三層パーセプトロンについて

erutuf

2015 年 12 月 30 日

1 Introduction

三層パーセプトロンの定義と基本的な性質について述べていく.

1.1 Definitions

L, M, $H \in \mathbb{N}$ に対し，パラメータ空間を

$$\Theta_H = \mathbb{R}^{M(H+1)+H(L+1)} = \{\theta = (w_{11}, \ldots, w_{MH}, \eta_1, \ldots, \eta_M, u_{11}, \ldots, u_{HL}, \zeta_1, \ldots, \zeta_H)\}$$

と定める．また，$\mathbf{u}_j = (u_{j1}, u_{j2}, \ldots, u_{jL})^T$ とおく.

定義 1.1. $\rho : \mathbb{R} \to \mathbb{R}$ を C^1 関数とする．ユニット数 $H \in \mathbb{N}$，活性化関数 ρ の三層フィードフォワードネットワークとは $\theta \in \Theta_H$ をパラメータに持つ以下の関数 $f^H(\cdot \mid \theta) = (f_1^H(\cdot \mid \theta), \ldots, f_M^H(\cdot \mid \theta)) : \mathbb{R}^L \to \mathbb{R}^M$ のことを言う：

$$f_i^H(x \mid \theta) = \sum_{j=1}^{H} w_{ij} \rho\left(u_{jk} x_k + \zeta_j\right) + \eta_i, \quad (1 \leq i \leq M).$$

また，活性化関数としてシグモイド関数 $\rho(x) = \sigma(x) = \frac{1}{1+e^{-t}}$ を用いた三層フィードフォワードネエットワークを三層パーセプトロンネットワークと呼ぶ.

ここで，以下の定理により，三層パーセプトロンネットワークはコンパクト台を持つ連続写像を近似するのに十分な写像であると言える ([1])：

定理 1.2. 三層パーセプトロンネットワークの集合

$$\left\{\left(f_1^H(x \mid \theta), \ldots, f_M^H(x \mid \theta)\right) \mid \theta \in \Theta_H,\ H \in \mathbb{N}\right\}$$

はコンパクト台を持つ連続写像の空間 $C_c(\mathbb{R}^L; \mathbb{R}^M)$ の中で sup ノルムで稠密.

機械学習において教師データから目標とする関数を求める手法として誤差伝播法がある．これは，教師データが与えられたとき，ユニット数 H を選びパラメータ θ をランダムに初期化して得られる三層パーセプトロン $f^H(\cdot \mid \theta)$ から出発し，パラメータ θ を逐次的に更新することで $f^H(\cdot \mid \theta)$ を目標とする関数に近似させる．ここでユニット数 H は初めに自分で選ばなければいけないことに注意が必要である．H の値は小さすぎると目標とする関数を近似するのに不十分であり，また大きすぎると過学習の問題が生じてしまう．よって H は

どの程度まで小さく取ってよいかという議論が重要になる．そのため，三層パーセプトロンのユニット数が極小であるという概念を以下で定義する：

定義 1.3. $f^H(\cdot \mid \theta)$ を三層パーセプトロンとする．$f^H(\cdot \mid \theta)$ が極小であるとは，ユニット数が H より少なく関数として等しい別の三層パーセプトロンが存在しないときに言う．すなわち，$H' < H$ のとき任意の三層パーセプトロン $f^{H'}(\cdot \mid \omega)$, $(\omega \in \Theta_{H'})$ に対して，ある点 $x \in \mathbb{R}^L$ で $f^H(x \mid \theta) \neq f^{H'}(x \mid \omega)$ をみたすときに言う．

極小でない三層パーセプトロンの重要な例を考えてみよう．

例 1. パラメータ $\theta = (w_{11}, \ldots, w_{MH}, \eta_1, \ldots, \eta_M, u_{11}, \ldots, u_{HL}, \zeta_1, \ldots, \zeta_H) \in \Theta_H$ が $w_{1H} = \cdots = w_{MH} = 0$ をみたしていたとする．このときユニット数 H，パラメータ θ を持つ三層パーセプトロン $f^H(\cdot \mid \theta)$ は

$$f_i^H(x \mid \theta) = \sum_{j=1}^{H-1} w_{ij} \rho \left(\sum_{k=1}^{L} u_{jk} x_k + \zeta_j \right) + \eta_i$$

となり，$w_{1H}, \ldots, w_{MH}, u_{H1}, \ldots, u_{HL}, \zeta_H$ は不要なパラメータであったことが分かる．したがって $\theta' = (w_{11}, \ldots, w_{M(H-1)}, \eta_1, \ldots, \eta_M, u_{11}, \ldots, u_{(H-1)L}, \zeta_1, \ldots, \zeta_{H-1})$ とおけば

$$f^H(\cdot \mid \theta) = f^{H-1}(\cdot \mid \theta')$$

となり，$f^H(\cdot \mid \theta)$ は極小でないことが分かった．

例 2. パラメータ $\theta = (w_{11}, \ldots, w_{MH}, \eta_1, \ldots, \eta_M, u_{11}, \ldots, u_{HL}, \zeta_1, \ldots, \zeta_H) \in \Theta_H$ が $u_{H1} = \cdots = u_{HL} = 0$ をみたしていたとする．このとき三層パーセプトロン $f^H(\cdot \mid \theta)$ は

$$f_i^H(x \mid \theta) = \sum_{j=1}^{H-1} w_{ij} \rho \left(\sum_{k=1}^{L} u_{jk} x_k + \zeta_j \right) + w_{iH} \rho(\zeta_H) + \eta_i$$

