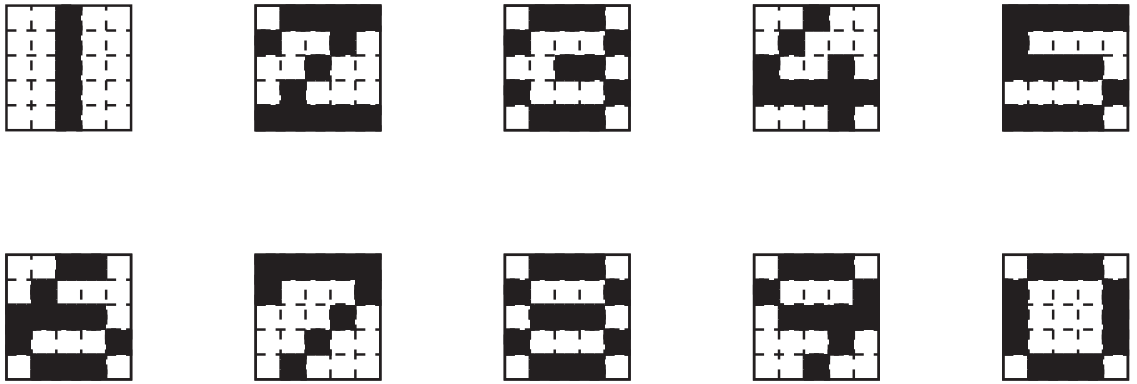
パターン認識

- パターン認識: 「パターンをそれが属すべきカテゴリに対応づける操作」
  - **パターン**: 「空間的または時間的に観測可能な事象であって，観測された対象どうしが同一であるか (または似ているか) 否かを判定できるような性質を備えているもの」
  - **カテゴリ** (または**クラス**): 「パターン認識の結果，同等とみなされるパターンの集合概念」
  - **パターンベクトル**: パターンを表す n 個の数値の組

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$$

  - **特徴ベクトル**: パターンに関する本質的な情報を含む少数個の数値の組．**位相**を崩さずに次元数を極力減らす.

特徴ベクトルと特徴空間の例



5×5メッシュによる2値パターン

- パターン

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_{25})$$

- 特徴量

$$\begin{cases} x_j = 1 & (\text{黒: 文字部分}) \\ x_j = 0 & (\text{白: 背景部分}) \end{cases} \quad (1 \leq j \leq 25)$$

- パターンの組み合わせ: $2^{25} = 33554432$通り
- 出力カテゴリ: 0 から 9 までの数字

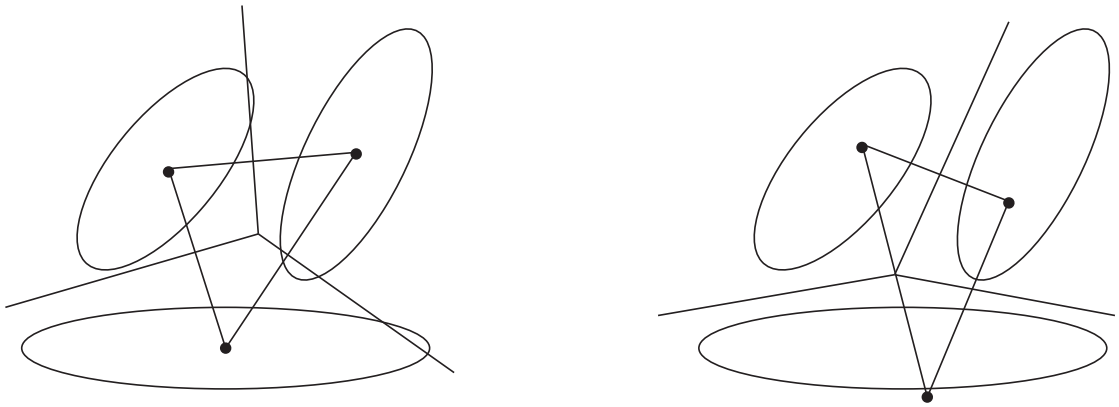
$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_{10}\} \quad (c = 10)$$

決定則の例

- **最近傍決定則**: 入力パターンともっとも距離の近いプロトタイプの属するカテゴリを選ぶ.

$$\min_{p=1,\cdots,n}\{D(\boldsymbol{x},\boldsymbol{x}_p)\} = D(\boldsymbol{x},\boldsymbol{x}_k) \;\Rightarrow\; \boldsymbol{x}\in\omega_k$$

