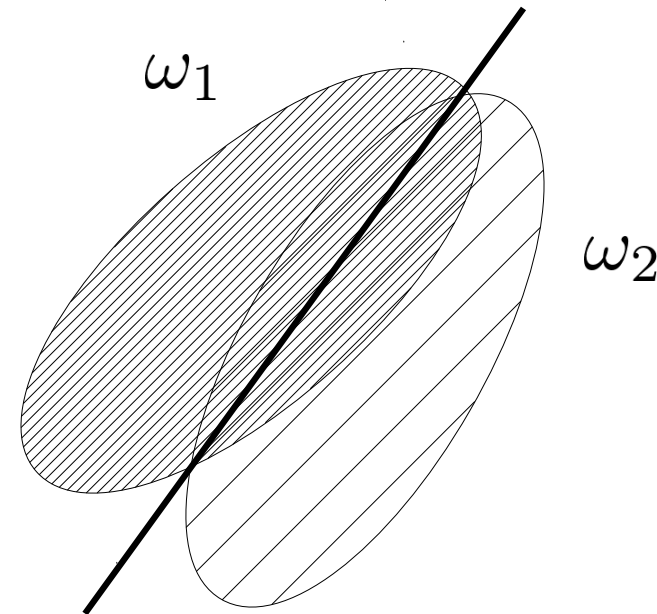


5. 誤差をできるだけ小さくしよう

- パーセプトロンの学習規則の欠点
 - 線形分離不可能である場合には利用できない
 - 一般に線形分離可能性を事前に確認するのは困難



- 評価関数最小化法
 - 線形分離不可能な場合にも適用可能



5.1 誤差評価に基づく学習とは

- 学習パターン $\chi = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$
- \boldsymbol{x}_p ($1 \leq p \leq n$) に対する c 個の識別関数の出力
 $(g_1(\boldsymbol{x}_p), \dots, g_c(\boldsymbol{x}_p))^T$
- \boldsymbol{x}_p に対する教師ベクトル (教師信号)
 $(b_{1p}, \dots, b_{cp})^T$
 - 正解クラスの要素が 1、他は 0
- 入力パターン \boldsymbol{x}_p に対する識別関数の出力と、教師信号との誤差 ϵ_{ip} ($i = 1, \dots, c$) が小さくなるように重みベクトル \boldsymbol{w} を定める

5.1 誤差評価に基づく学習とは

- 誤差 $\epsilon_{ip} = g_i(\mathbf{x}_p) - b_{ip}$
- ϵ_{ip} の全クラスに対する二乗和を評価関数 J_p とする

$$\begin{aligned} J_p(\mathbf{w}_1, \dots, \mathbf{w}_c) &= \frac{1}{2} \sum_{i=1}^c \epsilon_{ip}^2 \\ &= \frac{1}{2} \sum_{i=1}^c (g_i(\mathbf{x}_p) - b_{ip})^2 \\ &= \frac{1}{2} \sum_{i=1}^c (\mathbf{w}_i^T \mathbf{x}_p - b_{ip})^2 \end{aligned}$$

\mathbf{x}_p に対する誤差

5.1 誤差評価に基づく学習とは

- 全パターンに対する二乗誤差 J

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

この値を最小にする $\boldsymbol{w}_1, \dots, \boldsymbol{w}_c$ を求める

5.1 誤差評価に基づく学習とは

- 2クラス問題を考える場合

- 識別関数を1つにできる

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$$

- 教師信号はクラス1のとき1、クラス2のとき-1

- 誤差関数

$$J(\boldsymbol{w}) = \frac{1}{2} \sum_{p=1}^n (\boldsymbol{w}^T \boldsymbol{x}_p - b_p)^2$$

5.2 解析的な解法

- パターン行列

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$$

教師信号ベクトル

$$\mathbf{b} = (b_1, \dots, b_n)^T$$

とすると

$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{b}\|^2$$

$$\frac{\partial J}{\partial \mathbf{w}} = \mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{b}) = 0$$

