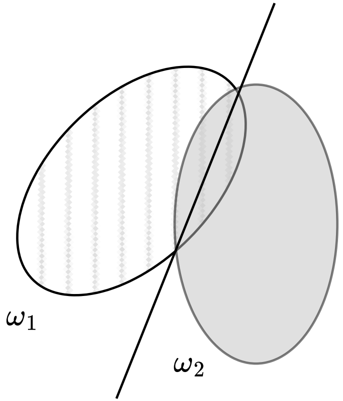
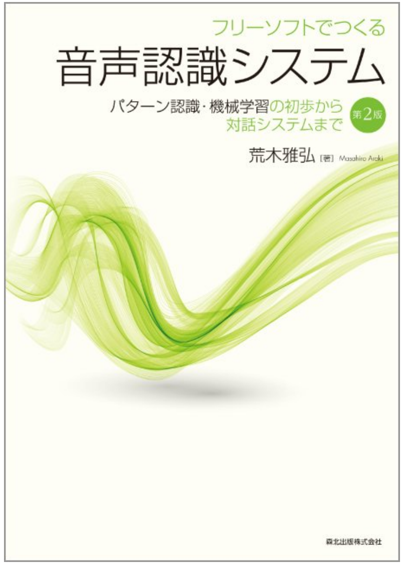
5. 誤差をできるだけ小さくしよう



- 5.1 誤差評価に基づく学習とは
- 5.2 解析的な解法
- 5.3 勾配降下法
- 5.4 パーセプトロンの学習規則との比較



- 荒木雅弘:『フリーソフトでつくる音声認識システム(第2版)』(森北出版, 2017年)
- スライドとJupyter notebook
- サポートページ

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
  - 学習データが線形分離不可能である場合には適用できない
  - 学習データが高次元である場合、線形分離可能性を事前に確認するのは一般には困難
- 誤差最小化法
  - **評価関数**で定義される「識別器の出力」と「正解」との**誤差**を最小化する
  - 学習データが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 (1/3)

- 学習データ: $\chi = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$
- あるデータ $\boldsymbol{x}_p \in \chi$ に対する線形識別関数 $g_i(\boldsymbol{x}_p) = \boldsymbol{w}_i^T \boldsymbol{x}_p$ $(i = 1, \ldots, c)$ の出力

$$(g_1(\boldsymbol{x}_p), \ldots, g_c(\boldsymbol{x}_p))^T$$

- $\boldsymbol{x}_p$ に対する教師ベクトル(教師信号)

$$(b_{1p}, \ldots, b_{cp})^T$$

  - 正解クラスの要素が 1、他は 0
- 入力 $\boldsymbol{x}_p$ に対する「識別関数の出力」と「教師信号」との差が最小となるように、識別関数の重みベクトル $\boldsymbol{w}_i$ を定める

5.1 誤差評価に基づく学習とは (2/3)

- クラス i に関する誤差 : $\epsilon_{ip} = g_i(\boldsymbol{x}_p) - b_{ip}$
- ϵ_{ip} の全クラス $(i = 1, \ldots, c)$ 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 (3/3)

- 全データに対する二乗誤差 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 $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$ 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
  - 教師信号 b : クラス1は1、クラス2は-1

5.2 解析的な解法 (1/2)

- パターン行列(全特徴ベクトルをまとめた $n \times d$ 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 n 次元ベクトル) : $\boldsymbol{b} = (b_1, \ldots, b_n)^T$
- 二乗誤差を評価関数とする

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

- 評価関数の勾配が 0(極小値)となる $\boldsymbol{w}$ を求める

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

5.2 解析的な解法 (2/2)

- 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(識別関数の重み)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
    - n, d が大きい場合は逆行列を求める計算が大変

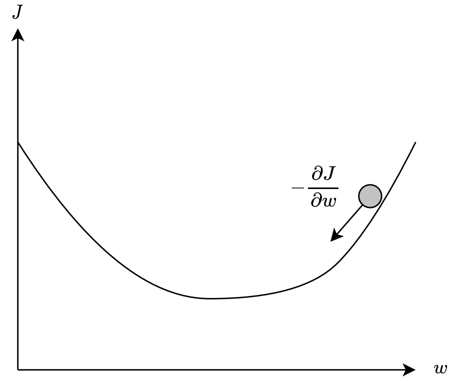
5.3 勾配降下法

5.3.1 勾配降下法による最適化

- $\boldsymbol{w}$ を J の傾き(微分係数)の逆方向に、学習係数 ρ で徐々に修正する

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

- 勾配降下法のイメージ



5.3.2 Widrow-Hoffの学習規則 (1/2)

- 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、勾配ベクトルを以下のように定義

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

5.3.2 Widrow-Hoffの学習規則 (2/2)

- 修正式の導出

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{w}} &= \sum_{p=1}^{n} \frac{\partial J_p}{\partial \boldsymbol{w}} \\ &= \sum_{p=1}^{n} (\boldsymbol{w}^T \boldsymbol{x}_p - b_p)\boldsymbol{x}_p\end{aligned}$$

