5章　ニューラルネットワーク

これまでの議論では、固定された基底関数の線形和で表されるモデルについて見てきた。基底関数を増やすほどモデルの表現能力は上がるが、基底関数を増やすとパラメータ数も増える（次元の呪い）問題や、適切な基底関数を選択しなければならない問題が生じる。

ニューラルネットワークは、固定された多数の基底関数を用いる代わりに、パラメトリックに決定する少数の基底関数を用いる手法の一つである。事前に基底関数の数を固定し、基底関数のパラメータを訓練中に決定する。

本章の構成は次の通り。
[5.1] まずニューラルネットワークの関数の形から見ていき、どのような点で優れているのかを示す。
[5.2] パラメータを学習する方法を示す。これには誤差逆伝播法なるものが用いられる。
[5.3] 正則化の話やその他の拡張について議論する。
[5.4] ベイズ理論の視点から見たニューラルネットワークを議論する。

5.1 フィードフォワードネットワーク関数

ニューラルネットワークは基底関数をパラメトリックに扱う方法であると述べた。この説では、ニューラルネットワークの最も基本形であるフィードフォワードネットワークについて、基底関数がどのように表現されるか、モデル全体がどのように表現されるか、また、なぜこれがどのように優れているのかについて解説する。

基底関数がどのように表現されるか

予測モデル：$y(x, w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right)$ (5.1)

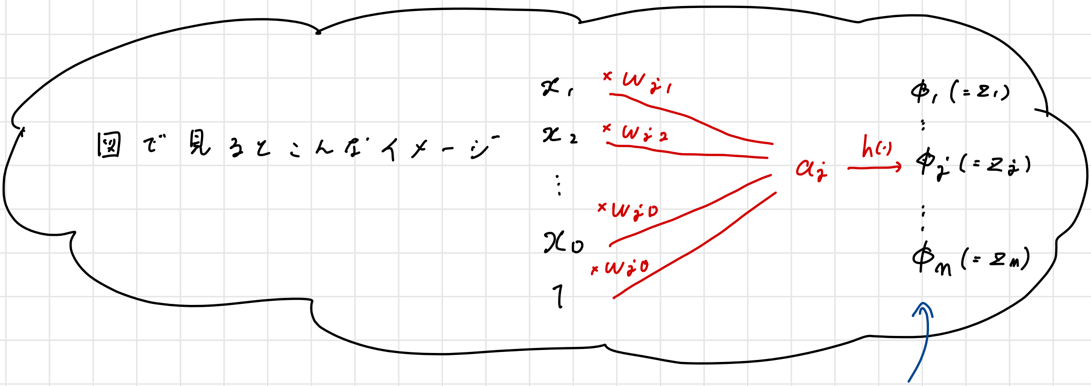
- $f(\cdot)$ は $\begin{cases} \text{分類問題のとき：非線形関数} \\ \text{回帰問題のとき：恒等写像} \end{cases}$
- $\phi_j(x)$ は基底関数

  ↳ これをパラメトリックに表現する。

$$\phi_j(x) = z_j$$

$$= \underbrace{h}_{\text{なんらかの非線形関数 (e.g. ロジスティックシグモイド)}}(a_j)$$

$$= h\left(\sum_{i=1}^{D} \underbrace{w_{ji}^{(1)}}_{x_i \text{が} \phi_j \text{に与える影響の重み。}} x_i + w_{j0}^{(1)}\right) \quad ①$$

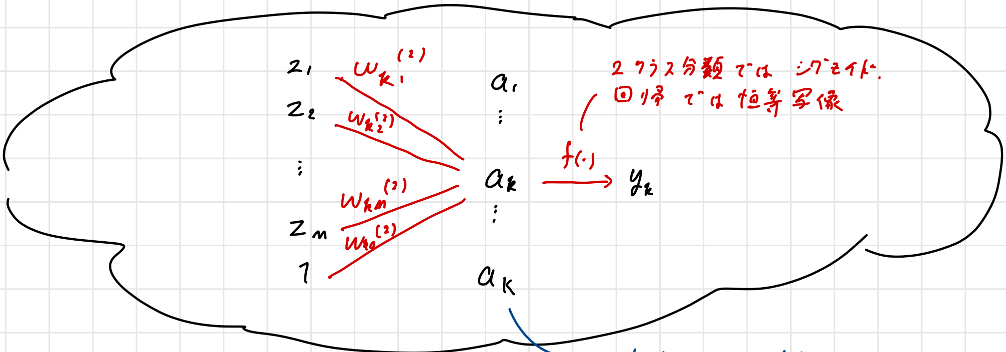


隠れユニットと呼ばれる。

出力の総数がKとすると、k番目の出力は次のようになる。

$$y_k(x, w) = f(a_k)$$

$$= f\left(\sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \right) \quad (2)$$