- 問題点
  - 特徴空間の分割法 (3 と 8 は似ている)
  - プロトタイプの選び方 (6 と回転した 9 の識別)



領域分割とプロトタイプ

重みベクトル

- 2分識別器 (正のクラスと負のクラスを識別):

$$f : X \subset \mathbf{R}^n \to \mathbf{R} \quad \text{s.t.} \quad \begin{cases} \text{if} \;\; \boldsymbol{x} \in P : \;\; f(\boldsymbol{x}) \geq 0 \\ \text{otherwise} : \;\; f(\boldsymbol{x}) < 0 \end{cases}$$

  ただし $P \subset X$ は正のクラス

- 教師データ:

$$y_i = \begin{cases} 1 & \text{if} \;\; \boldsymbol{x} \in P \\ -1 & \text{otherwise} \end{cases}$$

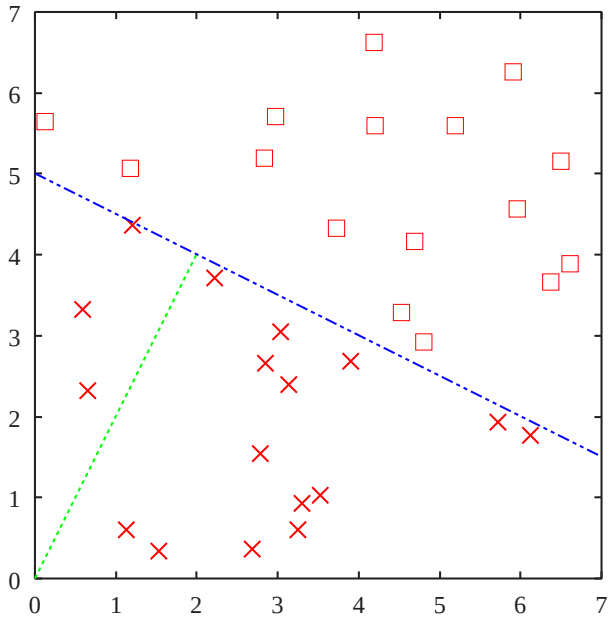
- **線形識別器**

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = \sum_{i=1}^{n} w_i x_i + b$$

  ただし $(\boldsymbol{w}, b) \in (\mathbf{R}^n, \mathbf{R})$

- 識別結果: $y_i f(\boldsymbol{x}) > 0$? (もし正なら正解)
- 線形識別可能性: 線形識別器でカテゴライズ可能かどうか
- 識別面: $\mathbf{R}^n$ における $\mathbf{R}^{n-1}$ 超平面
  - パラメータ数: $n + 1$
  - n 次元は $\mathbf{R}^n$ における平面の法線ベクトル $\boldsymbol{w}$ の次元
  - 1次元は平面の原点からの距離 b

