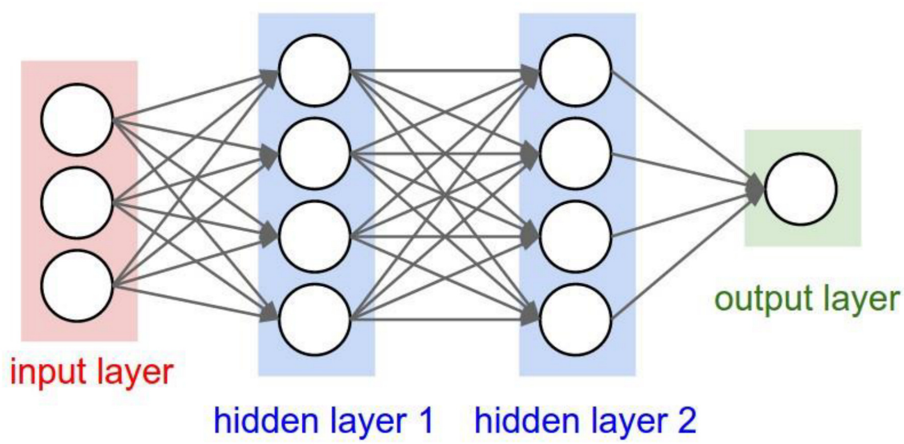

Figure 1: Abbildung eines Multilayer Perceptron (MLP).

Wird für das letzte Layer des MLPs die korrekte Aktivierungsfunktion gewählt (dazu mehr in einer kommenden Vorlesung), kann unser NN eingeschränkt werden nur Funktionen $f_\theta : \mathbb{R}^n \to [0,1]^m$ mit $\sum_{i=1}^m f(x)_i = 1$ (Summe über alle Outputs) darzustellen. In diesem Fall können wir ein solches Neuronales Netzwerk auch als bedingte Wahrscheinlichkeitsfunktion $p_\theta(Y \mid M)$ über m Label interpretieren. Die Wahrscheinlichkeit $p(Y = i \mid X = x)$ ist dann gegeben durch $f_\theta(x)_i$. Wir sehen also dass es eine direkte Verbindung zwischen Neuronalen Netzen und Stochastik gibt. Das Ziel dieses Übungsblatts ist es, die Zusammenhänge zwischen den Grundlagen des Machine Learnings bzw. der Wahrscheinlichkeitstheorie und Neuronalen Netzwerken weiter zu vertiefen.

Theoretische Aufgaben (*15 Punkte*)

a) Bias/Varianz Trade-off und Generalisierung ***(6 Punkte)*** Der Mean squared error (MSE) eines Punktschätzers wird in der Vorlesung definiert als

$$\mathrm{MSE}(\hat{\theta}_m) := \mathbb{E}\left[\left(\hat{\theta}_m - \theta\right)^2\right]$$

und stellt ein gutes Maß für die Generalizisierungsfähigkeit eines Schätzers dar. Diese Eigenschaft wird damit begründet, dass MSE ein Tradeoff zwischen der erwarteten Abweichung zum tatsächlichen Wert und der Streuung der geschätzten Werte ist. Wir trainieren unser Neuronales Netzwerk f_θ auf den Trainingsdaten $x_1, ..., x_m$. Den Trainingsprozess unseres Neuronalen Netzwerks können wir als Punktschätzer $\hat{\theta}_m$ für den optimalen (aber unbekannten) Parameter θ unseres Neuronalen Netzwerks f_θ auffassen. Nach dem Trainingsprozess quantifizieren wir den Erfolg indem wir unser Performance-Maß auf Testdaten auswerten, die nicht Bestandteil der Trainingsdaten gewesen sind. Wir beobachten dass der Trainingsprozess nicht auf die Testdaten generalisiert.

