Technical Sheet
No. 22-28

データ解析入門 15 <回帰分析の概説>

キーワード：単回帰分析、重回帰分析、多重共線性、過剰適合、多変量解析

はじめに

回帰分析は連続値を予測するための解析手法の総称であり、化学分野では検量線などに用いられてきました。近年では、検量線（単回帰分析）よりも高度な技術が提案されており、材料開発などへの適用事例が増加しています。本稿では、検量線を例にとりながら回帰分析について概説し、解析時に注意すべき過剰適合について説明します。

機械学習手法の大別

近年、人工知能（AI）や機械学習という言葉を耳にする機会が多くなりました。機械学習とは、データの背後に潜むパターンを見出すためのコンピュータアルゴリズム、あるいはその研究領域のことを指します。機械学習手法は図 1 に示す 3 つの手法に大別されることがあります。

教師なし学習の「教師」とは、コンピュータが出力すべきもののことであり、教師なし学習では出力すべきものを予め定めることなく学習が行われます。主成分分析 [1]や階層的クラスタリング [2]などは教師なし学習に該当します。

教師あり学習には、データの分類や回帰分析などが該当します。例えば、溶液濃度測定のために検量線を作成する場合、溶液濃度が出力すべきものとして予め定められています。したがって、検量線の作成は非常にシンプルな教師あり学習（単回帰分析）を行っていると言えます。

強化学習は、あるスコアが向上するような行動を見出すための機械学習手法です。プロ棋士を打ち負かすような囲碁 AI の開発例が有名です。

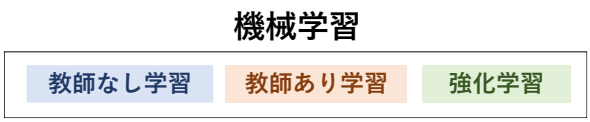

図 1　機械学習の大別

回帰分析の流れ

溶液濃度測定における検量線作成を例にとり、回帰分析の流れを説明します。まず、濃度既知の溶液を準備し、例えばクロマトグラフで測定を行うとピーク面積が得られます。このとき、溶液濃度は予測対象であり、目的変数とも呼ばれます。一方、予測に用いる入力データ（ピーク面積）は説明変数と呼ばれます。また、説明変数と目的変数はあわせて、訓練データと呼ばれます。ただし、文献等によっては異なる用語が用いられることがあります。

図 2　溶液濃度測定における単回帰分析

最小二乗法

一般的に、図 2 に示した回帰直線は最小二乗法により求められます。最小二乗法では、回帰直線と観測値との二乗誤差の和が最小になるように直線の傾き（および切片）を求めます。

解析的に計算できる最小二乗法は非常に有用ですが、問題を引き起こすこともあります。その 1 つが多重共線性の問題であり、重回帰分析などの予測精度に悪影響を及ぼします。多重共線性については後述します。

重回帰分析

上述の溶液濃度測定における検量線の例では、溶液濃度とピーク面積の間に非常に強い相関関係がありました。しかしながら、1 つの説明変数では十分な相関が得られないこともあります。

図 3 に、アルカンの沸点データセット [3]の単回帰分析結果を示します。RDKit により得た分子屈折率（MolMR）を説明変数として用いることで、ある

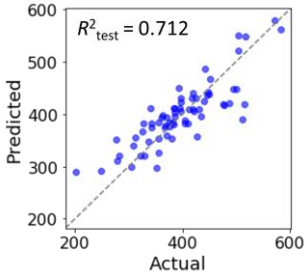

図 3　単回帰によるテストデータの予測結果

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

程度沸点を予測できていることがわかります。

さらに、複数の分子記述子を説明変数として用いると、予測精度が向上しました。ここでは、最小二乗法により回帰係数を求める重回帰分析などと呼ばれる線形回帰モデルを用いました。

回帰分析における予測精度の指標には、決定係数 R^2、相関係数、平均絶対誤差（MAE）、二乗平均平方根誤差（RMSE）などがあります。R^2 および相関係数は 1、MAE および RMSE は 0 に近ければ、予測精度が良好であるということを表します。