線形識別器

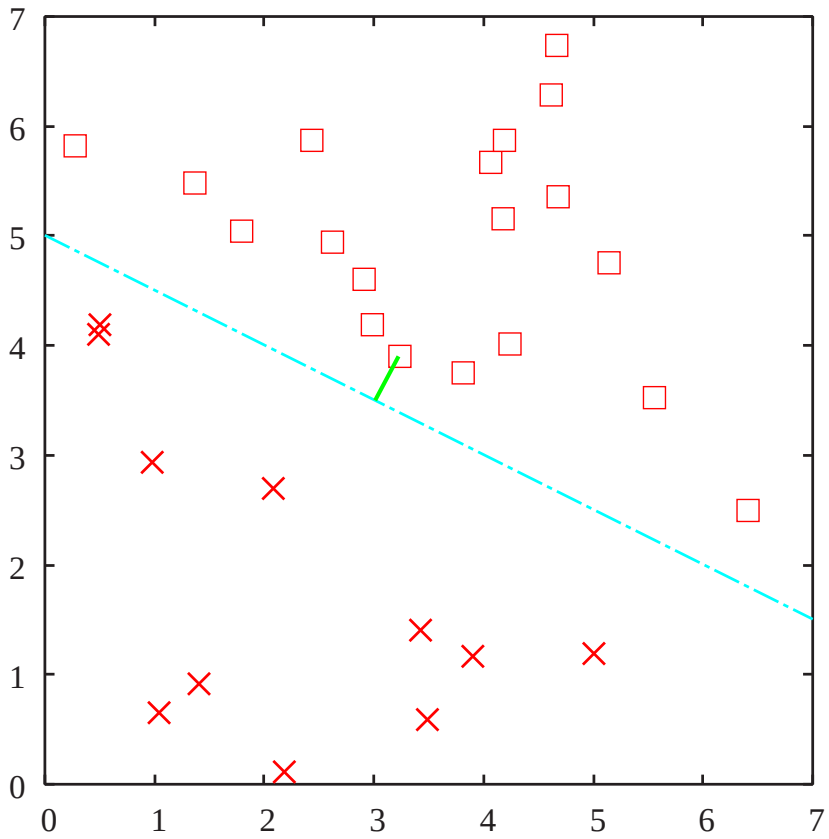


- 識別式 ($\|\boldsymbol{w}\| = 1$)

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = \sum_{i=1}^{n} w_i x_i + b$$

- 識別面の方程式: $f(\boldsymbol{x}) = 0$
- 原点から識別面までの距離: $|b|$
- 識別面の法線ベクトル: $\boldsymbol{w}$
- 識別面から点 $\boldsymbol{x}_j$ までの (符合つき) 距離: $f(\boldsymbol{x}_j)$

マージン



- 超平面 $(\boldsymbol{w}, b)$ に関するトレーニングデータ $(\boldsymbol{x}_i, y_i)$ のマージン (識別面からの距離)

$$\gamma_i = y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b)$$

- $\gamma_i > 0$ **ならば** $(\boldsymbol{x}_i, y_i)$ **に関して正しい識別**
- 超平面 $(\boldsymbol{w}, b)$ に関するトレーニングセット S のマージン: トレーニングデータに関するマージンの最小値
- トレーニングセット S のマージン: すべての超平面 $(\boldsymbol{w}, b)$ に関するトレーニングセット S のマージンの最小値

パーセプトロン

- F. Rosenblatt: The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review*, 65: 386-408, 1959 1960.
- M. L. Minsky and S. A. Papert: Perceptrons, MIT Press, 1969. Expanded Edition 1990.

- 入力空間: $X \subset \mathbf{R}^n$
- 出力カテゴリ: $Y = -1, 1$
- トレーニングデータ: $S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)) \subset (X \times Y)^\ell$
- 判別: $y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0$?
- 重み更新:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta y_i \boldsymbol{x}_i$$
$$b_{k+1} = b_k + \eta y_i R^2$$
$$R = \max_{1 \le i \le \ell} \|\boldsymbol{x}_i\|$$

アルゴリズム

Given linearly separable training set S and learning rate $\eta > 0$

$\boldsymbol{w}_0 = \mathbf{0}; b_0 = 0; k = 0$
$R = \max_{1 \leq i \leq \ell} \|\boldsymbol{x}_i\|$
do
　　for $i = 1$ to ℓ
　　　　if $y_i(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + b_k) \leq 0$ then
　　　　　　$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta y_i \boldsymbol{x}_i$
　　　　　　$b_{k+1} = b_k + \eta y_i R^2$
　　　　　　$k = k + 1$
　　　　end if
　　end for
until no mistakes made within the for loop
return $k, (\boldsymbol{w}_k, b_k)$

where k is the number of mistakes

収束定理

(Novikoff) Let S be a non-tribial training set, and let

$$R = \max_{1 \leq i \leq \ell} \|\boldsymbol{x}_i\|.$$

Suppose that there exists a vector $\boldsymbol{w}_{opt}$ such that $\|\boldsymbol{w}_{opt}\| = 1$ and

$$y_i(\langle \boldsymbol{w}_{opt}, \boldsymbol{x}_i \rangle + b_{opt}) \geq \gamma$$

for $1 \leq i \leq \ell$. Then the number of mistakes made by the on-line perceptron algorithm on S is at most

$$\left(\frac{2R}{\gamma}\right)^2.$$

- 線形識別可能性を仮定すると間違いの回数の上限が与えられる
- 線形識別可能でない場合は修正を繰り返し収束しない

証明

Define $\hat{\boldsymbol{x}}_i = (\boldsymbol{x}_i^T, R)^T, \hat{\boldsymbol{w}} = (\boldsymbol{w}^T, b/R)^T$. Let $\hat{\boldsymbol{w}}_{t-1}$ be the augmented weight vector prior to the tth mistake. The tth update is performed when

$$y_i \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{x}}_i \rangle = y_i (\langle \boldsymbol{w}_{t-1}, \boldsymbol{x}_i \rangle + b_{t-1}) \leq 0$$

