機械学習（4回目）

創域理工学部　情報計算科学科
桂田　浩一

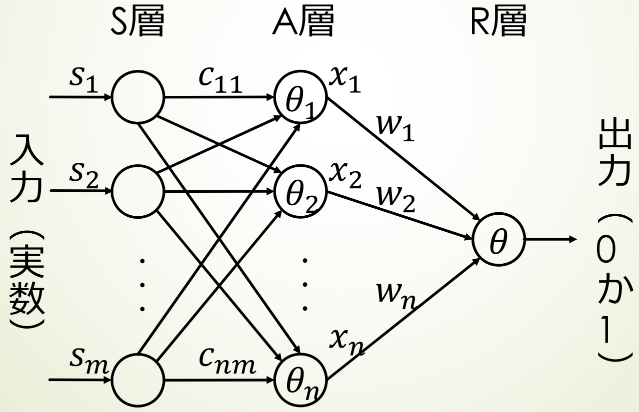
1

10/5/2023

2

前回の復習

- 単純パーセプトロンの学習可能性と収束定理

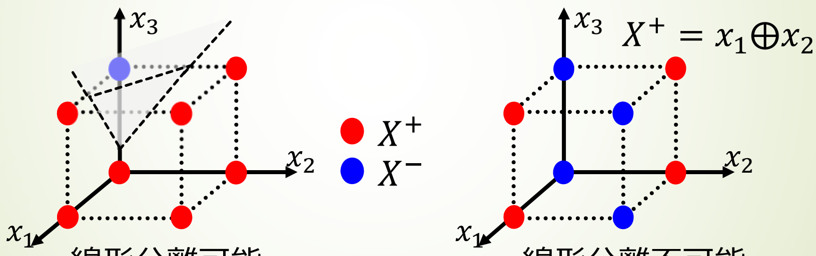


10/5/2023

3

単純パーセプトロンの問題点

- 線形分離可能な問題しか学習できない
  - XORが学習できない
    （⇒層を重ねるとXORも学習可能になるので多層化が必要）



線形分離可能　　線形分離不可能

- 入出力が0と1だけだと用途が限られる
  （⇒連続値を扱えるようにした方が良い）

10/5/2023

4

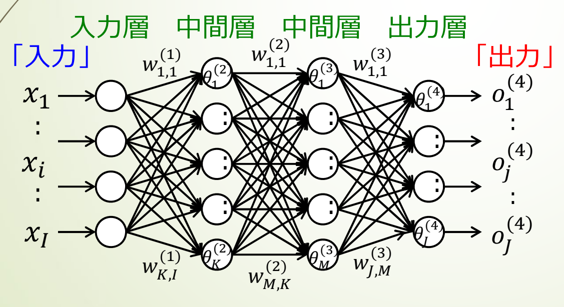
本日の内容

- 多層ニューラルネットワーク
  - 多層ニューラルネットワークとは？
  - 勾配降下法
  - 各層における微分値の求め方

10/5/2023

多層ニューラルネットワーク

- 多層ニューラルネットワーク
  - 3層以上の層（入力層＋出力層＋中間層を1層以上）を持つニューラルネットワーク
  - 出力値は連続値



第 l 層の出力 $o_n^{(l)}$ $(l \geq 2)$

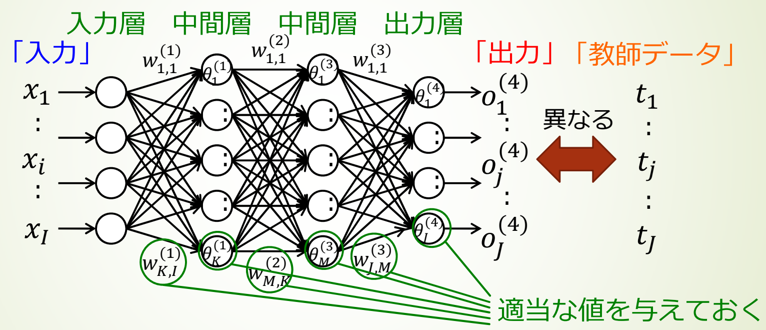
$$o_n^{(l)} = f\left(\sum_m o_m^{(l-1)} w_{n,m}^{(l-1)} - \theta_n^{(l)}\right)$$

f：活性化関数（微分可能な関数） ポイント

$o_i^{(1)} = x_i$

多層ニューラルネットワークの学習

- 多層ニューラルネットワークの学習過程



$o_j^{(L)}$ を t_j に近付ける（$o_j^{(L)}$ と t_j の差を小さくする）

$\boxed{E = \sum_{j=1}^{J}\left(t_j - o_j^{(L)}\right)^2}$ をできるだけ小さくする.

