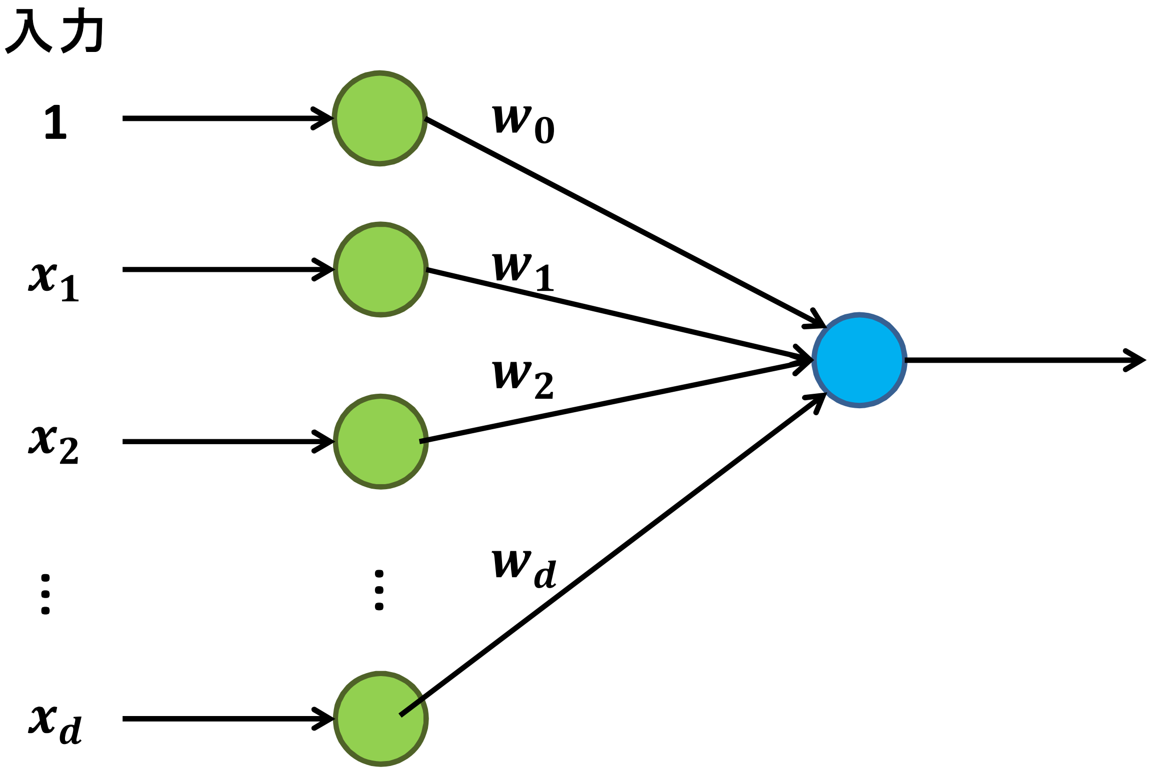
第7章 パーセプトロン型学習規則

12T4069L
佐鳥 恭太郎

パーセプトロン



出力

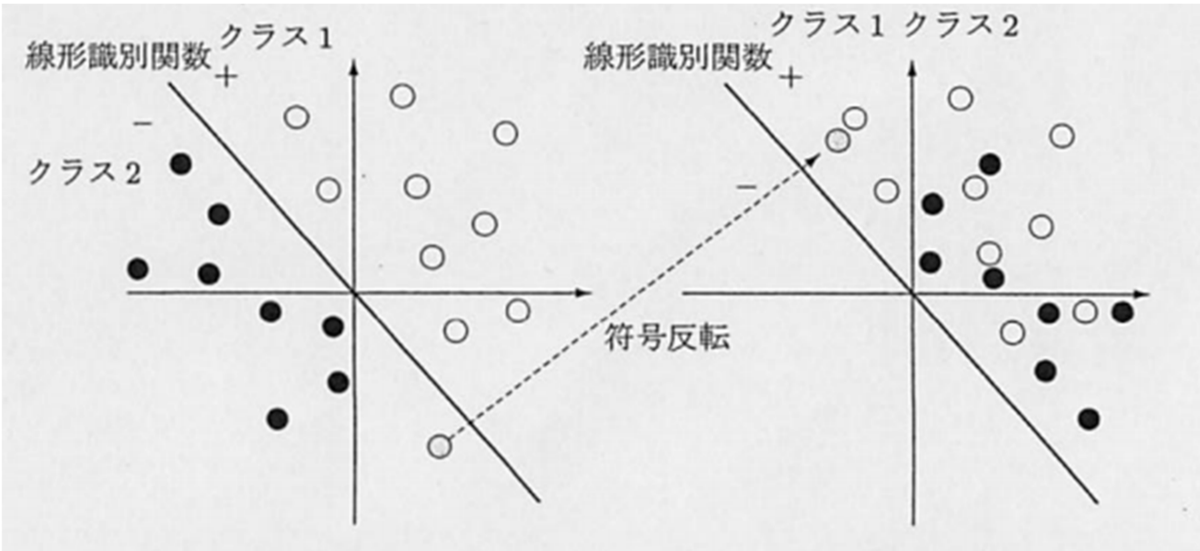
$$f(x) = w^T x$$

$$\begin{cases} f(x) \geq 0 : \; x \in C_1 \\ f(x) < 0 : x \in C_2 \end{cases}$$

$$x = (1, x_1, x_2, \dots, x_d)$$
$$w = (w_0, w_1, \dots, w_d)$$

パーセプトロンの学習規則

- データは線形分離可能であるとする。
- 以下の様にデータの符号を反転させると、全てのデータが同じ側の超平面に存在するようになる。



（a）2クラスの学習データ　（b）符号反転

学習規則の収束

- 学習データxが入力されたとき、$f(x)$の正負によってwを更新する。

$$\begin{cases} \boldsymbol{w_{new}} = \boldsymbol{w_{old}} \ (\boldsymbol{f}(\boldsymbol{x}) \geq \boldsymbol{0}) \\ \boldsymbol{w_{new}} = \boldsymbol{w_{old}} + \boldsymbol{\eta x} \ (\boldsymbol{f}(\boldsymbol{x}) < \boldsymbol{0}) \end{cases}$$

判断が正しい=>何もしない

判断が誤っている=>修正する

η : 学習の収束速度
特にη=1のとき、固定増分誤り訂正法という。

学習の難しさの尺度

- ノイズのない学習データで学習した識別関数は、ノイズありのテストデータで誤りやすくなる。
- 学習データが識別超平面に値$h(>0$, マージンという$)$より近い距離にあれば誤りとして、wを更新する。
- i番目の学習におけるw_iの変更量 Δw_iは、符号反転を行った学習データについて

$$\Delta w_i = \begin{cases} \eta x_i \ (h > w_i^T x_i / \|w_i\|) & \text{近いので修正} \\ 0 \ (\text{それ以外の場合}) & \text{遠いのでそのまま} \end{cases}$$

と書ける。

最大マージン

- 符号反転を行った学習データを超平面の法線ベクトル上へ射影した長さの最小値がマージンとなり、wが最大（クラスが分けられた）のとき、最大マージンとなる。

最大マージンD_{max}

$$D_{max} = \max_{w} D(w) = \max_{\mathrm{w}} \min_{x \in C_1, C_2} \frac{w^T x}{\|w\|}$$