となり，これは $\eta_i' = w_{iH} \rho(\zeta_j) + \eta_i$，$\theta' = (w_{11}, \ldots, w_{M(H-1)}, \eta_1', \ldots, \eta_M', u_{11}, \ldots, u_{(H-1)L}, \zeta_1, \ldots, \zeta_{H-1})$ と取れば

$$f^H(\cdot \mid \theta) = f^{H-1}(\cdot \mid \theta')$$

をみたす．

例 3. パラメータ θ が，$u_{(H-1)1} = u_{H1}, \cdots, u_{(H-1)L} = u_{HL}$, $\zeta_{H-1} = \zeta_H$ をみたすとする．このとき三層パーセプトロン $f^H(\cdot \mid \theta)$ は

$$f_i^H(x \mid \theta) = \sum_{j=1}^{H-2} w_{ij} \rho \left(\sum_{k=1}^{L} u_{jk} x_k + \zeta_j \right) + (w_{i(H-1)} + w_{iH}) \rho \left(\sum_{k=1}^{L} u_{(H-1)k} x_k + \zeta_{H-1} \right) + \eta_i$$

となり，$w_{i(H-1)}' = w_{i(H-1)} + w_{iH}$ とおけば新しいパラメータ $\theta' \in \Theta_{H'}$ を作ることができ，

$$f^H(\cdot \mid \theta) = f^{H-1}(\cdot \mid \theta')$$

をみたす．

また，パラメータ θ が $u_{(H-1)1} = -u_{H1}, \cdots, u_{(H-1)L} = -u_{HL}, \zeta_{H-1} = -\zeta_H$ をみたすとする．このとき三層パーセプトロン $f^H(\cdot \mid \theta)$ は

$$f_i^H(x \mid \theta) = \sum_{j=1}^{H-2} w_{ij}\rho\left(\sum_{k=1}^{L} u_{jk}x_k + \zeta_j\right) + (w_{i(H-1)} - w_{iH} + 1)\rho\left(\sum_{k=1}^{L} u_{(H-1)k}x_k + \zeta_{H-1}\right) + \eta_i$$

となり (ここで $\sigma(-x) = -\sigma(x) + 1$ を利用した)，$w'_{i(H-1)} = w_{i(H-1)} - w_{iH} + 1$ とおけば新しいパラメータ $\theta' \in \Theta_{H'}$ を作ることができ，

$$f^H(\cdot \mid \theta) = f^{H-1}(\cdot \mid \theta')$$

をみたす．

上の三つの例の中の $f^H(\cdot \mid \theta)$ はどれもユニット数が一つ少ない $f^{H-1}(\cdot \mid \theta')$ に簡約できることが分かった．そこで，上の例をふまえて三層パーセプトロンが既約 (それ以上簡約できない) ことを以下で定義する：

定義 1.4. 三層パーセプトロン $f^H(\cdot \mid \theta)$ が既約であるとは以下の三つの条件をみたすときに言う：

1. 各 $1 \leq j \leq H$ で $(w_{1j}, \ldots, w_{Mj}) \neq 0$
2. 各 $1 \leq j \leq H$ で $(u_{j1}, \ldots, u_{jL}) \neq 0$
3. 各 $1 \leq j_1, j_2 \leq H,\ j_1 \neq j_2$ で $(u_{j_1 1}, \ldots, u_{j_1 L}, \zeta_{j_1}) \neq \pm(u_{j_2 1}, \ldots, u_{j_2 L}, \zeta_{j_2})$.

このように三層パーセプトロンに極小と既約という二つの概念が定義された．これらはどちらも三層パーセプトロンがそれ以上小さくできないということを表現したものであった．実は，この二つの条件は互いに同値であることが示される：

定理 1.5. 三層パーセプトロンが極小であることと既約であることは同値.

Proof. (TODO). □

1.2 Fisher Information

三層パーセプトロンに対する確率密度関数を定義し，さらにその Fisher 情報行列について見ていく．

定義 1.6. $f(\cdot \mid \theta)$ をフィードフォワードネットワークとする．q を $\mathbb{R}^L$ 上の確率密度関数とし，V を $M \times M$ 次の正定値対称行列とする．このとき $f(\cdot \mid \theta)$ に対する確率密度関数を

$$p(x, y \mid \theta) = \frac{1}{(2\pi)^{M/2} \det V^{1/2}} \exp\left\{-\frac{1}{2}(y - f(x \mid \theta))^T V^{-1}(y - f(x \mid \theta))\right\} q(x)$$

と定義する．

一般に，C^1 級のパラメータ付き確率密度関数 $p(z \mid \theta),\ z \in \mathbb{R}^n,\ \theta \in \Theta$ に対してその Fisher 情報行列 $I(\theta) = (I_{ab}(\theta))_{ab}$ は

$$I_{ab}(\theta) = \int_{\mathbb{R}^n} \frac{\partial \log p(z \mid \theta)}{\partial \theta_a} \frac{\partial \log p(z \mid \theta)}{\partial \theta_b} p(z \mid \theta) dx$$

で定義される．Fisher 情報行列は一般に半正定値であるが，正定値であるとは限らない．

次の節では以下の定理を示すことが目標とする：

定理 1.7. $\mathbb{R}^L$ 上の確率密度関数 q が正かつ連続であるとする．このとき三層パーセプトロン $f^H(\cdot \mid \theta)$ について，$f^H(\cdot \mid \theta)$ が極小であることは $f^H(\cdot \mid \theta)$ の確率密度関数の Fisher 情報行列が正定値であることに同値である．

この定理は，三層パーセプトロンが極小であることが Fisher 情報行列という解析的な量から決まることを意味し，ユニット数決定の議論において非常に重要な事実となる．

参考文献

[1] G. Cybenco. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems, 2(4):3030314, 1989.

[2] K. Fukumizu. ”A regularity condition of the information matrix of a multilayer perceptron network.” Neural networks 9.5 (1996): 871-879.