誤差関数（一例）

多層ニューラルネットワークの学習

活性化関数 f がシグモイド関数 $f(x) = \frac{1}{1+e^{-\beta x}}$ の場合,

層 l の出力 $o_j^{(l)} = \frac{1}{1+e^{-\beta\left(\sum_i o_i^{(l-1)} w_{j,i}^{(l-1)} - \theta_j^{(l)}\right)}}$ となる.

したがって，３層のネットワークの場合

$$E = \sum_{j=1}^{J}\left(t_j - \frac{1}{1+e^{-\beta\left(\sum_{m=1}^{M}\left(\frac{1}{1+e^{-\beta\left(\sum_{k=1}^{K}\left(\frac{1}{1+e^{-\beta(\cdots)}}\right)w_{m,k}^{(2)} - \theta_m^{(3)}\right)}}\right)w_{j,m}^{(3)} - \theta_j^{(4)}\right)}}\right)^2$$

この辺に $w_{1,1}^{(1)}$ が出てくる

⇩

このまま E を最小にする $w_{1,1}^{(1)}$ を求めるのは困難

多層ニューラルネットワークの学習

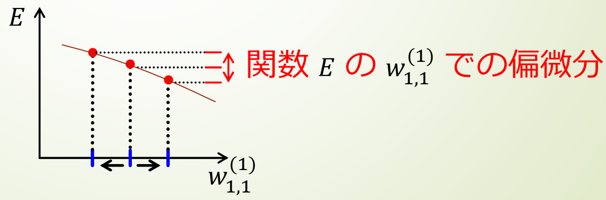
$$E = \sum_{j=1}^{J}\left(t_j - \frac{1}{1+e^{-\beta\left(\sum_{m=1}^{M}\left(\frac{1}{1+e^{-\beta\left(\sum_{k=1}^{K}\left(\frac{1}{1+e^{-\beta(\cdots)}}\right)w_{m,k}^{(2)} - \theta_l^{(3)}\right)}}\right)w_{j,m}^{(3)} - \theta_j^{(4)}\right)}}\right)^2$$

この辺に $w_{1,1}^{(1)}$ が出てくる

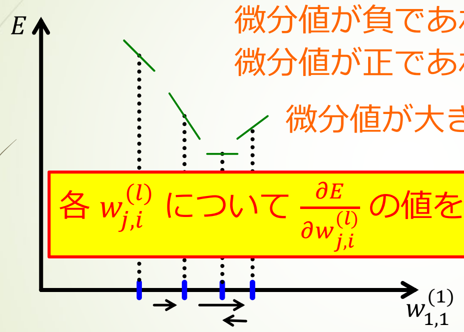
⇩

このまま E を最小にする $w_{1,1}^{(1)}$ を求めるのは困難

幸いなことに $w_{1,1}^{(1)}$ を少し増減した時の E の傾きは求まる



勾配降下法



微分値が負であれば $w_{1,1}^{(1)}$ を正の方向に，
微分値が正であれば $w_{1,1}^{(1)}$ を負の方向に動かす

微分値が大きければ大きく動かす

各 $w_{j,i}^{(l)}$ について $\frac{\partial E}{\partial w_{j,i}^{(l)}}$ の値を求める必要がある

$$w_{j,i}^{(l)} \leftarrow w_{j,i}^{(l)} - \eta \frac{\partial E}{\partial w_{j,i}^{(l)}}$$

これを全ての $w_{j,i}^{(l)}$ に対して繰り返し行えば，E を小さくする
（「教師データ」に近い値を出力する）ような $w_{j,i}^{(l)}$ が求まる

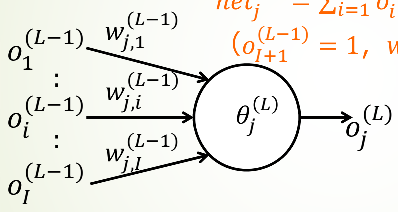
誤差逆伝播法（１）

- 誤差逆伝播法
  - 上位層から下位層に向かって（ニューラルネットの入出力と逆方向に）
    誤差関数の重み，閾値に対する偏微分と更新値を求めていく方法
- これからの説明でのネットワーク設定
  - 誤差関数： 二乗誤差（$E = \frac{1}{2}\sum_{j=1}^{J}\left(t_j - o_j^{(L)}\right)^2$）
  - 活性化関数： $f(x)$
  - 層の数： L
  - 各ノードの出力： $o_j^{(l)}$（第 l 層の j 番目のノードの出力）
  - 結合荷重： $w_{j,i}^{(l-1)}$
    （第 $l-1$ 層の i 番目のノードと第 l 層の j 番目のノードの間のリンク）