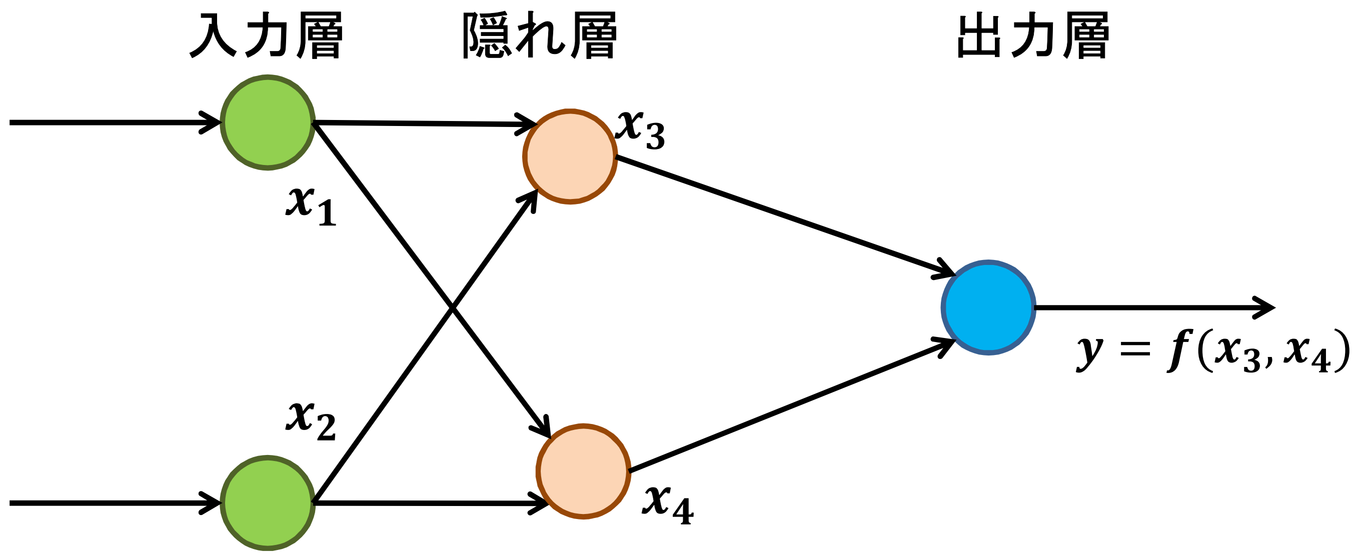
パーセプトロンの収束定理

- パーセプトロンの収束定理
  - 2クラス間の学習データが線形分離可能ならば、パーセプトロンの学習規則は有限の学習回数で収束する。

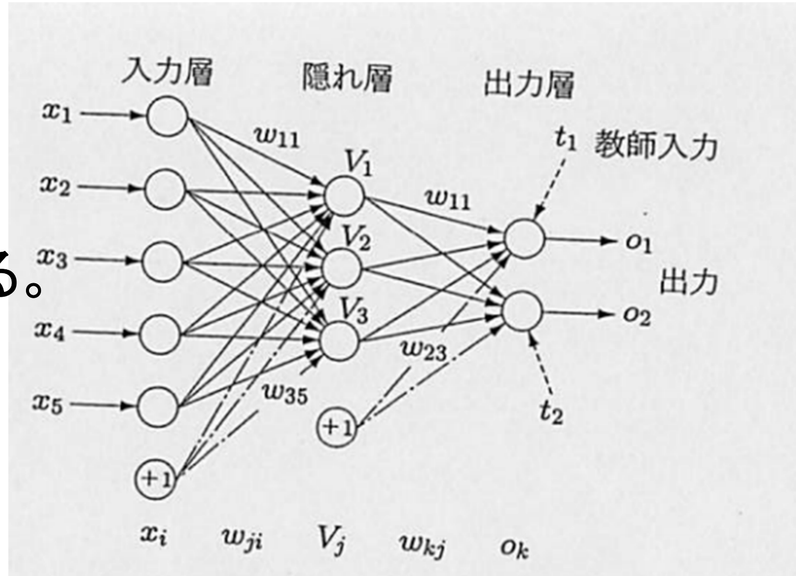
多層回路

- 線形識別関数では正しく識別できない例（排他的論理和関数など）もある。

=>入力を用いて、第三の入力を定義すると線形分離可能になる。



誤差逆伝搬法



- 学習データ : $\boldsymbol{x^n}(\boldsymbol{n}=\mathbf{1},\dots,\boldsymbol{N})$
- 学習データの次元 : d(図ではd=5)
- よって入力は、$\boldsymbol{x^n}=(\mathbf{1},\boldsymbol{x_1^n},\dots,\boldsymbol{x_d^n})^{\boldsymbol{T}}$となる。

