ロジスティック回帰

長谷川駿一

2022 年 5 月 27 日

1　概要

ロジスティック回帰とは，複数の説明変数から分析を行う**多変量解析**の一種であり，二値分類（0 か 1 か，成功か失敗か など）を行う**「分類」**に属する手法である．

2 つの状態において，**正事象** (予測したい事象) の起こる確率を求め，一般的に確率が 0.5 以上なら正事象に属し，0.5 未満であればもう一方の事象に属すると判定する．

具体例を見ていく．ある企業で，他の企業に対して訪問し，契約が成立するか成立しないかを事前に予測したいと考える．ここで，説明変数は「その企業に対しての訪問回数」としているが，他にも「資料の出来栄え」「その日の天気の具合」など，複数の要因が考えられる．目的関数は契約が成立するかしないかの二値である（図 1）．

ここで，我々が予測したいのは，この営業で契約が成立する確率であり，確率は 0 以上 1 以下の値をとるため，図 2 のようなグラフで表現できることが想像できる (ロジスティック回帰のモデル関数（予測値を出力する関数）である**シグモイド関数**は，このような形をしている)．

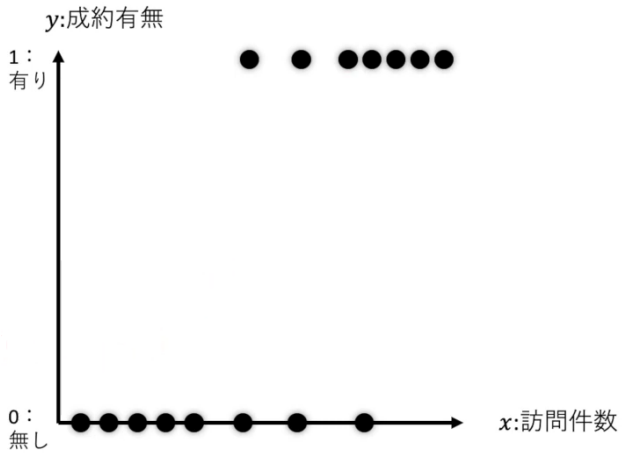

図 1　ある企業の営業データ

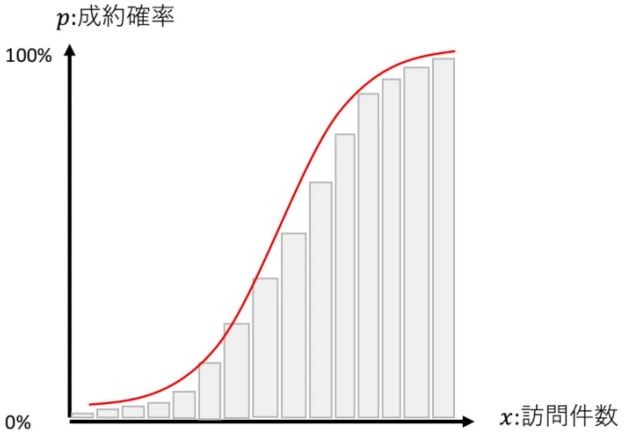

図 2　ある企業の営業データ（確率的に表したもの）

2　前提知識

2.1　シグモイド関数

ロジスティック回帰のモデル関数であるシグモイド関数を理解するために，まず**オッズ**というものを考える．オッズとは，事象の起こりやすさを表すもので，正事象の確率 p に対して $\frac{p}{1-p}$ で表すことができる．このオッズに対して対数をとったものを**ロジット**関数と呼び，以下のように定義される．

$$\mathrm{logit}(p) = \log \frac{p}{(1-p)} \tag{2.1}$$

このロジット関数だが，図 3 に示すように，実数全体に対して値をとりうる．よって，特徴量の値とロジット関数との間に線形関係を表すことができる．

$$\mathrm{logit}(p(y=1|\boldsymbol{x})) = w_0x_0 + w_1x_1 + \cdots + w_mx_m = \sum_{i=0}^{m} w_ix_i = \boldsymbol{w^t x} \tag{2.2}$$

ここで，$p(y=1|\boldsymbol{x})$ は，特徴量 $\boldsymbol{x}$ が与えられた場合にデータ点がクラス 1 に所属するという条件付き確率を示す（y は分類するクラスを表し，2 つに分類するため，1 か 0 の値をとる）．

