線形回帰とロジスティック回帰

中田和秀

東京工業大学 工学院 経営工学系

機械学習入門

http://www.iee.e.titech.ac.jp/~nakatalab/text/lecture.html

概要

ここでは線形モデルによって回帰や判別を行う方法について説明する。これらは昔から使われてきた単純な方法であるが、計算が容易である、解析しやすい、人間が理解しやすいなどの優れた特徴を持っている。

目次：

記号の使い方：

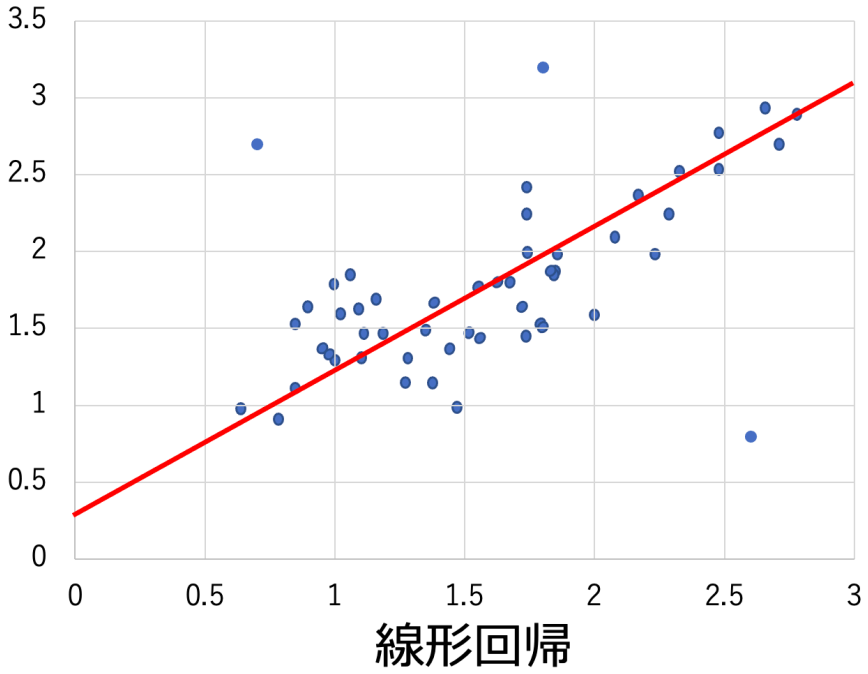
- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

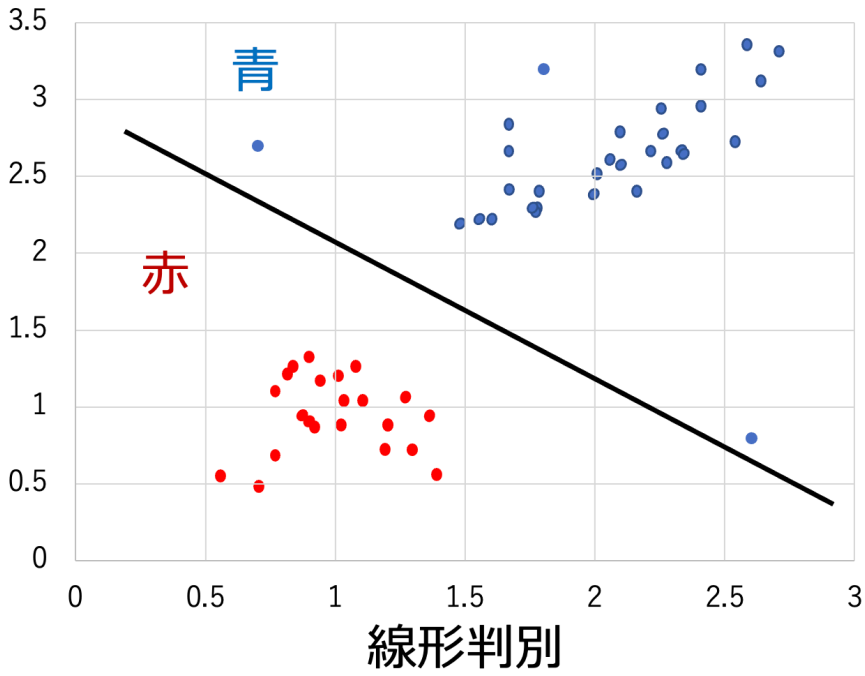
線形モデル

線形モデルによる教師あり学習

- 回帰：線形回帰
- 判別・分類：ロジスティック回帰

線形性：$f(\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2) = \alpha_1 f(\boldsymbol{x}_1) + \alpha_2 f(\boldsymbol{x}_2)$