Aufgabenstellung:

1. ***(3 Punkte)*** Zeigen Sie, dass der MSE tatsächlich einen Tradeoff zwischen dem Bias und der Varianz des Schätzer darstellt d.h.

   $$\mathrm{MSE}(\hat{\theta}_m) = \mathrm{Bias}(\hat{\theta}_m)^2 + Var(\hat{\theta}_m).$$

   Hinweis: Der Erwartungswert einer reellwertigen Konstante c ist $E[c] = c$, damit gilt auch $\mathbb{E}[\mathbb{E}[\hat{\theta}]] = \mathbb{E}[\hat{\theta}]$; der Sollwert θ ist eine Konstante. Verwenden Sie die Definition der Varianz.

2. ***(3 Punkte)*** Benennen Sie zwei Gründe aus der Vorlesung die erklären warum unser Netzwerk nach dem Training nicht gut auf die Testdaten generalisiert und erklären Sie wie sich der Bias und die Varianz unseres Trainingsprozesses in den beiden Fällen verhält. Nehmen Sie in Ihrer Erklärung Bezug auf den Hyperparameter der in der Vorlesung in diesem Zusammenhang besonders herausgestellt wird.

b) MAP für bedingte Wahrscheinlichkeiten *(5 Punkte)* Der Maximum A Posteriori (MAP) Schätzer für einen Parameter $\boldsymbol{\theta}$ mit gegebener Observationen $\boldsymbol{X}$ ist definiert als

$$\boldsymbol{\theta}_{\text{MAP}} := \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\boldsymbol{\theta}|\boldsymbol{X}).$$

In der Vorlesung wird, unter der Annahme von i.d.d. Stichproben $\{x^{(i)}\}_{i\in[1:m]}$, aus diesem MAP Schätzer der folgende Punktschätzer hergeleitet:

$$\boldsymbol{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \left[\sum_{i=1}^{m} \log p(x^{(i)}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right].$$

Wir betrachten nun einen MAP Schätzer bei dem zwei Observationen $(\boldsymbol{X}, \boldsymbol{Y})$ mit i.d.d. Stichproben $(x^{(i)}, y^{(i)})$ gegeben sind.

Aufgabenstellung: Zeigen Sie, dass der Schätzer

$$\boldsymbol{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \sum_{i=1}^{n} \log P(y^{(i)} \mid x^{(i)}, \theta) + \log P(\theta).$$

ein Punktschätzer auf den Stichproben $(x^{(i)}, y^{(i)})$ für den MAP-Schätzer

$$\boldsymbol{\theta}_{\text{MAP}} := \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})$$

ist. Sie dürfen dabei Annehmen, dass die Observationen X, Y unabhängig von den Parametern unsere Modells θ sind.

Hinweis: Verwenden Sie die Definition der bedingten Wahrscheinlichkeit für $P(\theta, X, Y)$.

c) Schätzer und Regularisierung *(4 Punkte)* In Vorlesung wird darauf hingewiesen, dass sich der Mean Squared Error (MSE) aus einem Maximum-Likelihood (ML) Schätzer ableiten lässt wenn angenommen wird, dass die Daten normalverteilt sind. Nehmen wir an, dass wir i.i.d. Stichproben $\{(x^{(i), y^{(i)}})\}_{i\in[1:m]}$ haben und dass unsere Daten durch ein Gaußmodell $p(y \mid x, \theta) = \mathcal{N}(y \mid f(x;\theta), \sigma^2)$ beschrieben werden, wobei σ^2 bekannt ist.

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m} \|f(x^{(i)};\theta) - y^{(i)}\|^2}_{\text{mean square error}} + \underbrace{\frac{1}{2}\|\theta\|_2^2}_{L_2-\text{reg.}}$$

Zeigen Sie, dass die Minimierung des MSE mit L_2-Regelmäßigkeit äquivalent zur maximalen a-posteriori-Schätzung mit einem Gauß-Prior für die Gewichte ist.

Praktische Aufgaben (***15 Punkte***)

Alle notwendigen Erläuterungen und die Aufgabenstellungen finden Sie in dem Juptyer Notebook "Übungsblatt04.ipynb" im Zusatzmaterial "Übungsblatt04.zip" zu dieser Übung.