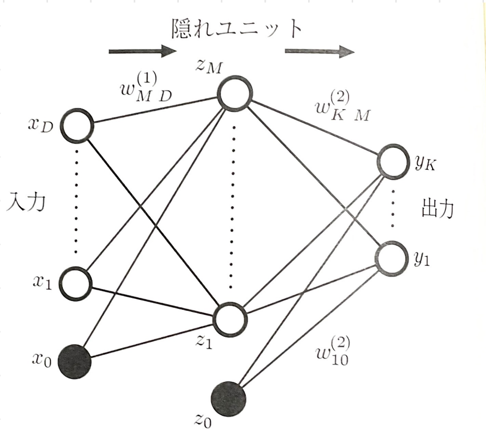
出力ユニット活性
と呼ばれる。

モデル全体がどのように表現されるか

①.② をまとめると、モデル全体は次の式になる。

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{5.7}$$

この関数をネットワーク図で表現すると右図のようになる。値を入力側から計算していく過程は順伝搬と呼ばれる。ニューラルネットワークは多層パーセプトロン（MLP）とも呼ばれるが、ステップ関数を用いていない点で厳密にはパーセプトロンではない点に注意。



上で扱ったネットワークを、本書では2層ネットワークと呼ぶ。これは重みパラメータを持つ層が2つあることからこのような名称としている。
また、本書ではネットワーク上に閉じた有向閉路を持たないネットワークについてのみ扱うこととする。このような性質を持つネットワークをフィードフォワードネットワークという。

【感想】
隠れ1層ネットワークという呼ばれ方がそこそこ一般的な印象である。
というのも、2層ネットワークと言われると、全体で2層なのか、隠れ2層なのか結局曖昧になるからである。
ユニットが3列並んでいることから、間違えて3層と呼ばないように注意

このモデルがどのように優れているのか

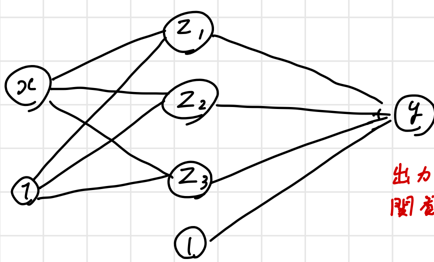
フィードフォワードネットワークは、多様な関数を近似できることが知られており、万能近似器と呼ばれている。これの実例を見てみる。