where $(\boldsymbol{x}_i, y_i)$ is the point incorrectly classified by $\hat{\boldsymbol{w}}_{t-1}$. The update is the following:

$$\hat{\boldsymbol{w}}_t = \begin{bmatrix} \boldsymbol{w}_t \\ b_t/R \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{t-1} \\ b_{t-1}/R \end{bmatrix} + \eta y_i \begin{bmatrix} \boldsymbol{x}_i \\ R \end{bmatrix} = \hat{\boldsymbol{w}}_{t-1} + \eta y_i \hat{\boldsymbol{x}}_i$$

where $b_t = b_{t-1} + \eta y_i R^2$. Since the margin is γ, we have

$$\langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle = \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{w}}_{opt} \rangle + \eta y_i \langle \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{w}}_{opt} \rangle \geq \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{w}}_{opt} \rangle + \eta\gamma$$

and this implies that

$$\langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle \geq t\eta\gamma$$

Similarly, we have

$$\begin{aligned} \|\hat{\boldsymbol{w}}_t\|^2 &= \|\hat{\boldsymbol{w}}_{t-1}\|^2 + 2\eta y_i \langle \hat{\boldsymbol{w}}_{t-1}, \hat{\boldsymbol{x}}_i \rangle + \eta^2 \|\hat{\boldsymbol{x}}_i\|^2 \\ &\leq \|\hat{\boldsymbol{w}}_{t-1}\|^2 + \eta^2 \|\hat{\boldsymbol{x}}_i\|^2 \leq \|\hat{\boldsymbol{w}}_{t-1}\|^2 + \eta^2 (\|\boldsymbol{x}_i\|^2 + R^2), \end{aligned}$$

which implies $\|\hat{\boldsymbol{w}}_t\|^2 \leq \|\hat{\boldsymbol{w}}_{t-1}\|^2 + 2\eta^2 R^2$. Thus we have

$$\|\hat{\boldsymbol{w}}_t\|^2 \leq 2t\eta^2 R^2.$$

証明(つづき)

The two inequalities combined give the 'squeezing' relations

$$\|\boldsymbol{w}_{opt}\|\sqrt{2t}\eta R \geq \|\boldsymbol{w}_{opt}\|\|\boldsymbol{w}_t\| \geq \langle \hat{\boldsymbol{w}}_t, \hat{\boldsymbol{w}}_{opt} \rangle \geq t\eta\gamma,$$

