目次

★13 パターン認識と機械学習 (4) — 識別（承前）

★13.4 ロジスティック回帰

識別のための手法の別の例として，ロジスティック回帰 (☆1) を紹介する．名前に「回帰」とあって紛らわしいが，「識別」のための手法である (☆2)．

☆1) ロジスティック回帰: logistic regression.

☆2) この手法の背景には，その他の多くの機械学習手法と同様に，データ等の確率的・統計的性質を考慮した問題設定や議論があるのだが，この授業では説明を省略する．

★13.4.1 2 クラス識別問題の場合の定式化

簡単のため，まずは識別すべきクラスの数が 2 に限定された場合を考える．2 つのクラスのうち一方を 'positive'（正）クラス，他方を 'negative'（負）クラスと呼ぶことにする．学習データは $\{(\boldsymbol{x}_n, y_n) | n = 1, 2, \ldots, N\}$ という形とする．$\boldsymbol{x}_n$ は D 次元のデータ（特徴ベクトル）である．y_n は，$\boldsymbol{x}_n$ の所属クラスの正解を表し，$\boldsymbol{x}_n$ が positive クラスに属すべきものなら $y_n = 1$，さもなくば $y_n = 0$ とする．

ここで，$(D+1)$ 個のパラメータ $w_0, w_1, \ldots, w_D$ を持つ次のような関数を考える．

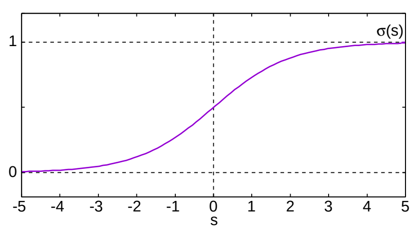
$$f(\boldsymbol{x}; w_0, w_1, \ldots, w_D) = \frac{1}{1 + \exp\left(-\left(w_0 + \sum_{d=1}^{D} w_d x_d\right)\right)} \tag{1}$$

$s = w_0 + \sum_{d=1}^{D} w_d x_d = w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ とおくと，式 (1) は

$$\sigma(s) = \frac{1}{1 + \exp(-s)} \tag{2}$$

という形をしている．この関数 $\sigma(s)$ のことを**シグモイド関数** (☆3) という．式の形と下のグラフからわかるように，この式は任意の s に対して 0 より大きく 1 より小さい値をとる．そこで，式 (1) の $f(\boldsymbol{x})$ の値が，「$\boldsymbol{x}$ は positive クラスに所属するものである」ということの確信度を表すと考えることにする．

☆3) シグモイド関数: sigmoid function. 次回の話題であるニューラルネットワークでも登場します．



このように考えると，学習データのうち $y_n = 1$ であるものについては positive クラス所属なのだから $f(\boldsymbol{x}_n)$ が 1 に近くなるように，$y_n = 0$ であるものについては逆に $f(\boldsymbol{x}_n)$ が 0 に近くなるようにうまくパラメータを調節すれば，式 (1) の値で 2 クラスの識別ができそうである．未知のデータ $\boldsymbol{x}$ に対しても，例えば $f(\boldsymbol{x}) > \frac{1}{2}$ なら positive クラス，さもなくば negative クラスと判断すればよい．

そこで，学習データからパラメータ $w_0, w_1, \ldots, w_D$ を定めるために，パラメータの「悪さ」を表す関数を定義する．ロジスティック回帰では，次式で表される**交差エントロピー** (☆ 4) を用いる．

$$H(w_0, w_1, \ldots, w_D) = \sum_{n=1}^{N} h_n \tag{3}$$

$$h_n = -y_n \log z_n - (1 - y_n) \log(1 - z_n) \tag{4}$$

ただし， $z_n = f(\boldsymbol{x}_n)$ である．この交差エントロピー H がなるべく小さくなるようなパラメータが「良い」と考えて，H を最小化するパラメータを探す．これが，2 クラス問題の場合のロジスティック回帰の考え方である．