化学分析などでは、目的変数と説明変数の直線的な関係が認められている範囲内で検量線が作成されることが一般的です。そのため、未知データに対する予測精度（汎化性能）の検証は省略されることが多いですが、材料探索などを目的としている場合は汎化性能の評価は重要になります。図 3 に、訓練データおよびテストデータに対する予測精度を示します。沸点が不明である真の未知データでは予測精度を評価できないため、全データのうちの一部をテストデータとして汎化性能の評価に用いています（図 4）。

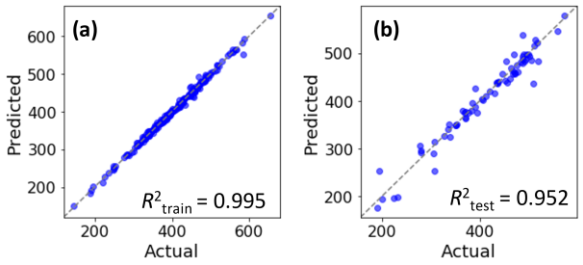

図 3　重回帰分析の結果（前処理あり）：
(a) 訓練データ、(b) テストデータ

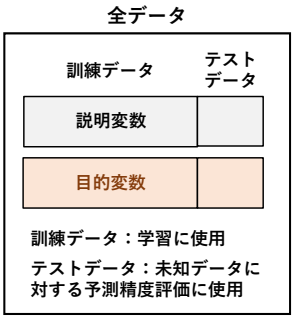

図 4　データ分割の概念図

多重共線性の問題

上述の回帰分析（図 3）では、データ前処理として相関の強い説明変数の一方を予め削除していました。なお、このような前処理を行わなくても重回帰分析は実行可能です。そこで、前処理なしの予測結果を確認してみます（図 5）。訓練データに対する予測精度は良好ですが、テストデータを上手

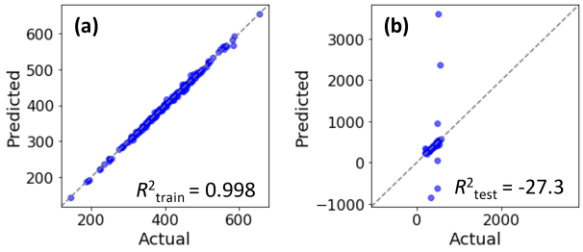

図 5　重回帰分析の結果（前処理なし）：
(a) 訓練データ、(b) テストデータ

く予測できていません。このような現象を過剰適合（オーバーフィッティング）といいます。この原因として多重共線性が挙げられます。

重回帰分析では、連立方程式 $\boldsymbol{y} = X\boldsymbol{\beta}$ の近似解を以下の式(1)から算出します。

$$\boldsymbol{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y} \qquad (1)$$

$\boldsymbol{y}$、X、$\boldsymbol{\beta}$ はそれぞれ沸点ベクトル、データ行列、回帰係数ベクトルです。多重共線性が生じると、式(1)の逆行列$(X^{\mathrm{T}}X)^{-1}$ の計算が不安定になり、極端に絶対値の大きな回帰係数が算出されることがあります。本解析においても、前処理を行わなければ極端な数値の回帰係数が算出されました（表 1）。そのため、3000 °C を超える沸点を出力するおかしな予測モデルが構築されてしまっています。

表 1　回帰係数の最小値および最大値

	最小値	最大値
前処理あり	-5.15	2.50
前処理なし	-1.77×10^{12}	1.77×10^{12}

おわりに

重回帰分析は表計算ソフトなどでも実行できる便利な線形回帰モデルですが、多重共線性などが問題となることがあります。また、変数の数がサンプル数を上回る状況では解が一意に求まりません。次稿では、このような問題を克服しうる線形回帰モデルを紹介します。

参考文献

1) 永廣卓哉：データ解析入門 2 <主成分分析によるデータの可視化>、ORIST テクニカルシート、No. 21-24 (2021)
2) 永廣卓哉：データ解析入門 5 <階層的クラスタリングの基礎>、ORIST テクニカルシート、No. 22-03 (2022)
3) E. S. Goll and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **39**, 974-983 (1999)

発行日　2023 年 2 月 1 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@orist.jp