また，w_0 は**バイアスユニット**（一次関数 $y = ax + b$ でいうところの切片 b）を表し，x_0 は 1 に設定される．

シグモイド関数とは，このロジット関数の逆関数であり，総入力 $z = \boldsymbol{w^t x}$ を入力値としている．以下のように定義され，グラフを図 4 に示す．

$$\phi(z) = \frac{1}{(1+e^{-z})} \tag{2.3}$$

よってシグモイド関数は，総入力 z を入力値，正事象（クラス 1）である確率 p を出力値とする．

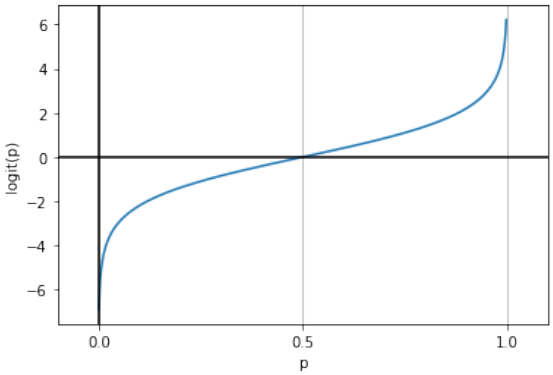

図 3　ロジット関数

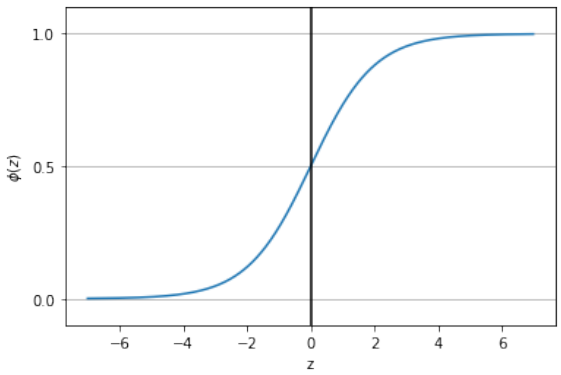

図 4　シグモイド関数

2.2　勾配降下法（コスト関数の最小化）

機械学習において，学習過程で最適化される**目的関数**のほとんどが，最小化したい**コスト関数**である．勾配降下法は，このコスト関数の最小値を見つける手法の一つである．
(ただし，コスト関数の特徴として，微分可能であること，凸関数であることが挙げられる．)

図 5 に示すように，勾配降下法は，大域的最小値に達するまで坂を下るイメージである．

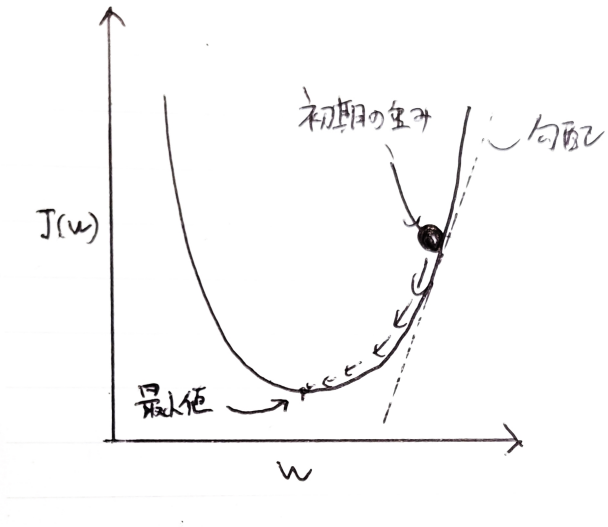

図 5　勾配降下法のイメージ

多くのモデルでは，コスト関数 J は，以下のような**誤差平方和**で定義される．この関数は微分可能かつ凸関数であり，最小値が誤差の最小をとるので，コスト関数として適切である．

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \tag{2.4}$$

ここで，$\hat{y}; \boldsymbol{w}$ である．また，添え字 i は，i 番目の訓練データの真のクラスラベル $y^{(i)}$ 及び予測されたクラスラベル $\hat{y}^{(i)}$ を表している．

勾配降下法を使って重みを更新するには，コスト関数 $J(\boldsymbol{w})$ の勾配 $\nabla J(\boldsymbol{w})$ に沿って動く．重みの変化である ∇w は，負の勾配に**学習率** η を掛けたものとして定義される（勾配を負にするのは，重みの更新は勾配と逆方向に進むためである）．

$$\boldsymbol{w} := \boldsymbol{w} + \boldsymbol{\nabla w} \tag{2.5}$$

$$\boldsymbol{\nabla w} = -\eta \boldsymbol{\nabla J}(\boldsymbol{w}) \tag{2.6}$$

重みの更新回数に対する措置として，以下のような方法がある．

i) あらかじめデータセットに対する訓練の回数（**エポック**）を決める．

ii) 重みの変化量が一定の値より小さくなったら更新をやめる．

また，学習率 η は，通常は 0.0 よりも大きく 1.0 以上の定数をとる．学習率が小さすぎると，重みの更新の変化量が小さいため，コスト関数の大域的最小値に収束させるために相当な数のエポックが必要になる．また，学習率が大きすぎると，大域的最小値を超えて更新されてしまうことがあり，発散してしまう可能性がある（図 6）．いずれにせよ最適な学習率を設定することが重要である．