☆ 4) 交差エントロピー: cross entropy. これが何者かは，この授業では説明しません．パターン認識や情報理論を勉強するとわかるかも．

★ 13.4.2 勾配法によるパラメータの逐次修正

上記のような問題設定は，最小二乗法による回帰の場合にも考えた．最小二乗法の場合，学習データに対する二乗誤差の和で誤差関数を定義し，それを最小にするパラメータを求めたのだった．このように，最小二乗法もロジスティック回帰も，目的とする関数（二乗誤差や交差エントロピー）を最小化する解（特定のパラメータの値）を探す問題である，というところは共通である．このような問題は，**最適化問題** (☆ 5) と呼ばれる．

最小二乗法の場合，最適化問題を解くことは比較的容易である．誤差関数をパラメータについて微分して $\mathbf{0}$ とおくと線形の連立方程式が得られるので，それを解けばよいのだった．しかし，ロジスティック回帰の場合，H の微分を $\mathbf{0}$ とおいて得られる連立方程式は非線形であり，簡単に解くことはできない．それでも，ロジスティック回帰の場合，H のパラメータに関する微分が求まるので，それを利用した最適化手法である**勾配法** (☆ 6) が使える．ここでは，その最も単純な方法の一つである**最急降下法** (☆ 7) を説明し，これによってパラメータを逐次修正していく学習アルゴリズムを構成できることを示す．

☆ 5) 最適化: optimization. 負号を付けるだけなので，最大化でも同じこと．機械学習，パターン認識，コンピュータビジョンではよく出てくるが，それ以外の幅広い分野で頻出の問題である．

☆ 6) 勾配法: gradient method.

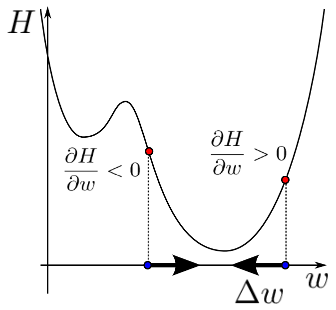
☆ 7) 最急降下法: steepest descent method.

2 クラス問題のロジスティック回帰における $(D+1)$ 個のパラメータのうちの一つを w と表して考える．H を最小にするパラメータ w の値を求めるための最急降下法の手順は次のようになる．

1. $w(t)$ の初期値（$t = 0$ における値）を適当に定める．
2. ステップ t における値 $w(t)$ を用いて微係数 $\left.\frac{\partial H}{\partial w}\right|_{w=w(t)}$ を計算し，次のステップ $t+1$ における値 $w(t+1)$ を次式により求める．

$$\begin{cases} w(t+1) = w(t) + \Delta w \\ \Delta w = -\eta \left.\frac{\partial H}{\partial w}\right|_{w=w(t)} \end{cases} \tag{5}$$

   Δw は w の修正量を表す．η は正の小さな定数である．
3. H の値が十分小さくなっていれば終了，さもなくば 2. を繰り返す．



上図からわかるように，最急降下法は「現在地での傾きを調べ，下り方向にちょっとだけ進む」ことを繰り返す手法である．初期値によっては極小解に陥って最小解にたどり着けないこともある．

上記のロジスティック回帰の場合，微係数を具体的に計算してみると，

$$\frac{\partial H}{\partial w_d} = \sum_{n=1}^{N} (z_n - y_n) x_{n,d} \qquad (d = 0, 1, \dots, D) \tag{6}$$

となる．ただし，$x_{n,d}$ はベクトル $\boldsymbol{x}_n$ の d 番目の要素（$d = 0$ のときは 1 とみなす）である．これを式 (5) に当てはめれば，学習アルゴリズムが得られる．

以下の Q は，式 (6) を導出するためのヒントである．

Q1. シグモイド関数の微分は，シグモイド関数自身を用いて

$$\frac{d\sigma(s)}{ds} = \sigma(s) \times \mathrm{hoge} \tag{7}$$

と表せる．ただし，hoge は $\sigma(s)$ を用いた式である．hoge を求めなさい．

Q2. 上記の結果を利用して，$d = 1, 2, \dots, D$ の場合の

$$\frac{\partial z_n}{\partial w_d} = \frac{\partial}{\partial w_d} \sigma \left(- \left(w_0 + \sum_{d=1}^{D} w_d x_{n,d} \right) \right) \tag{8}$$

