原子力研究開発機構夏季実習生 repor

2023 年 12 月 31 日

目次

第 I 部
機械学習入門

Day1

1　線形回帰モデル

機械学習手法のうち最も基本的なモデルである線形回帰モデルについて述べる.

1.1 線形回帰モデル

線形回帰とは，入力 $\vec{x}$ と出力 y の間に線形的な関係があると仮定し，訓練データ集合 $\mathcal{D} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_N, y_N)\}$ から，線形モデル

$$y_i = \vec{w}^t \vec{x}_i = w_1(\vec{x}_i)_1 + w_2(\vec{x}_i)_2 + \cdots + w_d(\vec{x}_i)_d \tag{1}$$

の未知パラメータ $\vec{w}$ を推定し，未知入力に対する予測を行うモデルである．ここで，入力 $\vec{x}$ は d 次元ベクトル，出力はスカラーである．式中の表記 $(\vec{x}_i)_j$ は，訓練データ集合 i 番目の入力ベクトルの j 番目の要素である．これは出力 y_i を基底 x で展開し，その展開係数を推定することが目的であるとも言える．

1.2 特徴量を入れた線形回帰モデル

入力として，直接訓練集合の入力ベクトル $\vec{x}$ を入力するのではなく，入力ベクトル $\vec{x}$ を特徴量ベクトル $\vec{\phi}(\vec{x})$ へ変換し，特徴量ベクトルに対する線形モデル

$$y_i = \vec{w}^t \vec{\phi}(\vec{x}_i) \tag{2}$$

を考えることで，非線形的な関係性を表現することができる．入力ベクトルが非線形変換されているが，パラメータ $\vec{w}$ に対して，線形になっていれば，線形回帰モデルである．

■例 1: 入力と出力がともにスカラーの場合を考えよう．入力 x に対する出力 y の関係を多項式によってフィッティングする場合を考える；

$$y = \sum_{k=0}^{d} w_k x^k \tag{3}$$

ここで，x^k は x の k 乗を表し，$x^0 = 1, w_0 = b$ である．これは，入力が 1 次元であったが特徴量ベクトルに変換されたことで，次元が d に増えていることに注意せよ．このとき特徴量ベクトルは次のようにかける：

$$\vec{\phi}(x) = \begin{pmatrix} 1 \\ \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_d(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x^1 \\ x^2 \\ \vdots \\ x^d \end{pmatrix} \tag{4}$$

訓練データの入力，出力ベクトルの次元がともに d 次元のときを考える．このとき，入力ベクトルと出力ベクトルはそれぞれ $\vec{x} = (x_1\ x_2\ x_3\ \cdots x_d)$, $\vec{y} = (y_1\ y_2\ y_3\ \cdots y_d)$ である．このとき，特徴量ベクトルへの変換は，

$$\vec{y}^{\,t} = \vec{w}^{\,t}\hat{\phi} = \vec{w}^{\,t}(\vec{\phi}(x_1)\ \vec{\phi}(x_2)\ \cdots\ \vec{\phi}(x_d))$$

$$= (w_1\ w_2\ \cdots\ w_d) \begin{pmatrix} \phi_0(x_1) & \phi_0(x_2) & \cdots & \phi_0(x_d) \\ \phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_d) \\ \phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_2(x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(x_1) & \phi_n(x_2) & \cdots & \phi_n(x_d) \end{pmatrix} \tag{5}$$

1.3 多次元の線形回帰モデル

高次元の線形回帰モデル

$$\vec{y}_i = {}^t\vec{w}\vec{x}_i \tag{6}$$

を考える．このモデルに対して，訓練データ集合 $\mathcal{D} = \{(\vec{x}_1, \vec{y}_1), (\vec{x}_2, \vec{y}_2), \cdots, (\vec{x}_N, \vec{y}_N)\}$ が与えられたとき，予測誤差の二乗和

$$L(\vec{w}) = \sum_{i=1}^{N} \|\vec{y}_i - {}^t\vec{w}\vec{x}_i\|^2 \tag{7}$$

を損失関数として，損失関数が最小となるようなパラメータ $\vec{w}^*$ を求めれば良い．これは数式で書けば

$$\vec{w}^* = \arg\min_{\vec{w}} L(\vec{w}) \tag{8}$$

である．最小二乗法による最適パラメータ $\vec{w}^*$ は，訓練データ集合から行列

$$\hat{X} = \begin{pmatrix} {}^t\vec{x}_1 \\ {}^t\vec{x}_2 \\ {}^t\vec{x}_3 \\ \vdots \\ {}^t\vec{x}_N \end{pmatrix} = \begin{pmatrix} (\vec{x}_1)_1 & (\vec{x}_1)_2 & \cdots & (\vec{x}_1)_d \\ (\vec{x}_2)_1 & (\vec{x}_2)_2 & \cdots & (\vec{x}_2)_d \\ (\vec{x}_3)_1 & (\vec{x}_3)_2 & \cdots & (\vec{x}_3)_d \\ \vdots & \vdots & \ddots & \vdots \\ (\vec{x}_N)_1 & (\vec{x}_N)_2 & \cdots & (\vec{x}_N)_d \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} \tag{9}$$