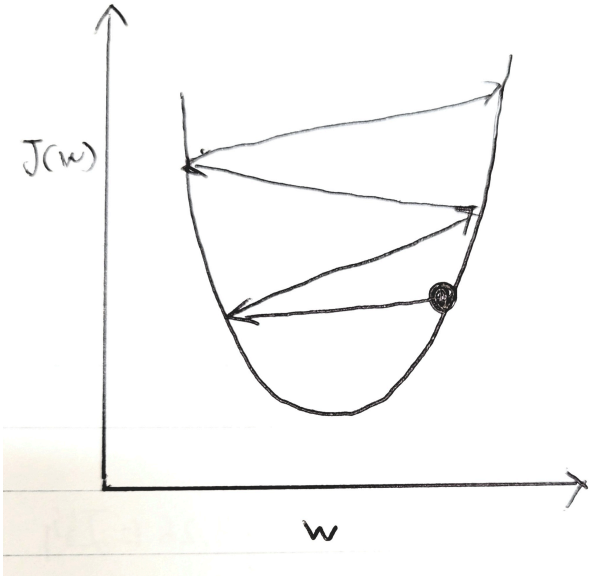

図 6　勾配降下法で発散する様子

3　ロジスティック回帰によるモデルの構築

3.1　ロジスティック回帰による分類

ロジスティック回帰による予測値 $\hat{y}$ は，クラス 1 に属する確率から分類されたクラスラベルである「0」か「1」をとるため，以下のような出力値をとる．ここで，z は総入力 $\boldsymbol{w}^{\boldsymbol{t}}\boldsymbol{x}$ である．

$$\hat{y} = \begin{cases} 1 & (\phi(z) \geq 0.5) \\ 0 & (\phi(z) < 0.5) \end{cases} \tag{3.1}$$

これは以下と等値であることがわかる．

$$\hat{y} = \begin{cases} 1 & (z \geq 0.0) \\ 0 & (z < 0.0) \end{cases} \tag{3.2}$$

3.2　ロジスティック回帰の重みの学習

ロジスティック回帰の重みの学習において，コスト関数を定義する．これは，式 (2.4) の予測値 $\hat{y}$ のところを，シグモイド関数に置き換えたものである．

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i}\left(y^{(i)} - \phi\left(z^{(i)}\right)\right)^2 \tag{3.3}$$

しかし，式 (3.3) は凸関数ではないため，勾配降下法による最小値の求め方を適応することができない．

したがって，ロジスティック回帰の重みの学習では，モデルの構築時に最大化したい**尤度**関数 $L(\boldsymbol{w})$ を定義する．尤度とは，結果から見たところの条件のもっともらしさを表す．式は次のようになる．

$$\begin{aligned} L(\boldsymbol{w}) &= P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w}) \\ &= \prod_{i=1}^{n} P\left(y^{(i)}|x^{(i)};\boldsymbol{w}\right) \\ &= \prod_{i=1}^{n}\left(\phi\left(z^{(i)}\right)\right)^{y^{(i)}}\left(1-\phi\left(z^{(i)}\right)\right)^{1-y^{(i)}} \end{aligned} \tag{3.4}$$

式 (3.4) の総乗の中身だが，$y^{(i)} = 1$ のとき $\phi(z^{(i)})$（クラス 1 である確率），$y^{(i)} = 0$ のとき $1 - \phi(z^{(i)})$（クラス 0 である確率）を出力するため，その総乗の最大値が尤度であることは直感的に理解できる．

よって，この関数の最大化をするのだが，計算をしやすくするために対数をとる．これは**対数尤**

度関数と呼ばれており，式は次のようになる．

$$
\begin{aligned}
l(\boldsymbol{w}) &= \log L(\boldsymbol{w}) \\
&= \sum_{i=1}^{n} \left[y^{(i)} \log \left(\phi \left(z^{(i)} \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - \phi \left(z^{(i)} \right) \right) \right]
\end{aligned} \tag{3.5}
$$

この関数は上に凸の関数であることが知られている．よって，対数尤度関数を最大化するためには，全体に-1 を掛け，勾配降下法による最小化をすることと同じである．したがって，コスト関数 J は以下のように定義できる．

$$
J(\boldsymbol{w}) = -\sum_{i=1}^{n} \left[y^{(i)} \log \left(\phi \left(z^{(i)} \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - \phi \left(z^{(i)} \right) \right) \right] \tag{3.6}
$$