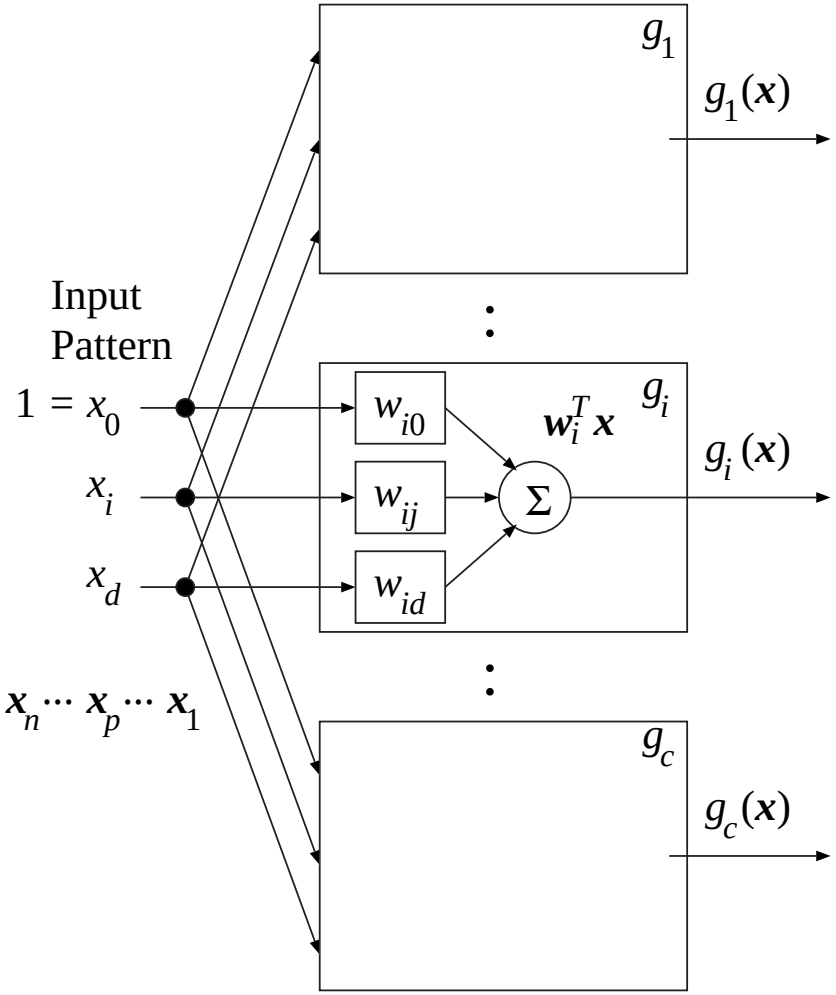
which together imply the bound

$$t \leq 2\left(\frac{R}{\gamma}\right)^2 \|\hat{\boldsymbol{w}}_{opt}\|^2 \leq \left(\frac{2R}{\gamma}\right)^2$$

since $b_{opt} \leq R$ and thus $\|\hat{\boldsymbol{w}}_{opt}\|^2 \leq \|\boldsymbol{w}_{opt}\|^2 + 1 = 2$.

- 線形識別可能性を仮定すれば間違いの回数に上限が存在
- つまり有限回の繰り返しで収束
- 一方，線形識別可能でなければ修正を繰り返し収束しない
- したがって，線形識別可能性の判断には使えない
- 途中で打ち切った場合，そのときの重みが最適かどうかわからない
- そもそも最適性の評価基準は？

誤差評価に基づく学習



- カテゴリ: $\omega_1, \cdots, \omega_c$
- 学習パターン: $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$
- 教師ベクトル: $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$

$$\boldsymbol{b}_p = [b_{1p}, \cdots, b_{cp}]^T$$

- 出力ベクトル:

$$\boldsymbol{g}(\boldsymbol{x}_p) = [g_1(\boldsymbol{x}_p), \cdots, g_c(\boldsymbol{x}_p)]^T$$

- 線形識別関数

$$g_i(\boldsymbol{x}_p) = w_{i0} + \sum_{j=1}^{d} w_{ij} x_{pj} = \boldsymbol{w}_i^T \boldsymbol{x}_p$$

$$\boldsymbol{w}_i = [w_{i0}, w_{i1}, \cdots, w_{id}]^T,$$
$$\boldsymbol{x} = [1, x_{p1}, \cdots, x_{pd},]^T$$

評価関数

- 誤差: $\boldsymbol{\varepsilon}_p = [\varepsilon_{1p}, \cdots, \varepsilon_{cp}]^T = \boldsymbol{g}(\boldsymbol{x}_p) - \boldsymbol{b}_p$