線形モデルの特徴

特徴

- 計算（学習）や理論的解析が容易
- 人間が理解しやすい
- 複雑な現象（非線形性を持つもの）を説明できない

データ： $\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}, \quad \boldsymbol{x}_d \in \mathbb{R}^n,\ y_d \in \mathbb{R}$

線形関数（アフィン関数）：$y = \boldsymbol{w}^T\boldsymbol{x} + b$

$$\tilde{\boldsymbol{x}} := \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} \in \mathbb{R}^{1+n},\ \tilde{\boldsymbol{w}} := \begin{pmatrix} b \\ \boldsymbol{w} \end{pmatrix} \in \mathbb{R}^{1+n} \text{ とすると、} \boldsymbol{w}^T\boldsymbol{x} + b = \tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}$$

以下では、１次元を増やした $\tilde{\boldsymbol{x}}_d, \tilde{\boldsymbol{w}}$ を $\boldsymbol{x}_d, \boldsymbol{w} \in \mathbb{R}^n$ として考える。

線形回帰

予測器

$$y \simeq \hat{y} = \boldsymbol{w}^T \boldsymbol{x} \qquad (\text{パラメタは } \boldsymbol{w})$$

誤差関数

予測誤差の２乗

$$L(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2 \qquad y\text{: 目標} \quad \hat{y}\text{: 予測値}$$

学習

$$\min_{\boldsymbol{w}} \sum_{d \in [D]} (y_d - \boldsymbol{w}^T \boldsymbol{x}_d)^2$$

目的関数の $1/2D$ は省略

線形回帰は最小２乗法や重回帰とも呼ばれる

学習アルゴリズム

$$\boldsymbol{X} := \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_D^T \end{pmatrix} \in \mathbb{R}^{D \times n}, \quad \boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{pmatrix} \in \mathbb{R}^D \quad \text{とする。}$$

学習の行列表現

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

$\boldsymbol{X}$ が列フルランク $\mathrm{rank}(\boldsymbol{X}) = n$ と仮定

学習アルゴリズム

最適解は解析的に与えられる

$$\boldsymbol{w}^* := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

証明

$$F(\boldsymbol{w}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

最適解の必要条件 $\nabla F(\boldsymbol{w}^*) = \boldsymbol{0}$ を考える。

$$\nabla F(\boldsymbol{w}^*) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{0}$$
$$\iff \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^T\boldsymbol{y}$$
$$\iff \boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (\because \mathrm{rank}(\boldsymbol{X}) = n \text{ より逆行列は存在})$$

また、$\nabla^2 F(\boldsymbol{w})$ は半正定値である。なぜなら、任意の $\boldsymbol{v} \in \mathbb{R}^n$ に対して

$$\boldsymbol{v}^T\nabla^2 F(\boldsymbol{w})\boldsymbol{v} = 2\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} = 2(\boldsymbol{X}\boldsymbol{v})^T(\boldsymbol{X}\boldsymbol{v}) \geq 0$$

よって、$F(\boldsymbol{w})$ は凸関数となり、$\nabla F(\boldsymbol{w}^*) = \boldsymbol{0}$ は最適解の必要十分条件

以上より、$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ が最適解。

確率モデルに基づいた解析

データ $\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$ が次の確率モデルから生成されたサンプル（標本）とする。

$$y_d = \boldsymbol{w}^T \boldsymbol{x}_d + \epsilon_d \quad (d \in [D])$$
$$\iff \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$\boldsymbol{\epsilon}$ はノイズ（測定誤差など）を表す確率変数

ここで確率変数 $\boldsymbol{\epsilon} \in \mathbb{R}^D$ が従う仮定として、次のものを考える。

誤差項に対する仮定

- 仮定 1：$\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$
- 仮定 2：$\mathbb{V}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}$　　$\boldsymbol{I}$ は単位行列
- 仮定 3：$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

※仮定 3 は仮定 1,2 を含むより強い仮定

不偏推定量

不偏推定量

仮定１のもとで、推定量 $\boldsymbol{w}^* := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ は不偏推定量である。

$$\mathbb{E}[\boldsymbol{w}^*] = \boldsymbol{w}$$

証明：

$$\begin{aligned}
\boldsymbol{w}^* &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\boldsymbol{w}^*] &= \mathbb{E}[\boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] \\
&= \boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{w} \qquad (\because \text{仮定 } \mathbf{1})
\end{aligned}$$