誤差逆伝播法（２）

- 出力層（第 L 層）とその一つ前の層（第 $L-1$ 層）の間の重み $w_{j,i}^{(L-1)}$ について

$$net_j^{(L)} = \sum_{i=1}^{I+1} o_i^{(L-1)} w_{j,i}^{(L-1)}$$
$$(o_{I+1}^{(L-1)} = 1,\ w_{j,I+1}^{(L-1)} = -\theta_j^{(L)})$$
$$o_j^{(L)} = f\left(net_j^{(L)}\right)$$



t_j：教師データ

$$E = \frac{1}{2}\sum_{j=1}^{J}\left(t_j - o_j^{(L)}\right)^2$$

$$\frac{\partial E}{\partial w_{j,i}^{(L-1)}} = \frac{\partial E}{\partial net_j^{(L)}} \underbrace{\boxed{\frac{\partial net_j^{(L)}}{\partial w_{j,i}^{(L-1)}}}}_{= o_i^{(L-1)}}$$

←チェーンルール
（ノード j の $net_j^{(L)}$ 以外が $w_{j,i}^{(L-1)}$ に依存しないことに注意！！）

誤差逆伝播法（３）

- 出力層（第 L 層）とその一つ前の層（第 $L-1$ 層）の間の重み $w_{j,i}^{(L-1)}$ について

$$net_j^{(L)} = \sum_{i=1}^{I+1} o_i^{(L-1)} w_{j,i}^{(L-1)}$$
$$(o_{I+1}^{(L-1)} = 1,\ w_{j,I+1}^{(L-1)} = -\theta_j^{(L)})$$
$$o_j^{(L)} = f\left(net_j^{(L)}\right) \qquad E = \frac{1}{2}\sum_{j=1}^{J}\left(t_j - o_j^{(L)}\right)^2$$

$$\frac{\partial E}{\partial w_{j,i}^{(L-1)}} = \boxed{\frac{\partial E}{\partial net_j^{(L)}}}\ \boxed{\frac{\partial net_j^{(L)}}{\partial w_{j,i}^{(L-1)}}} = o_i^{(L-1)}$$

$$\boxed{\frac{\partial E}{\partial net_j^{(L)}}} = \underbrace{\boxed{\frac{\partial E}{\partial o_j^{(L)}}}}_{= -\left(t_j - o_j^{(L)}\right)}\ \boxed{\frac{\partial o_j^{(L)}}{\partial net_j^{(L)}}} = f'\left(net_j^{(L)}\right)$$

誤差逆伝播法（4）

- 出力層（第 L 層）とその一つ前の層（第 $L-1$ 層）の間の重み $w_{j,i}^{(L-1)}$ について

$$\frac{\partial E}{\partial w_{j,i}^{(L-1)}} = \boxed{\frac{\partial E}{\partial net_j^{(L)}}}\boxed{\frac{\partial net_j^{(L)}}{\partial w_{j,i}^{(L-1)}}} \quad o_i^{(L-1)}$$

$$\boxed{\frac{\partial E}{\partial net_j^{(L)}}} = \boxed{\frac{\partial E}{\partial o_j^{(L)}}}\boxed{\frac{\partial o_j^{(L)}}{\partial net_j^{(L)}}} \quad f'\left(net_j^{(L)}\right) \quad -\left(t_j - o_j^{(L)}\right)$$

$$= -f'\left(net_j^{(L)}\right)\left(t_j - o_j^{(L)}\right)$$

ここで $f'\left(net_j^{(L)}\right)\left(t_j - o_j^{(L)}\right) = \delta_j^{(L)}$ と置くと $\boxed{\frac{\partial E}{\partial net_j^{(L)}}} = -\delta_j^{(L)}$

$$\frac{\partial E}{\partial w_{j,i}^{(L-1)}} = -\delta_j^{(L)} o_i^{(L-1)}$$

14

出題予定の演習課題

- 多層ニューラルネットワークについて

10/5/2023