隠れ素子$\boldsymbol{V_j}(\boldsymbol{j}=\mathbf{1},\dots,\boldsymbol{M})$

入力:$\boldsymbol{h_j^n}=\sum_{i=0}^{d}\boldsymbol{w_{ji}x_i^n}=\boldsymbol{w_j^T x^n}$

出力:$\boldsymbol{V_j^n}=\boldsymbol{g}\left(\boldsymbol{h_j^n}\right)$

出力素子$\boldsymbol{o_k}(\boldsymbol{k}=\mathbf{1},\dots,\boldsymbol{K})$

入力:$\boldsymbol{h_k^n}=\sum_{j=0}^{M}\boldsymbol{w_{kj}V_j^n}=\sum_{j=0}^{M}\boldsymbol{w_{kj}g}(\sum_{i=0}^{d}\boldsymbol{w_{ji}x_i^n})$

出力:$\boldsymbol{o_k^n}=\tilde{\boldsymbol{g}}(\boldsymbol{h_k^n})=\tilde{\boldsymbol{g}}\left(\sum_{j=0}^{M}\boldsymbol{w_{kj}V_j^n}\right)=\tilde{\boldsymbol{g}}\left(\sum_{j=0}^{M}\boldsymbol{w_{kj}g}(\sum_{i=0}^{d}\boldsymbol{w_{ji}x_i^n})\right)$

$\boldsymbol{g}(\boldsymbol{u})$:非線形出力関数

良く使用されるのはシグモイド関数

$$\boldsymbol{g}(\boldsymbol{u})=\frac{\mathbf{1}}{\mathbf{1}+\mathbf{exp}(-\boldsymbol{\beta u})}$$

出力素子用非線形出力関数

$$\tilde{\boldsymbol{g}}(\boldsymbol{o_k^n})=\mathbf{exp}\boldsymbol{o_k^n}/\sum_{\boldsymbol{l}=\mathbf{1}}^{\boldsymbol{k}}\mathbf{exp}\boldsymbol{o_l^n}$$

または、$\tilde{g}(o_k^n)=p(t_k^n=1|x^n)$

誤差逆伝搬法の学習規則

n番目の学習データによる誤差の評価関数は、

$$E_n(w) = \frac{1}{2}\sum_{k=1}^{K}(t_k^n - o_k^n)^2 = \frac{1}{2}\sum_{k=1}^{K}\left(t_k^n - \tilde{\boldsymbol{g}}\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{M}}\boldsymbol{w_{kj}}\boldsymbol{g}\left(\sum_{\boldsymbol{i=0}}^{\boldsymbol{d}}\boldsymbol{w_{ji}}\boldsymbol{x_i^n}\right)\right)\right)^2$$

となる。学習データ全体では、

$$E(w) = \sum_{n=1}^{N} E_n(w)$$

出力素子係数の更新

$$w_{kj}(\tau+1) = w_{kj}(\tau) + \Delta w_{kj}(\tau)$$

$$\Delta w_{kj}(\mathrm{\tau}) = \sum_{n=1}^{N}\left(-\eta\frac{\partial E_n(w)}{\partial w_{kj}}\right) =$$

$$-\eta\sum_{n=1}^{N}\left(\frac{\partial E_n(w)}{\partial o_k^n}\frac{\partial o_k^n}{\partial w_{kj}}\right) = \eta\sum_{n=1}^{N}(t_k^n - o_k^n)\tilde{g}'(h_k^n)V_j^n = \eta\sum_{n=1}^{N}\delta_k^n V_j^n$$

$\delta_k^n = (t_k^n - o_k^n)\tilde{g}'(h_k^n)$：誤差信号

確率降下法（オンライン学習）:w_{kj}の修正量を次式で与える方法。

$$\Delta w_{kj}(\mathrm{\tau}) = \eta\delta_k^n(\mathrm{\tau})V_j^n(\mathrm{\tau})$$