この意味でこの計算に正当性があるといえる

最良線形不偏推定量

仮定 2 のもとで

$$\begin{aligned}
\mathbb{V}[\boldsymbol{w}^*] &= \mathbb{V}[\boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{V}[\boldsymbol{\epsilon}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \qquad (\because \text{仮定 } \mathbf{2}) \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

ガウス・マルコフの定理

仮定 1,2 のもとで、$\boldsymbol{w}^* := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ は、$\boldsymbol{y}$ の線形変換で表される不偏推定量 $\tilde{\boldsymbol{w}}$ の中で最も分散が小さい。

$$\mathbb{V}[\tilde{\boldsymbol{w}}] \succeq \mathbb{V}[\boldsymbol{w}^*]$$

※ $\boldsymbol{A} \succeq \boldsymbol{B}$ は、$\boldsymbol{A} - \boldsymbol{B}$ が半正定値行列であることを意味する

この性質を最良線形不偏推定量や BLUE(best linear unbiased estimztor) と呼ぶ

証明

推定量 $\tilde{\boldsymbol{w}}$ として、$\boldsymbol{y}$ の線形変換で表せるものを考える。

$$\tilde{\boldsymbol{w}} = \boldsymbol{C}\boldsymbol{y}$$

任意の $\boldsymbol{w}$ に対して不偏推定量であるためには、

$$\boldsymbol{w} = \mathbb{E}[\tilde{\boldsymbol{w}}] = \mathbb{E}[\boldsymbol{C}(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon})] = \boldsymbol{C}\boldsymbol{X}\boldsymbol{w}$$

である必要があるため、$\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$.

$$\mathbb{V}[\hat{\boldsymbol{w}}] = \mathbb{V}[\boldsymbol{C}(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon})] = \boldsymbol{C}\mathbb{V}[\boldsymbol{\epsilon}]\boldsymbol{C}^T = \boldsymbol{C}\sigma^2\boldsymbol{I}\boldsymbol{C}^T = \sigma^2\boldsymbol{C}\boldsymbol{C}^T$$

$$\mathbb{V}[\boldsymbol{w}^*] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$\boldsymbol{P} := \boldsymbol{C} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ とすると、

$$\begin{aligned}
\boldsymbol{P}\boldsymbol{P}^T &= (\boldsymbol{C} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)(\boldsymbol{C} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\
&= \boldsymbol{C}\boldsymbol{C}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{C}^T - \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \boldsymbol{C}\boldsymbol{C}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \boldsymbol{C}\boldsymbol{C}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

証明（続き）

つまり、

$$\mathbb{V}[\tilde{\boldsymbol{w}}] - \mathbb{V}[\boldsymbol{w}^*] = \sigma^2 \boldsymbol{C}\boldsymbol{C}^T - \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \sigma^2 \boldsymbol{P}\boldsymbol{P}^T$$

となるため、

$$\mathbb{V}[\tilde{\boldsymbol{w}}] \succeq \mathbb{V}[\boldsymbol{w}^*]$$

この意味で $\boldsymbol{w}^*$ は線形不偏推定量 $\hat{\boldsymbol{w}}$ の中で最も分散が小さいといえる。

また、$\mathrm{Tr}(\mathbb{V}[\tilde{\boldsymbol{w}}]) \geq \mathrm{Tr}(\mathbb{V}[\boldsymbol{w}^*])$ が成り立つ。そして、

$$\begin{aligned}
\mathrm{Tr}(\mathbb{V}[\tilde{\boldsymbol{w}}]) &= \mathrm{Tr}(\mathbb{E}[(\tilde{\boldsymbol{w}} - \boldsymbol{w})(\tilde{\boldsymbol{w}} - \boldsymbol{w})^T]) = \mathbb{E}[\mathrm{Tr}((\tilde{\boldsymbol{w}} - \boldsymbol{w})(\tilde{\boldsymbol{w}} - \boldsymbol{w})^T)] \\
&= \mathbb{E}[\mathrm{Tr}((\tilde{\boldsymbol{w}} - \boldsymbol{w})^T(\tilde{\boldsymbol{w}} - \boldsymbol{w}))] = \mathbb{E}\left[\|\tilde{\boldsymbol{w}} - \boldsymbol{w}\|^2\right].
\end{aligned}$$