を構成したとき，方程式

$$ {}^t\hat{X}\hat{X}\vec{w}^* = {}^t\hat{X}\vec{y} \tag{10} $$

を満たす．この方程式を正規方程式という．この方程式は ${}^t\hat{X}\hat{X}$ が正規行列，つまり逆行列が存在すれば $\vec{w}^* = ({}^t\hat{X}\hat{X})^{-1}\ {}^t\hat{X}\vec{y}$ と解ける．正規方程式の導出を次に示す．

$$ L(\vec{w}) = \frac{1}{2}\|\vec{y} - {}^t\vec{w}\vec{x}\|_2^2 = \frac{1}{2}(\vec{y} - \hat{X}\vec{w})^t(\vec{y} - \hat{X}\vec{w}) \tag{11} $$

$$ = \frac{1}{2}\left\{\vec{y}^t\vec{y} - \vec{y}^t(\hat{X}\vec{w}) - (\hat{X}\vec{w})^t\vec{y} + \vec{w}^t\hat{X}^t\hat{X}\vec{w}\right\} \tag{12} $$

$$ = \frac{1}{2}\left\{\vec{y}^t\vec{y} - 2(\hat{X}\vec{w})^t\vec{y} + \vec{w}^t\hat{X}^t\hat{X}\vec{w}\right\} = \frac{1}{2}\vec{y}^t\vec{y} - (\hat{X}\vec{w})^t\vec{y} + \frac{1}{2}\vec{w}^t\hat{X}^t\hat{X}\vec{w} \tag{13} $$

ここで，$(\hat{X}\vec{w})^t = \vec{w}^t\hat{X}^t$ という関係を用いて，

$$ (\hat{X}\vec{w})^t\vec{y} = \vec{w}^t\hat{X}^t\vec{y} \tag{14} $$

を用いた．

ここで，内積と標準形に関する微分の公式**??**，**??**をそれぞれ使うと，

$$ \nabla_{\vec{w}}(\vec{y}^t\hat{X}\vec{w}) = \hat{X}^t\vec{y} \tag{15} $$

$$ \nabla_{\vec{w}}(\vec{w}^t\hat{X}^t\hat{X}\vec{w}) = \hat{X}^t\hat{X}\vec{w} + (\hat{X}^t\hat{X})^t\vec{w} = 2(\hat{X}^t\hat{X})^t\vec{w} \tag{16} $$

ここで，$(\hat{X}^t\hat{X})^t = (\hat{X}^t\hat{X})^t$ であり，また，このような行列をという．したがって，$L(\vec{w})$ の $\vec{w}$ 方向についての gradient は

$$ \nabla_{\vec{w}}L(\vec{w}) = -\hat{X}^t\vec{w} + (\hat{X}^t\hat{X})\vec{y} \tag{17} $$

と求まる．よって，

$$ \nabla_{\vec{w}}L(\vec{w}) = -\hat{X}^t\vec{w} + (\hat{X}^t\hat{X})\vec{y} = 0 \tag{18} $$

$$ (\hat{X}^t\hat{X})\vec{y} = \hat{X}^t\vec{w} \tag{19} $$

1.4 具体的な実装・結果

2

最小ノルム解は n 個のデータ $\vec{y} \in \mathbb{R}^N$, $\hat{X} \in \mathbb{R}^{n \times N}$，$N$ 個のパラメータ $\vec{w}$ を用いて以下のように表された：

$$\vec{w}^* = (\hat{X}^t \hat{X})^{-1} \hat{X}^t \vec{y} \tag{20}$$

ここでは，(20) の意味について考察していく．ベクトル $\vec{y}$ はデータを表しているため，問題を解くためのヒントを表している．また，方程式の個数 n を表している．そして，パラメータ $\vec{w}$ はパラメータの数，すなわち，モデルの複雑さを表している．方程式 (20) は連立方程式の解を表現していると言える．すなわち，(20) から，自分のモデルをフィッティングさせるということは，連立方程式を解くことに帰着できることを意味している．

実際，$n = N$，すなわち，データ数とパラメータ数が等しい場合，$\hat{X}$ は正方行列となる．このとき，$\hat{X}$ には逆行列が存在して，

$$(\hat{X}^t \hat{X})^{-1} = \hat{X}^{-1} (\hat{X}^t)^{-1} \tag{21}$$

より，

$$\vec{y} = (\hat{X}^t \hat{X})^{-1} \hat{X}^t \vec{w} = \hat{X}^{-1} (\hat{X}^t)^{-1} \hat{X}^t \vec{w} = \hat{X}^{-1} \vec{w} \tag{22}$$

を得る．これは，通常の N 個の連立方程式を解く問題に帰着していることを意味している．

逆行列は通常正方行列に対して定義される．しかし，$n \neq N$ の場合，逆行列を定義できない．そこで，このような場合，行と列が異なる $(n \neq N)$ 行列に対しては疑似逆行列が定義される．最小ノルム解 (20) のなかで，$(\hat{X}^t \hat{X})^{-1} \hat{X}^t$ を疑似逆行列と呼ぶ．