$$\varepsilon_{ip} = g_i(\boldsymbol{x}_p) - b_{ip}$$

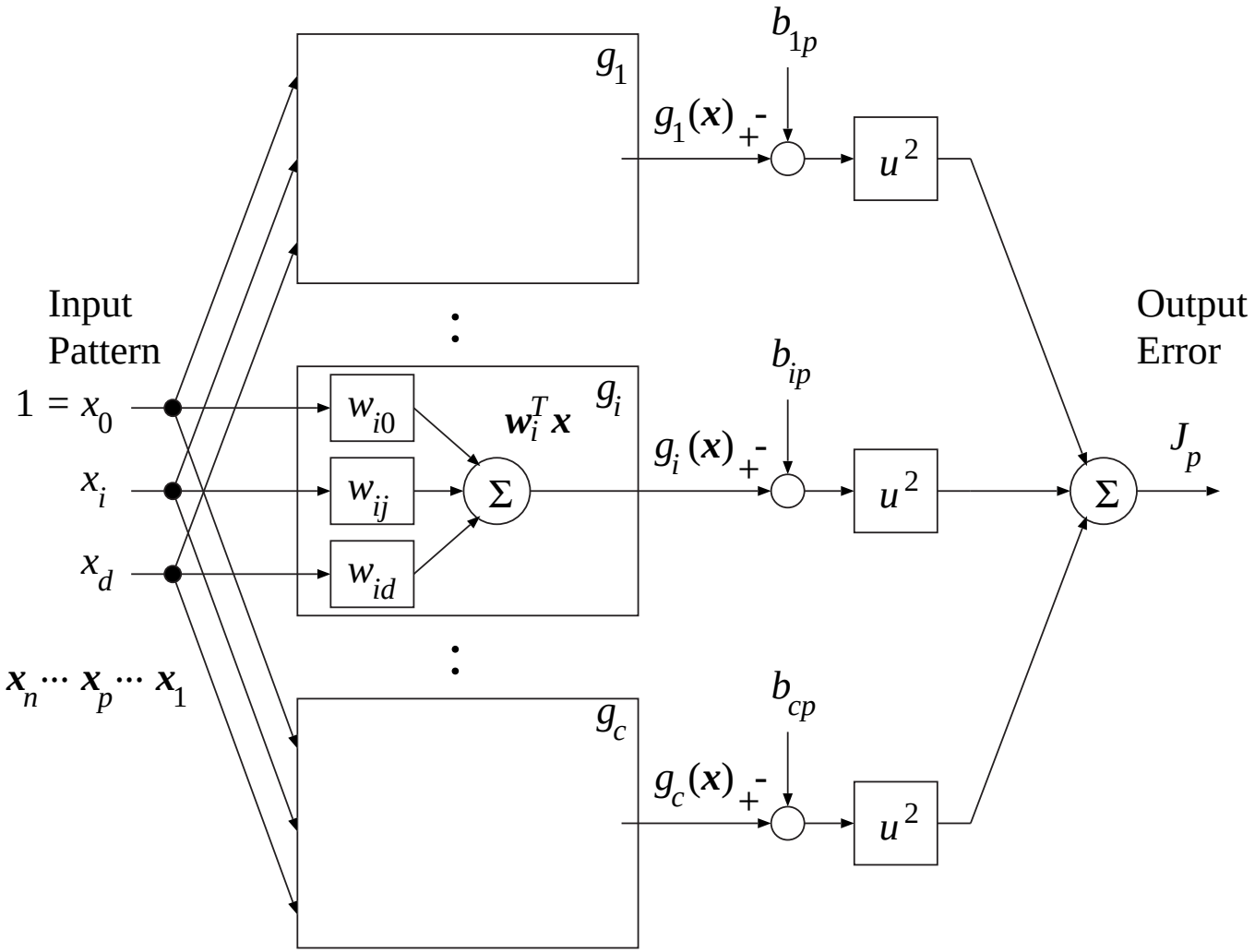
- パターン $\boldsymbol{x}_p$ に対する評価関数: J_p

$$J_p(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_c) = \frac{1}{2}\|\boldsymbol{\varepsilon}_p\|^2 = \frac{1}{2}\sum_{i=1}^{c}\varepsilon_{ip}^2 = \frac{1}{2}\sum_{i=1}^{c}(\boldsymbol{w}_i^T\boldsymbol{x}_p - b_{ip})^2$$

- 全パターンに対する評価関数: J

$$J(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_c) = \sum_{p=1}^{n} J_p(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_c) = \frac{1}{2}\sum_{p=1}^{n}\sum_{i=1}^{c}(\boldsymbol{w}_i^T\boldsymbol{x}_p - b_{ip})^2$$

評価機構をもつ識別器


b_{1p}
g_1
$g_1(\boldsymbol{x})$
u^2
Input
Pattern
Output
Error
b_{ip}
g_i
$\boldsymbol{w}_i^T \boldsymbol{x}$
$g_i(\boldsymbol{x})$
J_p
$1 = x_0$
x_i
x_d
w_{i0}
w_{ij}
w_{id}
Σ
$\boldsymbol{x}_n \cdots \boldsymbol{x}_p \cdots \boldsymbol{x}_1$
b_{cp}
g_c
$g_c(\boldsymbol{x})$

最適重み

- 最適重み: 評価関数 J を最小にする $\boldsymbol{w}_i$ $(i = 1, \ldots, c)$

$$\frac{\partial J}{\partial \boldsymbol{w}_i} = 0 \quad (i = 1, \ldots, c)$$

- 全パターンに対する評価関数: J

$$J(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_c) = \frac{1}{2} \sum_{p=1}^{n} \sum_{i=1}^{c} (\boldsymbol{w}_i^T \boldsymbol{x}_p - b_{ip})^2$$