隠れ素子係数の更新

出力素子と同様に、

$$\Delta w_{ji}(\tau) = \sum_{n=1}^{N}\left(-\eta\frac{\partial E_n(w)}{\partial w_{ji}}\right) = -\eta\sum_{n=1}^{N}\left(\frac{\partial E_n(w)}{\partial V_j^n}\frac{\partial V_j^n}{\partial w_{ji}}\right)$$

$$= -\eta\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\left(\frac{\partial E_n(w)}{\partial o_k^n}\frac{\partial o_k^n}{\partial V_j^n}\right)\frac{\partial V_j^n}{\partial w_{ji}}\right)$$

$$= \eta\sum_{n=1}^{N}\sum_{k=1}^{K}(\mathrm{t_k^n} - o_k^n)\tilde{g}'(h_k^n)w_{kj}g'\left(h_j^n\right)x_i^n$$

$$= \eta\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{\delta_k^n}w_{kj}g'\left(h_j^n\right)x_i^n$$

隠れ素子jの誤差信号：$\delta_j^n = g'(h_j^n)\sum_{k=1}^{K}\delta_k^n w_{kj}$

確率降下法では、$\Delta w_{ji}^n(\tau) = \eta\delta_j^n(\tau)x_i^n(\tau)$

誤差逆伝搬法の学習特性

- 初期依存性がある
  - 誤差評価関数に局所最小値がたくさんあるため、局所最適解が結合係数の初期値により決まる。
- 隠れ素子の数
  - 増やしすぎると汎化誤差が増加する。
    =>過学習
  - ホールドアウト法,交差確認法で最適な数を求める。

過学習と正則化

- 隠れ素子の数が膨大=>過学習　ではない。
- 学習が進む=>結合係数が大きくなる=>シグモイド関数が非線形動作し始める=>ノイズ成分が混じる=>学習がうまくいかない（過学習）
- 結合係数の大きさを抑えるのに正則化を用いる。
- 正則化
  - 荷重減衰ペナルティ:誤差の評価関数にペナルティ項を加える。

$$\tilde{E}(w) = E(w) + \lambda R(w)$$

$$R(w) = \sum_{i=0}^{d}\sum_{j=0}^{M} w_{ji}^2 + \sum_{j=0}^{M}\sum_{k=0}^{K} w_{kj}^2$$

λ:正則化の大きさを決める正則化パラメータ

学習規則 は、 $\Delta w_{kj}(\tau) = \eta \sum_{n=1}^{N} \delta_k^n V_j^n - 2\lambda w_{kj}$ となる。

学習回路の尤度(1)

- 誤差逆伝搬法の評価関数を確率で表す。

$$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^n \log o_k^n$$

- 最尤度推定法により結合係数の更新式は、
- 出力o_kをK個の無関係な確率とみなす場合は、ベルヌーイ試行

$$p(t|x,w) = \prod_{k=1}^{K} o_k^{t_k}(1-o_k)^{1-t_k}$$

と解釈できる。負の対数尤度は交差エントロピー型誤差関数

$$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left(t_k^n \ln o_k^n + (1-t_k^n)\ln(1-o_k^n)\right)$$

となる。

学習回路の尤度(2)

- 出力端子の学習は、

$$\Delta w_{kj} = -\eta \frac{\partial E(w)}{\partial w_{kj}} = \eta \sum_{n=1}^{N} \frac{t_k^n - o_k^n}{o_k^n(1 - o_k^n)} g'\left(\sum_{j=0}^{M} w_{kj} V_j^n\right) V_j^n$$

となる。出力関数がシグモイド関数の場合は、

$$g'\left(\sum_{j=0}^{M} w_{kj} V_j^n\right) = \beta o_k^n (1 - o_k^n)$$

なので、

$$\Delta w_{kj} = \eta \sum_{n=1}^{N} \delta_k^n V_j^n, \qquad \delta_k^n = \beta(t_k^n - o_k^n)$$

となり、出力関数の微分が消えたので非線形部分でも学習できる。