2.1 行列のランク

行列のランクは特異値の数を意味する．以下のようなグラム行列の場合，固有値の数が特異値の数になる：

$$\underset{N \times n}{\hat{X}^t} \underset{n \times N}{\hat{X}} = \hat{U} \hat{\Lambda} \hat{U}^{-1} \tag{23}$$

ここで，固有値は以下で表され，内側の次元 n は固有値の数を表す：

ここから，データの数 n が少ない場合，固有値が 0 を取ってしまうことを意味する．未知数 N の数に対して，方程式 n の数が一致しなければ，連立方程式を解くことはできない．これは長方行列において，逆行列が存在しないことに対応する．

$$\hat{X} = \hat{U}\hat{\Sigma}\hat{V}^T \tag{24}$$

2.1.1 劣決定系：$n < N$

$n < N$，すなわちデータ数が未知数に対して少ない場合，ランク落ち，劣決定系という．固有値が N よりも少ないため，$(\hat{X}^t\hat{X})^{-1}$ が存在しない．そのため，再考を行う必要がある．

2.1.2 優決定系：$n > N$

$n > N$，すなわちデータ数が未知数に対して多い場合，フルランク，優決定系という．固有値が N よりも多いため，$(\hat{X}^t\hat{X})^{-1}$ が存在する．このとき，疑似逆行列 $(\hat{X}^t\hat{X})^{-1}\hat{X}^t$ を計算でき，近似解を求めることができる．データがあるため，連立方程式を解くことができる．

2.2 リッジ回帰

逆行列が存在する，しないを議論することは連立方程式が解けるか，解けないかを議論している．逆行列が存在するようにすることを正則化という．ここでは正則化を行うための手法の一つであるリッジ回帰について学ぶ.[*1]

データ数が少ない，$n < N$ の場合に用いられる正則化の一つがリッジ回帰である．このとき，グラム行列の固有値にはゼロ固有値が存在していた．リッジ回帰の基本的アイデアは，このゼロ固有値を取り得る対角成分に何かを埋めることである．

ランク落ちが起こる別の場合，定数倍になっている場合，線形独立ではなくなって，ランク落ちが起こる場合がある．

リッジ回帰はパラメータ空間で，二乗誤差がリッジ上の箇所で同じ解をもつために，解がユニークに定まらない場合に，

現在,

リッジ回帰を行うためにパラメータベクトルのノルム

$$\vec{w}||_2^2 = \vec{w}^t\vec{w} \tag{25}$$

[*1] データ拡張　データをある程度まねできるモデルからデータを生成し水増しすること本当はうまくいっていないのではないかという意見があるデータの素性を知っているデータを吐き出せるくらいモデルがデータを知っているじゃあ，拡張しなくてもいいのでは？

に係数 $\lambda > 0$ をかけて，二乗誤差に加えたものの最小化を考えてみる：

$$L(\vec{w}) = \min_{\vec{w}} \left\{ \frac{1}{2} \|\vec{y} - {}^t\vec{w}\vec{x}\|_2^2 + \frac{\lambda}{2} \|\vec{w}\|_2^2 \right\} \tag{26}$$

$$f(\vec{w}) = \frac{1}{2}\vec{w}^t\vec{w} = \frac{1}{2}\sum_k w_k \tag{27}$$

より，

$$\nabla_{\vec{w}} f(\vec{w}) = \vec{w} \tag{28}$$

したがって，

$$\nabla_{\vec{w}} L(\vec{w}) = (\hat{X}^t\hat{X})\vec{w}^* - \hat{X}^t\vec{y} + \lambda\vec{w} \tag{29}$$

$$= -\hat{X}^t(\vec{y} - \hat{X}\vec{w}^*) + +\lambda\vec{w} = 0 \tag{30}$$

よって，

$$-\hat{X}^t(\vec{y} - \hat{X}\vec{w}^*) + +\lambda\vec{w} = 0 \tag{31}$$

$$(\hat{X}^t\hat{X} + \lambda\hat{I})\vec{w} = \hat{X}^t\vec{y} \tag{32}$$

を得る．これを $\vec{w}$ について解くと，

$$\vec{w}^* = (\hat{X}^t\hat{X} + \lambda\hat{I})^{-1}\hat{X}^t\vec{y} \tag{33}$$

を得る．これがリッジ回帰の解である．$\hat{X}^t\hat{X}$ は正則ではない項を意味し，$\lambda\hat{I}$ は正則化する項を表している．

リッジ回帰を使うことでパラメータ $\vec{w}$ の絶対値を小さくすることができる．また，結果的に行列 $\hat{X}^t\hat{X}$ の対角成分に小さな α を足すことで，ゼロ固有値をなくし，逆行列の計算を安定化させることができる．

■$L2$ ノルム

$$\|\vec{a}\|_2^2 = \vec{a}^t\vec{a} = a_1^2 + a_2^2 + \cdots \tag{34}$$

$$\|\vec{a}\|_2 = \sqrt{\vec{a}^t\vec{a}} = \sqrt{a_1^2 + a_2^2 + \cdots} \tag{35}$$

■内積の微分

■標準形の微分

参考文献