- 最適重み $\boldsymbol{w}_i$ は次式をみたす

$$\begin{aligned} 0 &= \frac{\partial J}{\partial \boldsymbol{w}_i} = \sum_{p=1}^{n} (\boldsymbol{w}_i^T \boldsymbol{x}_p - b_{ip}) \boldsymbol{x}_p = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \begin{bmatrix} \boldsymbol{w}_i^T \boldsymbol{x}_1 - b_{i1} \\ \vdots \\ \boldsymbol{w}_i^T \boldsymbol{x}_n - b_{in} \end{bmatrix} \\ &= [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \left(\begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{w}_i - \begin{bmatrix} b_{i1} \\ \vdots \\ b_{in} \end{bmatrix} \right) = X^T (X \boldsymbol{w}_i - \bar{\boldsymbol{b}}_i) \end{aligned}$$

ただし (X は**パターン行列**と呼ばれる)

$$X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T, \quad \bar{\boldsymbol{b}}_i = [b_{i1}, \cdots, b_{in}]^T$$

最適重み

- **注意:** $\bar{\boldsymbol{b}}_i$ は教師ベクトル $\boldsymbol{b}_p$ とは以下の関係がある.

$$[\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n] = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{c1} & \cdots & b_{cn} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{b}}_1^T \\ \vdots \\ \bar{\boldsymbol{b}}_c^T \end{bmatrix}$$

- 最適重み

$$\boldsymbol{w}_i = (X^T X)^{-1} X^T \boldsymbol{b}_i$$

逐次法

- **最急降下法**(パターンが提示されるたびに修正を行う)

$$\boldsymbol{w}_i' = \boldsymbol{w}_i - \rho \frac{\partial J_p}{\partial \boldsymbol{w}_i} \quad (i = 1, \ldots, c)$$

$$J_p(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_c) = \frac{1}{2} \sum_{i=1}^{c} (\boldsymbol{w}_i^T \boldsymbol{x}_p - b_{ip})^2$$

ここで $\varepsilon_{ip} = \boldsymbol{w}_i^T \boldsymbol{x}_p - b_{ip}$ に注意すると

$$\frac{\partial J_p}{\partial \boldsymbol{w}_i} = \frac{\partial J_p}{\partial \varepsilon_{ip}} \frac{\partial \varepsilon_{ip}}{\partial \boldsymbol{w}_i} = \varepsilon_{ip} \boldsymbol{x}_p$$

であるから

$$\boldsymbol{w}_i' = \boldsymbol{w}_i - \rho \varepsilon_{ip} \boldsymbol{x}_p = \boldsymbol{w}_i - \rho (\boldsymbol{w}_i^T \boldsymbol{x}_p - b_{ip}) \boldsymbol{x}_p \quad (i = 1, \ldots, c)$$

を得る．これを **Widrow-Hoff の学習則**という．

パーセプトロン(再)

- **学習:** $\mathcal{X}_i$ に属するすべての $\boldsymbol{x}$ に対して

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad (j = 1, \cdots, c,\ j \neq i)$$

が成り立つように重み $\boldsymbol{w}_i$ を決定すること.

- 学習アルゴリズム

1. 重みベクトル $\boldsymbol{w}_i$ の初期値を適当に設定する.
2. $\mathcal{X}_i$ のなかから学習パターンをひとつ選ぶ.
3. 識別関数