あとは 2.2 節で示したように，コスト関数の負の勾配と学習率で重みを更新していけばよい．具体的には，要素ごとに重みを更新するため，

$$
w_j := w_j - \eta \nabla w_j \tag{3.7}
$$

$$
= w_j - \eta \frac{\partial}{\partial w_j} J(\boldsymbol{w}) \tag{3.8}
$$

$$
\nabla w_j := -\eta \frac{\partial J}{\partial w_j} = \eta \sum_{i=1}^{n} \left(y^{(i)} - \phi \left(z^{(i)} \right) \right) x_j^{(i)} \tag{3.9}
$$

このように表現できる．したがって，すべての重みを更新するため，以下のようなベクトルで表記できる．

$$
\boldsymbol{w} := \boldsymbol{w} + \boldsymbol{\nabla w} \tag{3.10}
$$

$$
\boldsymbol{\nabla w} = -\eta \boldsymbol{\nabla} \boldsymbol{J}(\boldsymbol{w}) \tag{3.11}
$$

これは，2.2 節で言及した勾配降下法の規則に相当する．

4　モデルの評価（混同行列）

最後に，ロジスティック回帰のモデルの評価方法として，**混同行列**を紹介する．混同行列は分類モデルの性能を測る指標として使われており，多クラス分類についても定義されるが，ここでは二値分類のクラス分類における混同行列について紹介する．

混同行列は，予測と実際の値について，行列形式にまとめたものである（図 7）．

		予測	
		真(Positive)	偽(Negative)
実際	真 (Positive)	真陽性 (True Positive/TP)	偽陰性 (False Negative/FN)
	偽 (Negative)	偽陽性 (False Positive /FP)	真陰性 (True Negative/TN)

図 7　混同行列のモデル

図 7 にあるように，混同行列では，以下の 4 パターンが存在する．

- **真陽性（True Positive/TP）**：予測で陽性と判断され，実際に陽性である．
- **真陰性（True Negative/TN）**：予測で陰性と判断され，実際に陰性である．
- **偽陽性（False Positive/FP）**：予測で陽性と判断されたが，実際は陰性である．
- **偽陰性（False Negative/FN）**：予測で陰性と判断されたが，実際は陽性である．

これら 4 つのパターンの値から，さまざまな評価指数が算出されるので，以下に紹介していく．

- **正解率 (Accuracy)**
  すべての予測のうち正しく分類できた割合を示す．高いほど性能が良い．
  メリット：直感的でわかりやすい．
  デメリット：Positive と Negative の出現の偏りが大きいデータには，不適切である．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **適合率 (Presision)**

  陽性と予測したもののうち，実際に陽性である割合を示す．高いほど性能が良い．

  メリット：偽陽性 FP を小さくすることを目標としており，誤って陽性と判断しては困る場合に有効．

  デメリット：偽陰性 FN が多いことが問題になるケースでは推奨されない．

$$Precision = \frac{TP}{TP + FP}$$

- **再現率 (Recall)**

  取りこぼしなく陽性であるデータを正しく陽性であると予測した割合を示す．高いほど性能が良い．

  メリット：偽陰性 FN を小さくすることを目標としており，誤って陰性と判断しては困る場合に有効．

  デメリット：偽陽性 FP が多いことが問題になるケースでは推奨されない．

$$Recall = \frac{TP}{TP + FN}$$

- **F 値**

  再現率と適合率の**調和平均**である（調和平均とは,「逆数の平均の逆数」で定義されている）．一般的に，適合率と再現率はトレードオフの関係であり，再現率と適合率を一方に偏らせずに均等に評価したい場合に F 値は適している．

$$F - measure = \frac{2 \times Precison \times Recall}{Precision + Recall}$$

1 章で挙げた例と同じく，他の企業に対して契約が成立するか成立しないかの予測と実績として，以下のような結果が得られたとする．

		予測	
		1（成立した）	0（成立しない）
実際	1 （成立した）	90	15
	0 （成立しない）	5	70

図 8　営業データの予測による混合行列

このモデルに対して，上記の評価指数を計算すると，正解率：約 89%，適合率：約 95%，再現率：約 86% ，F 値：0.89 である．よって，適合率と再現率から，このモデルでは陽性と予測したものの信憑性はかなり高く，その反面実際に陽性であるものを陰性として予測してしまう可能性が少し高いことがわかる．

5　まとめ

今回のゼミでは，ロジスティック回帰の概要からモデル構築の方法と予測方法，そしてモデルの評価方法について紹介した．以下に要点を示す．

- ロジスティック回帰は二値分類を行う分類手法の一種．
- モデル構築の際は**対数尤度**関数より**勾配降下法**の手法を用いて，最適な重み $\boldsymbol{w}$ を求める．
- 予測の際は，**シグモイド関数**を用いて確率的に分類する．
- モデルの評価方法として混同行列紹介した．二値分類を行うモデルの評価として採用される評価モデルである．