3個の隠れユニットを持つ2層ネットワークを使って、下記関数のデータにフィットさせる。

- $f(x) = x^2$
- $f(x) = \sin(x)$
- $f(x) = |x|$
- $f(x) = H(x)$

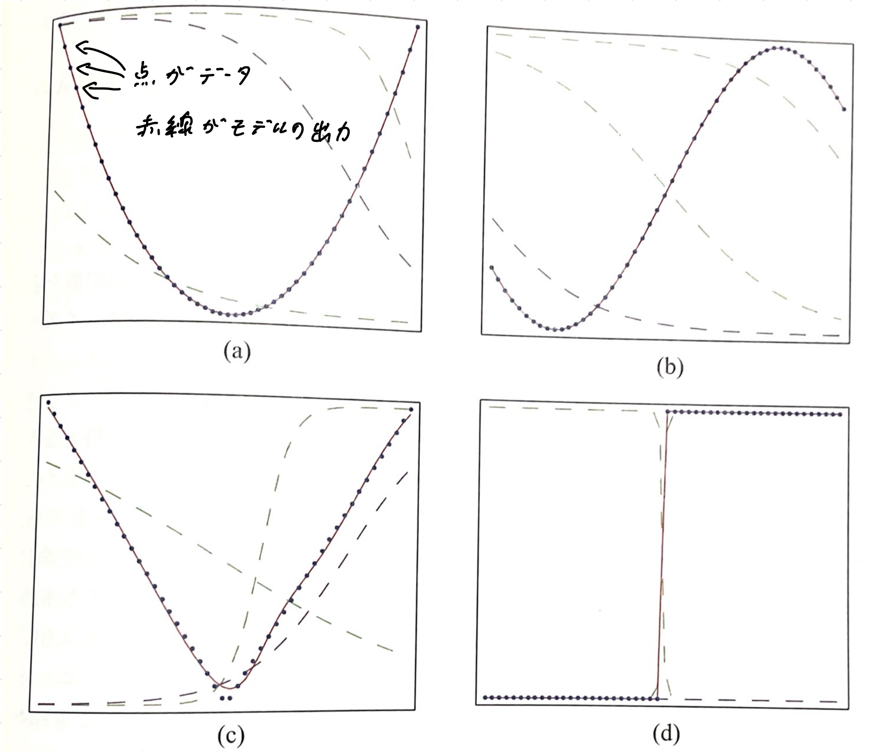
ヘヴィサイドステップ関数

ネットワークモデルは、右図の通り。



出力ユニット活性化関数は、恒等写像。

隠れユニット活性化関数は、tanh



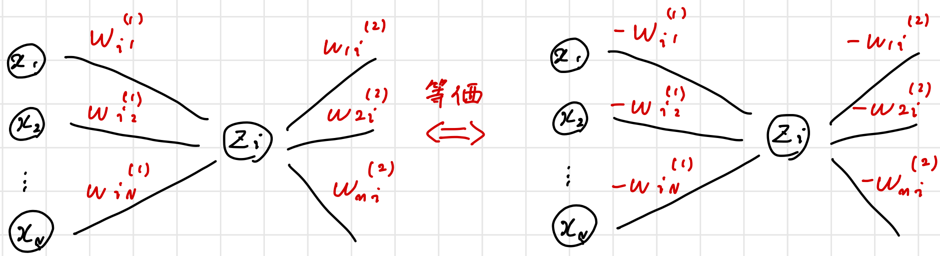
3つの破線は、隠れユニットの出力を示す。
すなわち、

$$z_1 = \tanh(w_{11}x + w_{10})$$
$$z_2 = \tanh(w_{21}x + w_{20})$$
$$z_3 = \tanh(w_{31}x + w_{30})$$

5.1.1 重み空間対称性

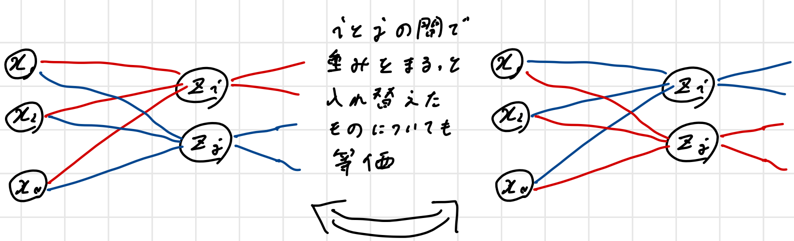
フィードフォワードネットワークでは、任意の入力に対して全く同じ出力をするが、重みパラメータが異なるようなケースが存在する。これは後にニューラルネットワークのモデルエビデンスを評価する際に注意しなければいけない。
ここではひとまず、等価な重みベクトルが存在することのみを示す。

隠れ層の活性化関数がシグモイド関数や、 tanh のように奇関数である場合を考える。隠れユニット z_i について、等価な出力をするものが二つ存在する。



したがって、M個の隠れ層について、2^M 個の等価な重みベクトルが存在する。

また、隠れユニットの順序を入れ替えたものについても等価な出力となる。



したがって、$M!$ 個の等価な重みベクトルが存在する。
全組み合わせを考えると、$M!\,2^M$ 個の等価な重みベクトルが存在する。

5.2 ネットワーク訓練

本節以降では、重みパラメータを実際に決定する手順を学んでいく。とはいえ、本節にぶら下がる 5.2.x の各項ではどのようにパラメータを決定すれば良いかの指針のみを示し、具体的な手順は 5.3 節以降で議論する。

5.2 では、最小化するべき誤差関数について議論するが、この内容はこれまでの回帰問題と分類問題にほぼ重複する内容のため、おさらい程度にまとめることとする。

【回帰問題の場合】

誤差項にガウス分布を仮定することで、尤度最大化の枠組みから二乗和誤差を導出できる。

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 \qquad (5.14)$$

ガウスノイズの分散を β^{-1} とすると、これの最尤推定量も次の式より求まる。

$$\beta_{ML}^{-1} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 \qquad (5.15)$$

【2クラス分類の場合】

データが得られた時のクラスの確率をベルヌーイ分布と仮定すると、交差エントロピー誤差関数を導出できる。

$$E(w) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\} \qquad (5.21)$$

この時、出力の活性化関数にはシグモイド関数が使われる。

【多クラス分類の場合】

2クラス同様に、多クラスの交差エントロピー誤差関数が導かれる。

$$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(x_n, w)$$

この時、出力の活性化関数には、ソフトマックス関数が使われる。