回帰分析の簡単なまとめ

2019 年 5 月 1 日

1　回帰分析の概要

回帰分析とは、簡単に言うと、「変数 x を用いて、変数 y を説明するモデルを考えること」である。
これは、数学的に表現すると、「$y = f(x)$ となるような、関数 f を求めること」 となる。
x を独立変数 (independent variable)、y を従属変数 (dependent variable) という。
回帰分析の主なモチベーションは「予測」である。
回帰は、単回帰と重回帰に分けられ、さらにそれぞれで、線形回帰と非線形回帰とに分けることができる。線形回帰は、最適化すべきパラメータに対して、線形であるモデルのことを言い、非線形回帰は、最適化すべきパラメータに対して、非線形であるモデルのことを言う。非線形回帰は、線形回帰に比べ、モデルの表現力が上がるが、線形回帰のように、解析的にパラメータを求めることができないため、計算に時間がかかる。また、モデルの理解も線形回帰に比べ、困難である。この文書では、線形回帰のみを扱う。

2　単回帰

単回帰では、1 つの独立変数を用いて、予測を行う。

$$\hat{y} = \beta\ x + \alpha$$

上の式は、最も単純な回帰モデルである、単線形回帰 (simple linear regression) である。
パラメータ α 、β の設定には最小二乗法を使う。

n 個のデータ対　$x = (x^1, x^2, \cdot\ \cdot\ \cdot, x^n)$ と $y = (y^1, y^2, \cdot\ \cdot\ \cdot, y^n)$ があるとする。
損失関数 $L(\,\alpha\,,\,\beta\,)$ を以下のように定める。

$$L(\,\alpha\,,\,\beta\,) = \sum_{i=1}^{n} (y^i - \hat{y}^i)^2$$

観測値 y^i と、予測値 $\hat{y}^i$ の二乗誤差を取っている。絶対誤差ではなく、二乗誤差を採用しているのは、後の微分計算を楽にするためである。

$$
\begin{aligned}
L(\alpha, \beta) &= \sum_{i=1}^{n} (y^i - \hat{y}^i)^2 \\
&= \sum_{i=1}^{n} (y^i - \beta\ x^i - \alpha)^2
\end{aligned}
$$

この $L(\alpha, \beta)$ を最小にするように、 α と β を設定する。

$$
\alpha, \beta\ = argmin_{\alpha,\ \beta} L(\alpha, \beta)
$$

まず $L(\alpha, \beta)$ を α と β で、それぞれ偏微分する。

$$
\begin{aligned}
\frac{\partial\ L(\alpha, \beta)}{\partial \alpha} &= -2 \sum_{i=1}^{n} (y^i - \beta\ x^i - \alpha) \\
\frac{\partial\ L(\alpha, \beta)}{\partial \beta} &= -2 \sum_{i=1}^{n} x^i (y^i - \beta\ x^i - \alpha)
\end{aligned}
$$

よって、次の連立方程式を解くことによって、 α と β が求められる。

$$
\begin{aligned}
n\ \alpha + \beta \sum_{i=1}^{n} x^i = \sum_{i=1}^{n} y^i \\
\alpha \sum_{i=1}^{n} x^i + \beta \sum_{i=1}^{n} x^i = \sum_{i=1}^{n} x^i y^i
\end{aligned}
$$

結果は、以下のようになる。

$$
\begin{aligned}
\beta\ &= \frac{\sigma_{xy}}{\sigma_x^2} \\
\alpha\ &= \overline{y} - \beta\, \overline{x}
\end{aligned}
$$