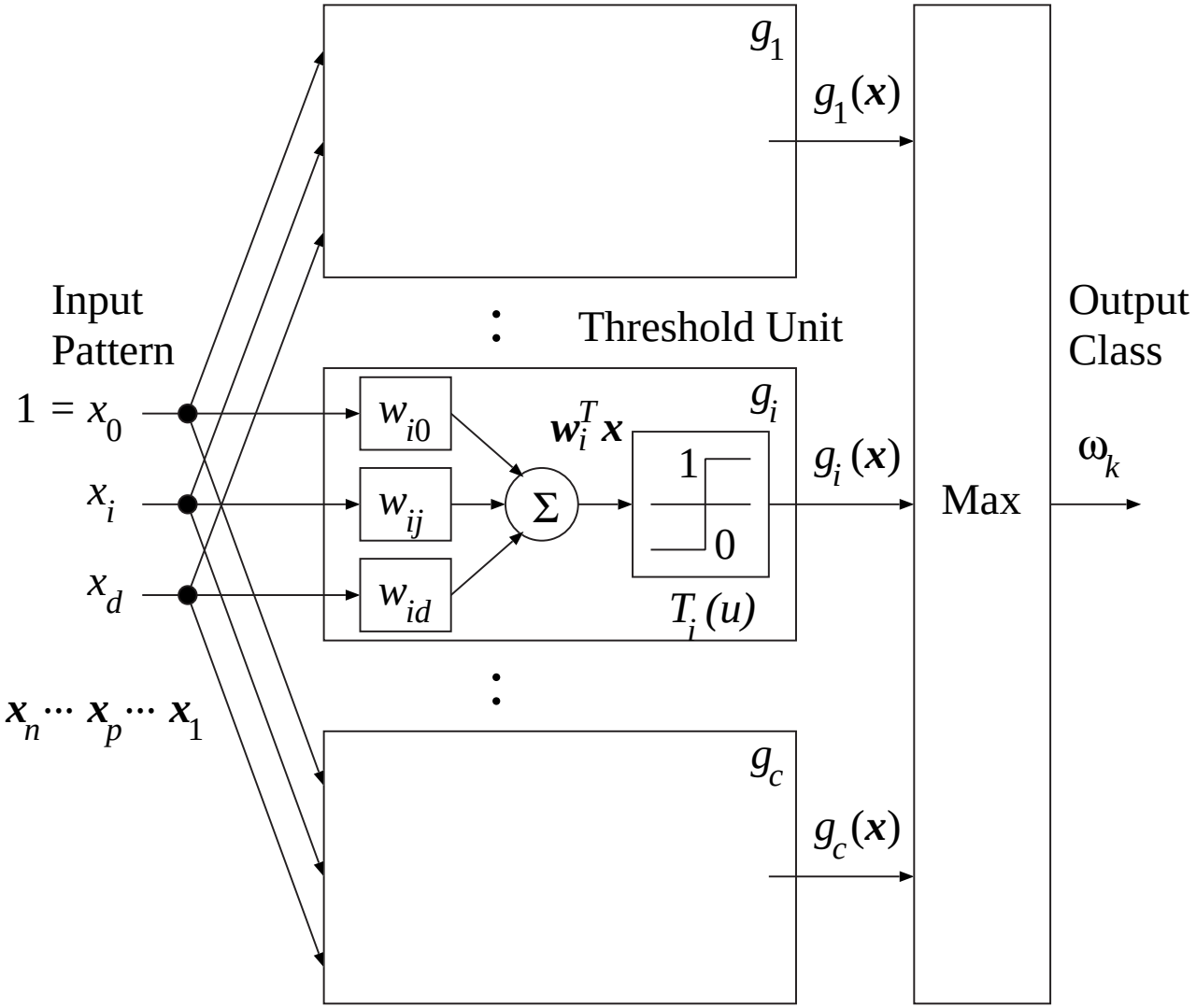
$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x}$$

によって識別を行い，誤識別を生じた場合 (ω_i に属するパターンを ω_j と誤ったとする) のみ次式により重みベクトルを修正する.

$$\begin{aligned} \boldsymbol{w}_i' &= \boldsymbol{w}_i + \rho \boldsymbol{x} \\ \boldsymbol{w}_j' &= \boldsymbol{w}_j - \rho \boldsymbol{x} \end{aligned}$$

4. 上の処理，2, 3 を $\mathcal{X}_i$ の全パターンに対して繰り返す.
5. $\mathcal{X}_i$ の全パターンを正しく識別できたら終了，誤りがあるときには 2 に戻る.

パーセプトロン(ブロック図)



誤差評価とパーセプトロン

- 教師信号

$$b_{ip} = \begin{cases} 1 & (\boldsymbol{x}_p \in \omega_i) \\ 0 & (\boldsymbol{x}_p \notin \omega_i) \end{cases} \qquad (i = 1, \cdots, c)$$

- 出力にしきい値関数T_iを施す

$$T_i(u) = \begin{cases} 1 & (u > 0) \\ 0 & (u \leq 0) \end{cases} \qquad (i = 1, \cdots, c)$$

- Widrow-Hoff 学習規則 (パターン$\boldsymbol{x}_p \in \omega_i$を$\omega_j$と誤認識したとき)

$$\begin{cases} \boldsymbol{w}_i' = \boldsymbol{w}_i + \rho(g_i(\boldsymbol{x}_p) - b_{ip})\boldsymbol{x}_p = \boldsymbol{w}_i + \rho\boldsymbol{x}_p \\ \boldsymbol{w}_j' = \boldsymbol{w}_j + \rho(g_j(\boldsymbol{x}_p) - b_{jp})\boldsymbol{x}_p = \boldsymbol{w}_j - \rho\boldsymbol{x}_p \end{cases}$$

- パーセプトロンは Widrow-Hoff 学習規則の一種 !?