1 Fehlerrückführung

1.1 Problemstellung und Notation

Die Trainingsdaten sind gegeben als

Eingangsmuster: $x_e^{1,\mu}$.

Sollantworten: s_n^μ.

μ bezeichnet ein Trainingspaar, ν bezeichnet eine Schicht im Netz und $e, ..., i, j, k, ..., n$ sind Indizes für einzelne Zellen in einer Schicht. Für die Eingangsschicht ist hier $\nu = 1$ gesetzt.

Wir betrachten jetzt nur ein Muster und lassen der Übersichtlichkeit halber den Index μ weg. Für eine Zelle j in der Schicht ν berechnen sich die Aktivitäten z_j^ν und Ausgänge y_j^ν gegeben die Eingänge x_i^ν und Gewichte w_{ji}^ν wie folgt:

$$z_j^\nu = \sum_i w_{ji}^\nu x_i^\nu \,, \tag{1}$$

$$y_j^\nu = \sigma(z_j^\nu) \,, \tag{2}$$

$$x_j^{\nu+1} = y_j^\nu \,. \tag{3}$$

Die Ausgänge dienen als Eingänge für die nachfolgende Schicht.

Den Fehler, den das Netz macht, können wir als mittlere quadratische Abweichung der Netzausgänge $y_n^{N,\mu}$ (N bezeichnet die Ausgangsschicht) von den Sollantworten s_n^μ definieren:

$$F = \frac{1}{2} \sum_\mu \sum_n (s_n^\mu - y_n^{N,\mu})^2 \,. \tag{4}$$

Dies ist jedoch nur eine von vielen möglichen Definitionen.

1.2 Gradientenabstiegsverfahren

Um den Fehler zu minimieren, können die Gewichte w_{ji}^ν schrittweise so verändert werden, dass F abnimmt. Die notwendigen Veränderungen Δw_{ji}^ν können durch partielle Ableitung von F nach w_{ji}^ν bestimmt werden:

$$\Delta w_{ji}^\nu = -\eta \frac{\partial F}{\partial w_{ji}^\nu} \,. \tag{5}$$

Wenn man die partiellen Ableitungen direkt bestimmt, dann hat das verschiedene Nachteile:

- teuer für große Netze mit vielen Schichten
- nicht lokal, d.h. um ein Gewicht zu verändern, muss Information über das ganze Netz vorliegen

Daher verwendet man eine lokale, effizientere Methode, die Fehlerrückführung.

1.2.1 Fehlerrückführung (backpropagation of error)

Wir betrachten wieder nur ein Muster und lassen den Index μ weg.

Für eine Zelle j in der Schicht ν ergibt die Kettenregel

$$\frac{\partial F}{\partial w_{ji}^\nu} = \frac{\partial F}{\partial z_j^\nu} \frac{\partial z_j^\nu}{\partial w_{ji}^\nu} \,,$$

da w_{ji}^ν direkt nur z_j^ν beeinflusst.

Wir definieren

$$\delta_j^\nu := \frac{\partial F}{\partial z_j^\nu} \,.$$

Wegen $z_j^\nu = \sum_i w_{ji}^\nu x_i^\nu$ gilt

$$\frac{\partial z_j^\nu}{\partial w_{ji}^\nu} = x_i^\nu \,.$$

Damit ist

$$\frac{\partial F}{\partial w_{ji}^\nu} = \delta_j^\nu x_i^\nu \,. \tag{6}$$

Die x_i^ν können leicht aus (1–3) iterativ von der Eingansschicht zur Ausgangsschicht hin berechnet werden. Es müssen also nur noch die δ_j^ν bestimmt werden.

Aus der Definition des Fehlers (Gl. 4) ergibt sich für die Ausgangsschicht

$$\delta_n^N := \frac{\partial F}{\partial z_n^N} = \frac{\partial F}{\partial y_n^N}\frac{\mathrm{d}y_n^N}{\mathrm{d}z_n^N} = -(s_n - y_n^N)\sigma'(z_n^N) \,. \tag{7}$$

Dies muss für jedes Muster einzeln ausgerechnet werden. Mit Index μ schreibt man $\delta_n^{N,\mu} = -(s_n^\mu - y_n^{N,\mu})\,\sigma'(z_n^{N,\mu})$.

Ist der Fehler nicht als mittlere quadratische Abweichung definiert, ändert sich der erste Faktor $\frac{\partial F}{\partial y_n^N}$ entsprechend.

Für die verborgenen Schichten gilt

$$\delta_j^\nu := \frac{\partial F}{\partial z_j^\nu} \quad = \quad \sum_k \frac{\partial F}{\partial z_k^{\nu+1}}\frac{\partial z_k^{\nu+1}}{\partial z_j^\nu}$$

NR:

$$\begin{aligned} z_k^{\nu+1} &= \sum_j w_{kj}^{\nu+1} x_j^{\nu+1} = \sum_j w_{kj}^{\nu+1}\sigma(z_j^\nu) \\ \frac{\partial z_k^{\nu+1}}{\partial z_j^\nu} &= w_{kj}^{\nu+1}\,\sigma'(z_j^\nu) \end{aligned}$$

$$\begin{aligned} \delta_j^\nu &= \sum_k \delta_k^{\nu+1}\, w_{kj}^{\nu+1}\,\sigma'(z_j^\nu) \\ &= \sigma'(z_j^\nu)\sum_k \delta_k^{\nu+1}\, w_{kj}^{\nu+1} \end{aligned} \tag{8}$$

Mit den Gleichungen (7, 8) können die δ_j^ν leicht iterativ von der Ausgangsschicht zur Eingansschicht hin berechnet werden.

Wenn die x_i^ν und δ_j^ν berechnet sind, kann mit Gleichung (6) der Gradient und mit Gleichung (5) die Änderung der Gewichte berechnet werden.

Die Änderung der Gewichte wird für jedes Trainingspaar neu durchgeführt. Mehrfache Iteration durch alle Trainingsmuster, oder einfache Iteration durch eine sehr lange Reihe von Trainingspaaren führt in der Regel zum Lernen der gesuchten Funktion.