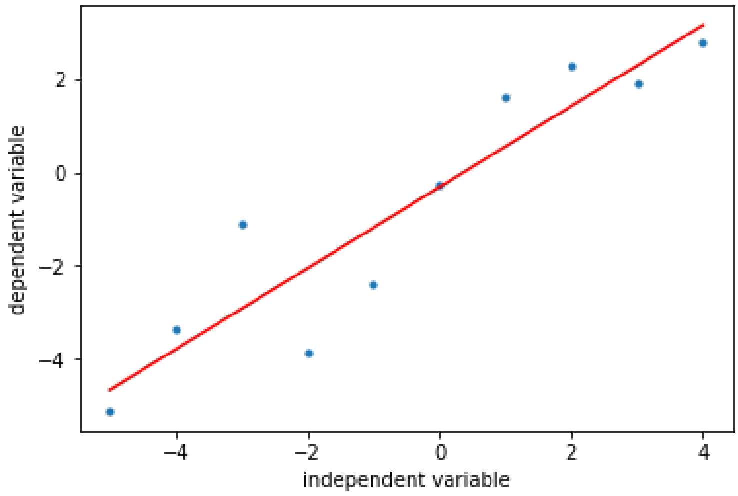

図 1　単線形回帰

3　重回帰

重回帰では、二つ以上の独立変数を用いて、予測を行う。

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \alpha$$

上の式は、重線形回帰 (multiple linear regression) である。
パラメータ α、β_1、β_2、$\cdots$、β_m の設定には、単回帰同様、最小二乗法を使う。
n 個のデータ対　$x_1 = (x_1^1, x_1^2, \cdots, x_1^n)$、$x_2 = (x_2^1, x_2^2, \cdots, x_2^n) \cdots x_m = (x_m^1, x_m^2, \cdots, x_m^n)$ と $y = (y^1, y^2, \cdots, y^n)$ があるとする。
重線形回帰での損失関数は以下のようになる。

$$\begin{aligned} L(\alpha, \{\beta_i\}_{i=1}^m) &= \sum_{i=1}^n (y^i - \hat{y}^i)^2 \\ &= \sum_{i=1}^n (y^i - \beta_1 x_1^i - \beta_2 x_2^i - \cdots - \beta_m x_m^i)^2 \end{aligned}$$

この $L(\alpha, \{\beta_i\}_{i=1}^m)$ を最小にするように、α、β_1、β_2、$\cdots$、β_m を設定する。

$$\alpha, \beta_1, \beta_2, \cdots, \beta_m = argmin_{\alpha, \{\beta_i\}_{i=1}^m} L(\alpha, \{\beta_i\}_{i=1}^m)$$

$L(\alpha, \{\beta_i\}_{i=1}^m)$ を α、β_1、β_2、$\cdots$、β_m で偏微分して、得られた連立方程式を解くという形でも、解は得られるが、それは非常に煩雑なプロセスである。行列とベクトルを用いて考えることにより、シンプルに計算が可能となる。以下、行列は大文字で、ベクトルは小文字で表す。

$$y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \\ \alpha \end{pmatrix}$$

$$X = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_m^1 & 1 \\ x_1^2 & x_2^2 & \dots & x_m^2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^n & x_2^n & \dots & x_m^n & 1 \end{pmatrix}$$

上記のように y、β、X を定めると、損失関数は以下のように書き直すことが出来る。「L2 ノルムの二乗」という表現を使っている。

$$L(\beta) = ||y - X\beta||_2^2$$

微分しやすいように $L(\beta)$ を変形しておく。

$$\begin{aligned} L(\beta) &= ||y - X\beta||_2^2 \\ &= (y - X\beta)^T(y - X\beta) \\ &= (y^T - \beta^T X^T)(y - X\beta) \\ &= (y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta) \end{aligned}$$

途中、スカラーのベクトル微分に関する公式をいくつか用いるが、それは割愛する。

$$\begin{aligned} \frac{\partial L(\beta)}{\partial \beta} &= \frac{\partial (y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta)}{\partial \beta} \\ &= -(y^T X)^T - X^T y + (X^T X + (X^T X)^T)\beta \\ &= -2X^T y + 2X^T X\beta \end{aligned}$$

よって、以下の式を解くことで、β が求められる。

$$X^T X\beta = X^T y$$

結果は以下のようになる。

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

重線形回帰でのパラメータの最適化ができた。しかし、この式は、$X^T X$ が正則でなければ、解が求まらないことを意味している。正則でないときには、正則化と呼ばれるテクニックを用いる。次の章では、正則化について、考える。