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

- 全データを用いた重みの修正を1エポックとよぶ
  - Widrow-Hoff の学習規則は1エポックで1回の重み修正が行われる

5.3.3 確率的勾配降下法（1/3）

- 勾配降下法の問題点
  - データ数やパラメータ数が多いと、1回の重み更新(=1エポック)に時間がかかる
  - 誤差最小に収束するには、多くのエポックが必要
- 確率的勾配降下法
  - 個々のデータの識別結果に基づき重みを更新
    - 1エポックで n 回重み修正が行われる
    - データは1エポック毎に順番をシャッフルし、出現順序をランダム化する
  - データが来る毎に学習するオンライン学習に適用が可能
  - 更新式

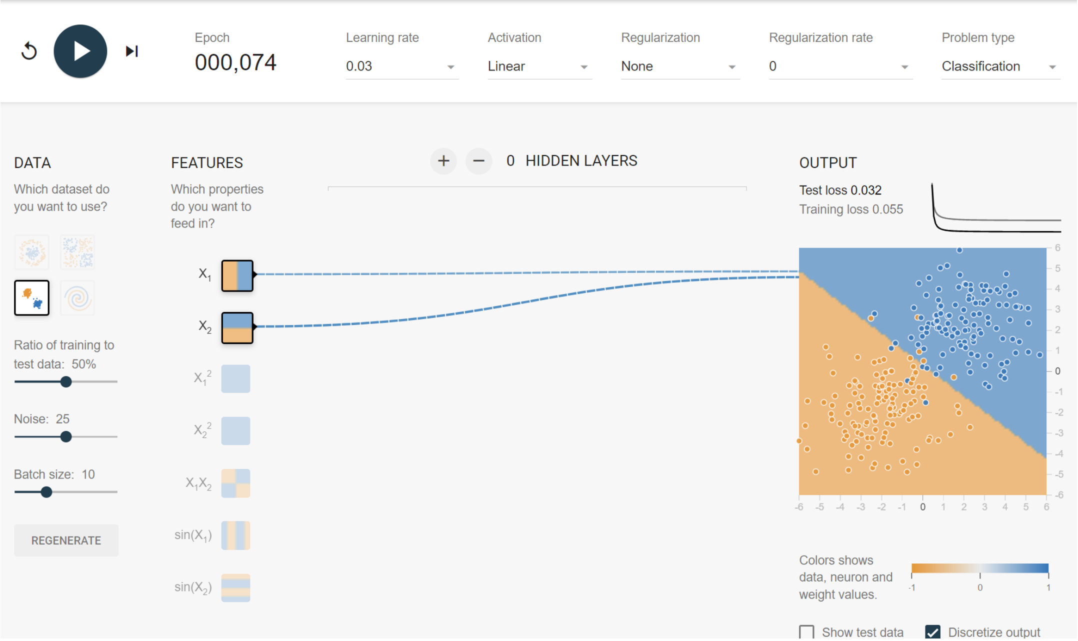
$$\boldsymbol{w}' = \boldsymbol{w} - \rho(\boldsymbol{w}^T \boldsymbol{x}_p - b_p)\boldsymbol{x}_p$$

5.3.3 確率的勾配降下法（2/3）

- ミニバッチ法
  - 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
  - これに対して確率的勾配降下法は、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〜数百個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
    - データは1エポック毎に順番をシャッフルし、エポック毎にミニバッチを構成するデータが異なるようにする
    - バッチサイズを s とすると、1エポックで n/s 回重み修正が行われる
    - 確率的勾配降下法よりも収束が安定する
    - GPU（graphics processing unit）を用いた行列の一括演算と相性がよい

5.3.3 確率的勾配降下法（3/3）

- デモ：https://playground.tensorflow.org/
  - HIDDEM LAYERS:0, Activation:Linear, DATA:Gaussian, Noise:25, Batch size:10



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
  - 確率的勾配降下法において
    - 教師信号 b を正解のときは1、不正解は0とする
    - 識別関数 $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$ の後ろに閾値論理ユニットを置き、出力を0または1に制限する

誤識別のパターン		重みの更新規則
$g(\mathbf{x}) = 0,\ \ b = 1$	$\rightarrow$	$\mathbf{w}' = \mathbf{w} + \rho \mathbf{x}$
$g(\mathbf{x}) = 1,\ \ b = 0$	$\rightarrow$	$\mathbf{w}' = \mathbf{w} - \rho \mathbf{x}$

$\rightarrow$ 誤差評価に基づく学習は、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
  - 全学習データに対して、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
  - 学習データが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
- 誤差評価に基づく学習
  - 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
  - 学習データが線形分離不可能な場合でも収束が保証されている
  - 学習データが線形分離可能な場合でも誤識別が 0 になるとは限らない
  - 全学習データに着目している

まとめ

- 誤差評価に基づく学習
  - 識別関数の出力と教師信号の二乗誤差の総和を最小化
- 解析的な解法
  - データ数が多くなると、計算量が膨大になる
- 勾配降下法
  - さまざまな工夫により、解析的な解法と同等の精度で短時間で解を求めることができる
- Jupyter notebook