を $x_{n,d}$ と z_n のみを用いた式で表しなさい（$z_n = \sigma(\dots)$，合成関数の微分…）．

Q3. 上記の結果を利用して，$d = 1, 2, \dots, D$ の場合の

$$\frac{\partial}{\partial w_d} (y_n \log z_n) = y_n \frac{\partial}{\partial w_d} \log z_n \tag{9}$$

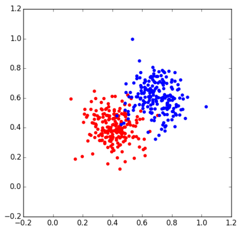
を $x_{n,d}, y_n$ と z_n のみを用いた式で表しなさい（やぱし合成関数の微分…）．

Q4. 上記の結果を利用して，$d = 1, 2, \dots, D$ の場合の $\frac{\partial h_n}{\partial w_d}$ を $x_{n,d}, y_n$ と z_n のみを用いた式で表しなさい．また，$x_{n,0} = 1$ と考えると，その式が $d = 0$ の場合にも当てはまることを確かめなさい．

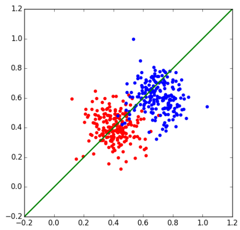
Q5. 上記の結果を利用して，式 (6) が成り立つことを示しなさい．

★ 13.4.3　2 クラスデータのロジスティック回帰の例

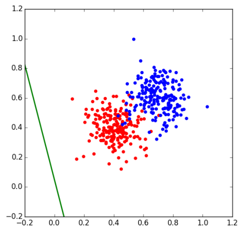
乱数を使って 2 次元の人工データを生成し，ロジスティック回帰によってそれらを 2 クラスに識別する実験を行った結果を示す．以下の図の赤と青の点は 2 つのクラスの学習データを表している．緑色は，$f(x_1, x_2; w_0, w_1, w_2) = \frac{1}{2}$ を満たす点 (x_1, x_2) の集合，すなわち赤クラスと青クラスの識別境界を表している．式 (1) からわかるようにこれらの点は $w_0 + w_1x_1 + w_2x_2 = 0$ を満たすので，境界は直線である．



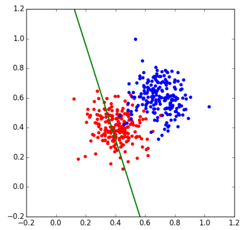
学習データ



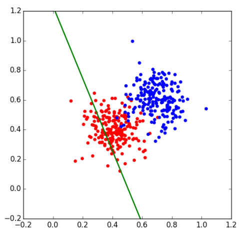
初期値 ($H = 276.4$)



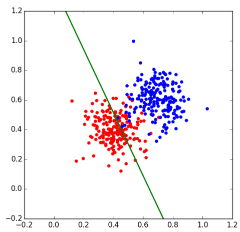
学習 1 回 ($H = 267.7$)



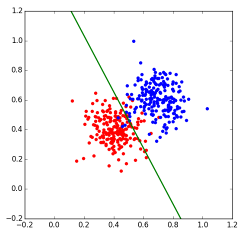
2 回 ($H = 260.0$)



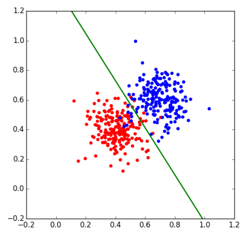
3 回 ($H = 252.9$)



5 回 ($H = 239.8$)

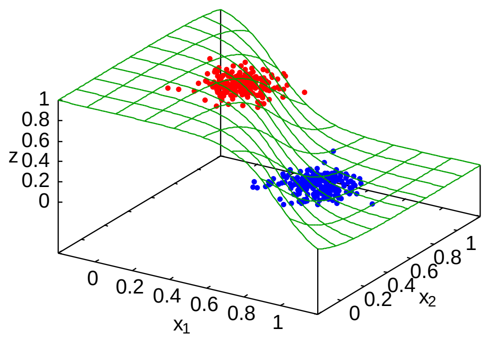


10 回 ($H = 212.5$)



100 回 ($H = 86.2$)

以下の図は，100 回学習後の $z = f(x_1, x_2)$ の表す曲面を可視化したものである．赤クラス青クラスのデータを，それぞれ $z = 1$ および $z = 0$ の平面上に重ねて表示してある．