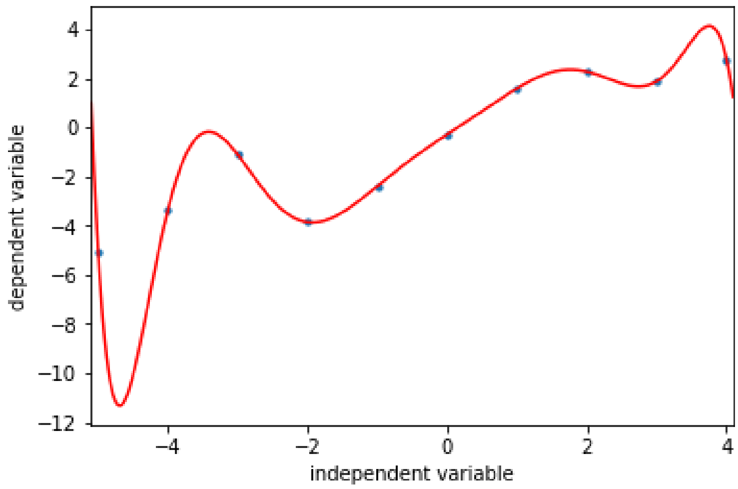

図 2　重線形回帰、9 次多項式によるモデル

4　正則化

正則化とは、先ほどの β の式において、X^TX が正則でないときに、おこなうテクニックである。まず、$L(\beta)$ に $\lambda \ ||\ \beta\ ||_2^2$ を加える。(λ はパラメータである)

$$L(\beta) = ||y - X\ \beta\ ||_2^2 + \lambda\ ||\ \beta\ ||_2^2$$

$\frac{\partial\ L(\beta)}{\partial \beta} = 0$ を計算することで、得られる β は以下のようになる。

$$\beta\ = (X^TX + \lambda\ E)^{-1}X^T\mathrm{y}$$

X^TX に $\lambda\ E$ を加わったことで、正則になっていることが分かる。これで、X^TX が正則でない場合も、β を求めることが可能になる。なお、λ は β 同様、パラメータであるが、これは、人間の手で設定すべきものである。(計算によって最適化されない。このようなパラメータをハイパーパラメータという。)

さきほど、$L(\beta)$ に $\lambda\ ||\ \beta\ ||_2^2$ を加えたが、こうすることで、β の値が大きくなるのを防ぐことができる。β の値が大きくなると、モデルは過学習 (over fitting) となる傾向にあり、β のノルムを損失関数に加えることによって、過学習を抑えることができる。今回は L2 ノルムを用いたが、一般に L p ノルムを用いることができる。L1 ノルムと L2 ノルムを用いるのが、ポピュラーであり、それぞれ、ラッソ回帰、リッジ回帰という。

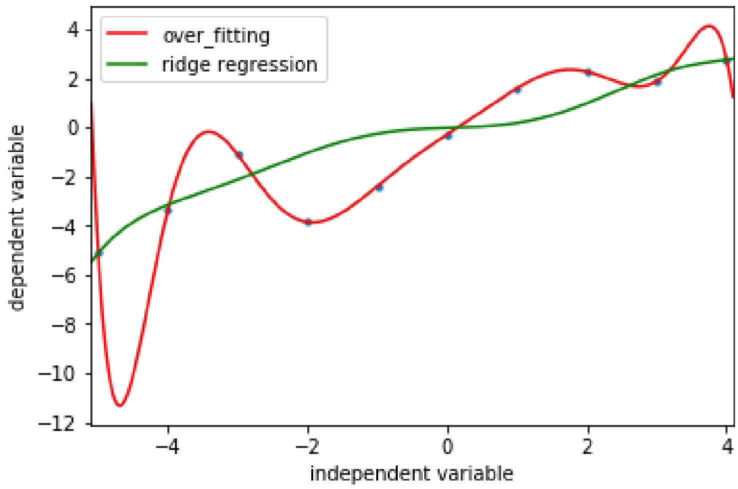

図 3　リッジ回帰、過学習が抑えられている

5　回帰モデルの評価

回帰モデルの評価には、決定係数 (R^2) を使う。決定係数にはいくつかの定義があるが、以下の定義がもっとも一般的である。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^i - \hat{y}^i)^2}{\sum_{i=1}^{n}(y^i - \bar{y})^2}$$

$\sum_{i=1}^{n}(y^i - \hat{y}^i)^2$ を残差変動、$\sum_{i=1}^{n}(y^i - \bar{y})^2$ を全変動という。R^2 の最大値は 1 であり、よっぽどひどいモデルを考えれば、負になることもあり得る。