解くべき式

5.2 解析的な解法

- 解くべき式

$$X^T X w = X^T b$$

$X^T X$ が正則であるとき

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

- 解が求まらない可能性

- $X^T X$ が正則であるとは限らない
- d が大きい場合は逆行列を求める計算が大変

正則：

逆行列が存在すること

逆行列：

$n \times n$ の正方行列Aに対して、
 $AB = BA = I$ となるB

最小二乗法

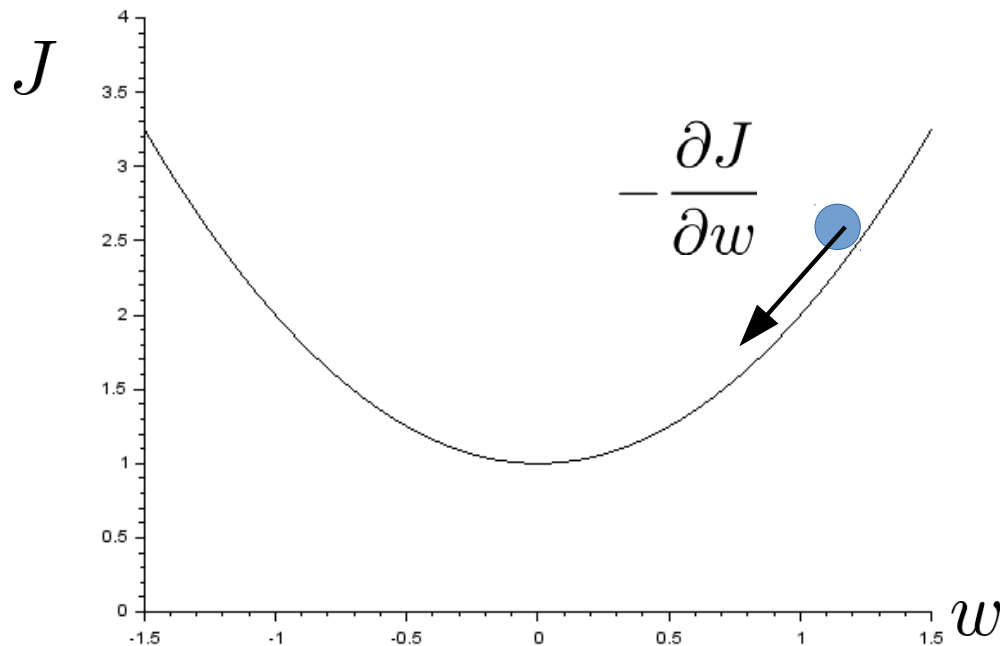
5.3 最急降下法

5.3.1 最急降下法による最適化

- w を J の傾きの方向に徐々に修正する

$$w' = w - \rho \frac{\partial J}{\partial w}$$

- 最急降下法のイメージ



5.3.2 Widrow-Hoffの学習規則

- 勾配ベクトルの定義
 - 重みベクトル

$$\boldsymbol{w} = (w_0, \dots, w_d)$$

の関数 $J(\boldsymbol{w})$ に対して、勾配ベクトルを

$$\nabla J = \frac{\partial J}{\partial \boldsymbol{w}} = \left(\frac{\partial J}{\partial w_0}, \dots, \frac{\partial J}{\partial w_d} \right)^T$$

と定義する

5.3.2 Widrow-Hoffの学習規則

- 修正式の導出

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{w}} &= \sum_{p=1}^n \frac{\partial J_p}{\partial \mathbf{w}} \\ &= \sum_{p=1}^n (\mathbf{w}^T \mathbf{x}_p - b_p) \mathbf{x}_p \\ \mathbf{w}' &= \mathbf{w} - \rho \frac{\partial J}{\partial \mathbf{w}} \\ &= \mathbf{w} - \rho \sum_{p=1}^n (\mathbf{w}^T \mathbf{x}_p - b_p) \mathbf{x}_p\end{aligned}$$

重みの修正式

5.3.3 確率的最急降下法

- 最急降下法の問題点
 - データ数やパラメータ数が多いと、重み更新に時間がかかる
- 確率的最急降下法
 - 個々のデータの識別結果に基づき、重みを更新
 - データが来る毎に学習するオンライン学習が可能
 - 更新式

$$\boldsymbol{w}' = \boldsymbol{w} - \rho(\boldsymbol{w}^T \boldsymbol{x}_p - b_p) \boldsymbol{x}_p$$

5.3.3 確率的最急降下法

- ミニバッチ法
 - Widorow-Hoffの学習規則のように、全データの誤差を用いて修正方向を決める方法をバッチ法とよぶ
 - 確率的最急降下法は1つのデータだけで修正方向を決める
 - これらの中間的手法として、数十～数百程度のデータで誤差を計算し、修正方向を決める方法を**ミニバッチ法**とよぶ
 - GPU (graphics processing unit) を用いた行列の一括演算と相性がよい