★ 13.4.4　クラス数が 3 以上の場合のロジスティック回帰

上述の定式化では 2 クラス問題しか扱うことができないが，3 クラス以上の識別ができるように拡張することは容易である．ただし，紙面と時間の都合で，具体的な定式化や実験結果については省略する．

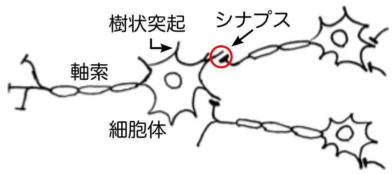
★14　パターン認識と機械学習 (5) — ニューラルネットワーク

ニューラルネットワークは，元々は脳における情報処理のモデルとして提案されたものであるが，現在では機械学習の手法として幅広く応用されている．

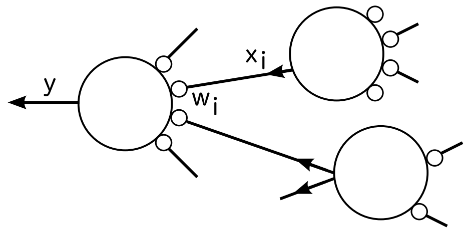
★14.1　ニューラルネットワークって？

★14.1.1　神経細胞とそのモデル

ヒトの脳には数百億の**ニューロン**（**神経細胞**, neuron）が存在している．ニューロンは信号を伝達する機能をもっており，多数のニューロンが相互につながりあって多様な情報処理を行なっている．生命維持から感情，思考にいたるまでの脳機能は，主にこれらニューロンの集団が担っているものと考えられている．



ニューロンのふるまいを思い切って単純化すると，次のようにまとめることができる．



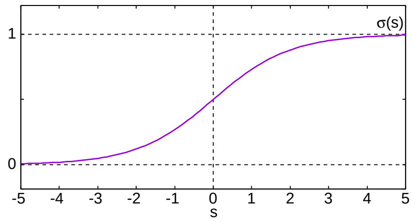
1. 他のニューロンの出力を受け取る．ただし，そのまま受け取るのではなく，ニューロン間のつながりの強さに応じて重みづけされた値を受け取る．
2. それらの和を求める
3. その値に応じて自身の出力を決める

このようなニューロンのふるまいは，次式のようにモデル化できる (☆ 8)．

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + \theta\right) \tag{10}$$

☆ 8) 1940 年代に McCulloch と Pitts が提案した．彼らのモデルではニューロンの出力は 0, 1 の二値である．

ただし， x_i はこのニューロンに信号を伝達する n 個のニューロンのうち i 番目のものの出力であり，w_i は x_i の**結合重み** (connection weight) を表すパラメータである．また，θ は**しきい値** (threshold value) と呼ばれるパラメータである．関数 $\sigma(x)$ としては，0, 1 二値の出力を与える**ステップ関数**（次式左）や (0, 1) の範囲の出力を与える**シグモイド関数**（次式右）がよく用いられる (☆ 9)．



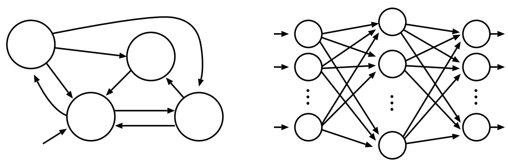
$$\sigma(s) = \begin{cases} 1 & s \geq 0 \\ 0 & otherwise \end{cases} \qquad \sigma(s) = \frac{1}{1+e^{-s}} \tag{11}$$

ニューロンの入力と出力の関係は一般に非線形であり，これらはその性質をモデル化したものとなっている．

☆ 9) ステップ関数: step function.　シグモイド関数: sigmoid function.

★14.1.2　ニューラルネットワーク，多層パーセプトロン

上記のようなニューロンモデルを複数つなぎあわせたものを**ニューラルネットワーク** (☆ 10) という．様々なつながり方のものが考えられるが，図右のようにニューロンが層を成し，層間のニューロンのみに単方向のつながりがあるタイプのものを**多層パーセプトロン** (☆ 11) と呼ぶ．



☆ 10) ニューラルネットワーク: neural network, 神経回路網とも．

☆ 11) 多層パーセプトロン: Multi-Layer Perceptron, MLP.