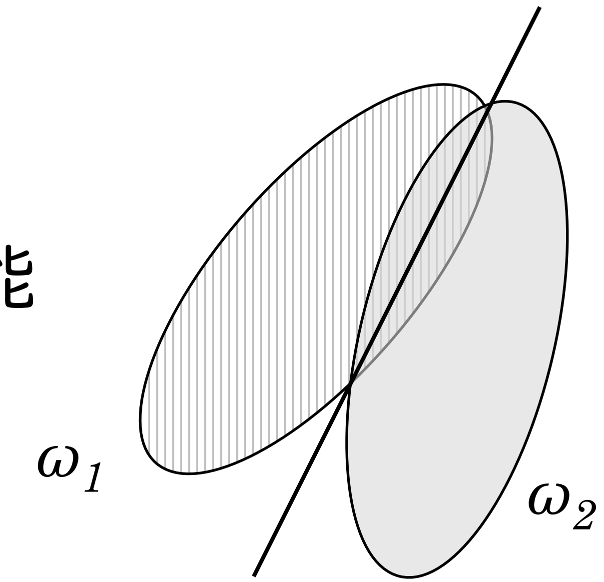
5. 誤差をできるだけ小さくしよう

- 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
  - ◆ 「誤差」を最小化する
  - ◆ 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　$\chi = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$
- $\boldsymbol{x}_p$ $(p = \{1, \dots, n\})$ 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: $\varepsilon_{ip} = g_i(\boldsymbol{x}_p) - b_{ip}$
- ε_{ip} の全クラスに対する二乗和を評価関数 J_p とする

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

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

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

  - ◆ この値を最小にする $\boldsymbol{w}_1, \dots, \boldsymbol{w}_c$ を求める
  - ◆ 以後2クラス問題として、$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ とする
  - ◆ デモ：https://playground.tensorflow.org/

5.2 解析的な解法

- パターン行列（全特徴ベクトルをまとめた行列）

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$$

- 教師信号ベクトル（全教師信号をまとめたベクトル）

$$\boldsymbol{b} = (b_1, \ldots, b_n)^T$$

- 二乗誤差

$$J = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{b}\|^2$$

- 二乗誤差の勾配が0（極小値）となる $\boldsymbol{w}$ を求める

$$\frac{\partial J}{\partial \boldsymbol{w}} = \underbrace{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{b}) = 0}_{\text{解くべき式}}$$

5.2 解析的な解法

- 解くべき式: $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{b}$
- 最小二乗法
  - $\boldsymbol{X}^T\boldsymbol{X}$ が正則であるとき、以下のように解が求まる

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

  - 解が求まらない可能性
    - $\boldsymbol{X}^T\boldsymbol{X}$ 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
- 例題5.1
  - https://github.com/MasahiroAraki/SpeechRecognition/blob/master/Python/chap05.ipynb

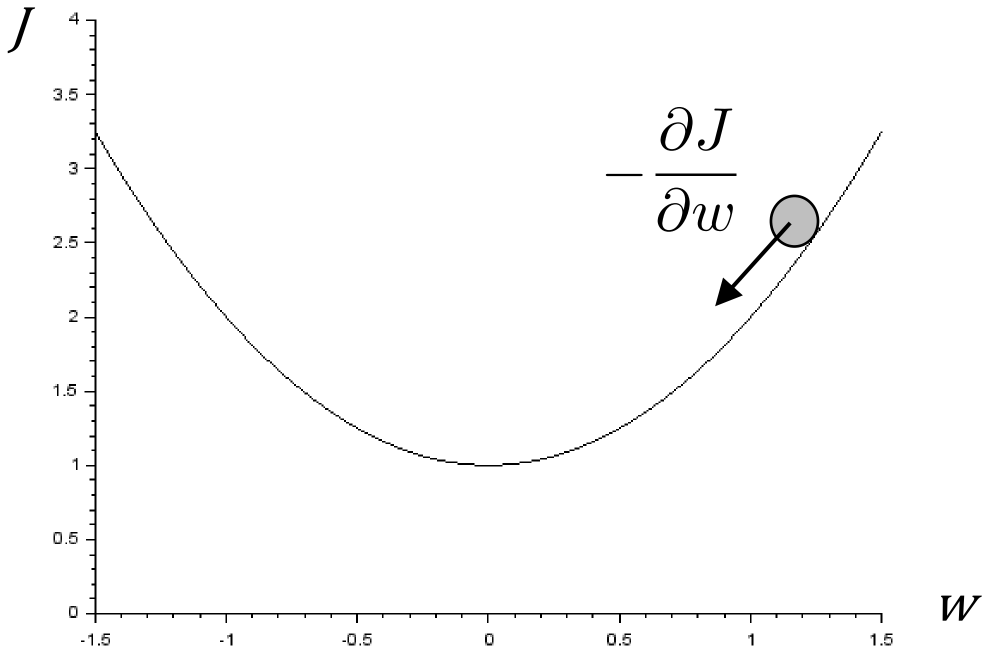
5.3 最急降下法

5.3.1 最急降下法による最適化

- $\boldsymbol{w}$ を J の傾きの方向に徐々に修正する

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

- 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

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

の関数 $J(\boldsymbol{w})$ に対して、勾配ベクトルを

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

と定義する

5.3.2 Widrow-Hoffの学習規則

- 修正式の導出

$$\frac{\partial J}{\partial \boldsymbol{w}} = \sum_{p=1}^{n} \frac{\partial J_p}{\partial \boldsymbol{w}}$$

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

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

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

重みの修正式

- 例題5.3

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
  - Widrow-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（→ 解への収束が安定しない）
  - 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

$$g(\boldsymbol{x}) = 0, b = 1$$
$$g(\boldsymbol{x}) = 1, b = 0$$

⇒

更新規則

$$\boldsymbol{w}' = \boldsymbol{w} + \rho \boldsymbol{x}$$
$$\boldsymbol{w}' = \boldsymbol{w} - \rho \boldsymbol{x}$$

→ 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