5.4 パーセプトロンの学習規則との比較

5.4.1 パーセプトロンの学習規則を導く

- 更新式の導出

- オンラインの学習規則において

- 教師信号を正解のときは1、不正解は0とする
- 識別関数の後ろに閾値論理ユニットを置き、出力を0と1に制限する

誤識別のパターン

$$g(\mathbf{x}_p) = 0, b_p = 1$$

$$g(\mathbf{x}_p) = 1, b_p = 0$$



更新規則

$$\mathbf{w}' = \mathbf{w} + \rho \mathbf{x}_p$$

$$\mathbf{w}' = \mathbf{w} - \rho \mathbf{x}_p$$

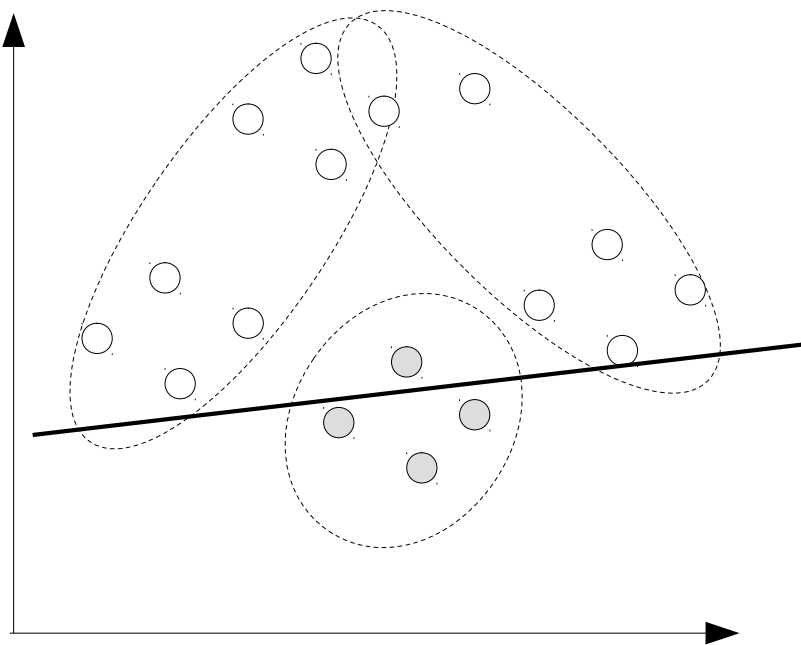
Widrow-Hoffの学習規則は、パーセプトロンの学習規則を特別な場合として含む

5.4.2 着目するデータの違い

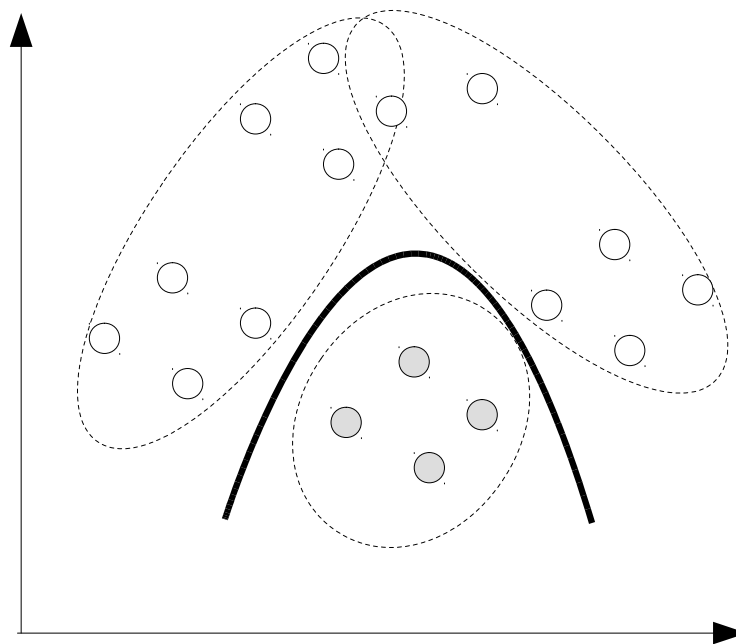
- パーセプトロンの学習規則
 - 識別関数、教師信号ともに 2 値
 - 全学習パターンに対して、識別関数の出力と教師信号が一致するまで重みの修正を繰り返す
 - 線形分離不可能な場合は収束しない
 - 誤識別を起こすデータに着目している
- Widrow-Hoffの学習規則
 - 識別関数の出力を連続値とし、教師信号との二乗誤差の総和を最小化
 - 線形分離不可能な場合でも収束が保証されている
 - 線形分離可能な場合でも誤識別 0 になるとは限らない
 - 全データの誤差に着目している

複雑な境界線は学習できないか

- 識別関数の次数を上げれば境界線は複雑にできる



線形分離不可能なデータ



2次関数を用いた場合

複雑な境界線は学習できないか

- 基底関数ベクトル

$$\phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \dots, \phi_b(\boldsymbol{x}))^T$$

- 例) $\phi(x) = (1, x, \dots, x^b)^T$

- 基底関数ベクトルを用いた識別関数の表現

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

- 重みに関して線形であれば、最小二乗法の閉じた解による解法や、最急降下法による調整が適用可能