第4章 うまくいく/いかない2択の王者

幅広く使われる確率的分類「ロジスティック回帰」を動かす

佐藤 聖

幅広く使われる確率的分類「ロジスティック回帰」の特徴

ロジスティック回帰は数量データが格納された説明変数(影響を及ぼす変数)を入力して2群のカテゴリデータが格納された目的変数(予測したい変数)を求めます．目的変数は例えば「買う」または「買わない」のようにどちらかの値をとる2クラス(もしくは2択)の分類問題(2値分類問題)に利用されます．単純パーセプトロンと同じように線形分離問題を解くことができます．説明変数をS字のロジスティック曲線に回帰させ，y軸(確率)により分類します(**図1**)．

● 用途

一般的にリスク分析に利用でき，企業や医学などの幅広い分野で応用されています．品質リスク・マネジメントの手法の1つで，リスク発生の可能性を見極めるのに利用されます．生産現場の品質管理はもちろん，融資先の与信，企業の賠償リスク分析など，広く応用が効くアルゴリズムです．

例えばマーケティングでは1年間の旅行回数と1回の旅行で使う金額を説明変数として利用し，世界一周クルーズ・ツアー商品を「買う」か「買わない」か(目的変数)を求めるのに利用します．ロジット関数により買う，買わないのいずれかになるよう確率的に分類します．こうすることで線形分離できない問題も2クラスに分類できるようになります．

プログラムの流れ

08_ロジスティック回帰で分類実験.ipynb(**リスト1**)を開くと，以下の5ステップがあります．

- トレーニング・データとテスト・データの準備
- 学習と予測
- データの標準化
- 再学習
- 予測

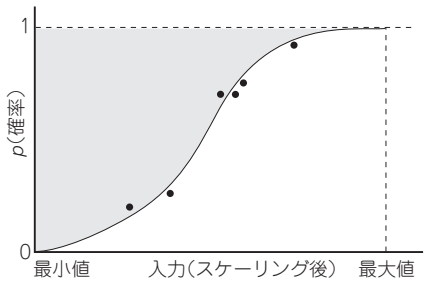

図1 確率的に分類するロジスティック曲線

トレーニング・データとテスト・データの準備

「トレーニング・データとテスト・データの準備」は，多層パーセプトロンと同じなので簡単に説明します．

● ライブラリの読み込み

ライブラリの読み込みはこれまで見てきた通りです(**リスト1**のIn[1])．scikit-learnのライブラリは`sklearn.linear_model`でロジスティック回帰の関数を，`sklearn.preprocessing`で標準化の関数を，`sklearn.metrics`で評価用の関数を読み込んでいます．

● CSVファイルの読み込み

CSVファイルの読み込みからデータ・セットの作成までは，多層パーセプトロンで使用したプログラムを流用しています(In[2])．説明は割愛しますがCSVファイルで読み込んだデータは加工する必要があるので処理の流れを確認してみてください．

● データの読み込み＆特徴からの予測

今回は2クラス分類問題ですが，データは3クラス(広告チラシ，新聞，フリーペーパ)です．特徴量からどのような予測が導き出されるのかを試してみます(In[3]，In[4]，In[5])．普通は2クラスの分類問題のときは2つのラベル付きデータを用意するかもしれませんが，どのような結果になるか試してみます．