同様に、$\mathrm{Tr}(\mathbb{V}[\boldsymbol{w}^*]) = \mathbb{E}\left[\|\boldsymbol{w}^* - \boldsymbol{w}\|^2\right]$ である。よって、

$$\mathbb{E}\left[\|\tilde{\boldsymbol{w}} - \boldsymbol{w}\|^2\right] \geq \mathbb{E}\left[\|\boldsymbol{w}^* - \boldsymbol{w}\|^2\right].$$

これは $\boldsymbol{w}^*$ は真の値 $\boldsymbol{w}$ からの距離が小さいことを意味する。

最尤推定法

仮定 3 では次の式に従う。

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

このとき、

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$$

尤度関数

$$l(\boldsymbol{y}; \boldsymbol{w}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right\}$$

対数尤度関数

$$\log l(\boldsymbol{y}; \boldsymbol{w}) = -\frac{D}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

対数尤度関数の最大化は線形回帰と等価

$$\max_{\boldsymbol{w}} \ \log l(\boldsymbol{y}; \boldsymbol{w})$$

つまり、線形回帰は最尤推定法とみなせる

決定係数

決定係数 R^2

- 線形回帰のデータへの当てはまりの良さの尺度
- $0 \leq R^2 \leq 1$ で、大きいほど当てはまりがよい。
- R^2 は寄与率とも言う。また、その平方根を重相関係数と言う。

定義：

- $\mathbb{V}[y]$： データ $\{y_d\}_{d\in[D]}$ の標本分散
- $\mathbb{V}[\hat{y}]$： 予測値 $\{\hat{y}_d\}_{d\in[D]}$ の標本分散
- E^2： 平均 2 乗予測誤差 $\dfrac{1}{D}\sum_{d\in[D]}(y_d-\hat{y}_d)^2$
- $\rho(y,\hat{y})$： $\{(y,\hat{y}_d)\}_{d\in[D]}$ の標本相関係数

決定係数

$$R^2 := \frac{\mathbb{V}[\hat{y}]}{\mathbb{V}[y]} = 1 - \frac{E^2}{\mathbb{V}[y]} = \rho(y,\hat{y})^2$$

証明

$$\boldsymbol{w}^* := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$\hat{\boldsymbol{y}} := \boldsymbol{X}\boldsymbol{w}^*$$
$$\boldsymbol{\epsilon} := \boldsymbol{y} - \hat{\boldsymbol{y}}$$

まず、

$$\boldsymbol{X}^T\boldsymbol{\epsilon} = \boldsymbol{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{0}$$

が成り立つ。このとき、

$$\hat{\boldsymbol{y}}^T\boldsymbol{\epsilon} = (\boldsymbol{X}\boldsymbol{w}^*)^T\boldsymbol{\epsilon} = \boldsymbol{w}^{*T}\boldsymbol{X}^T\boldsymbol{\epsilon} = 0 \tag{1}$$

である。また、$\boldsymbol{X}^T$ の１行目は $\boldsymbol{e}^T := (1, 1, \cdots, 1)$ なので、$\boldsymbol{X}^T\boldsymbol{\epsilon} = \boldsymbol{0}$ の一番目は

$$\boldsymbol{e}^T\boldsymbol{\epsilon} = 0 \tag{2}$$

となる。このとき、次の計算より y と $\hat{y}$ の標本平均は等しくなる。

$$\mathbb{E}[y] = \frac{1}{D}\boldsymbol{e}^T\boldsymbol{y} = \frac{1}{D}\boldsymbol{e}^T(\hat{\boldsymbol{y}} + \boldsymbol{\epsilon}) = \frac{1}{D}\boldsymbol{e}^T\hat{\boldsymbol{y}} = \mathbb{E}[\hat{y}]$$

証明（続き）

$$
\begin{aligned}
\mathbb{V}[y] &= \frac{1}{D}(\boldsymbol{y} - \mathbb{E}[y]\boldsymbol{e})^T(\boldsymbol{y} - \mathbb{E}[y]\boldsymbol{e}) \\
&= \frac{1}{D}(\hat{\boldsymbol{y}} + \boldsymbol{\epsilon} - \mathbb{E}[\hat{y}]\boldsymbol{e})^T(\hat{\boldsymbol{y}} + \boldsymbol{\epsilon} - \mathbb{E}[\hat{y}]\boldsymbol{e}) \\
&= \frac{1}{D}(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e})^T(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e}) + \frac{2}{D}(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e})^T\boldsymbol{\epsilon} + \frac{1}{D}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\
&= \mathbb{V}[\hat{y}] + E^2 \qquad (\because (1)(2))
\end{aligned}
$$

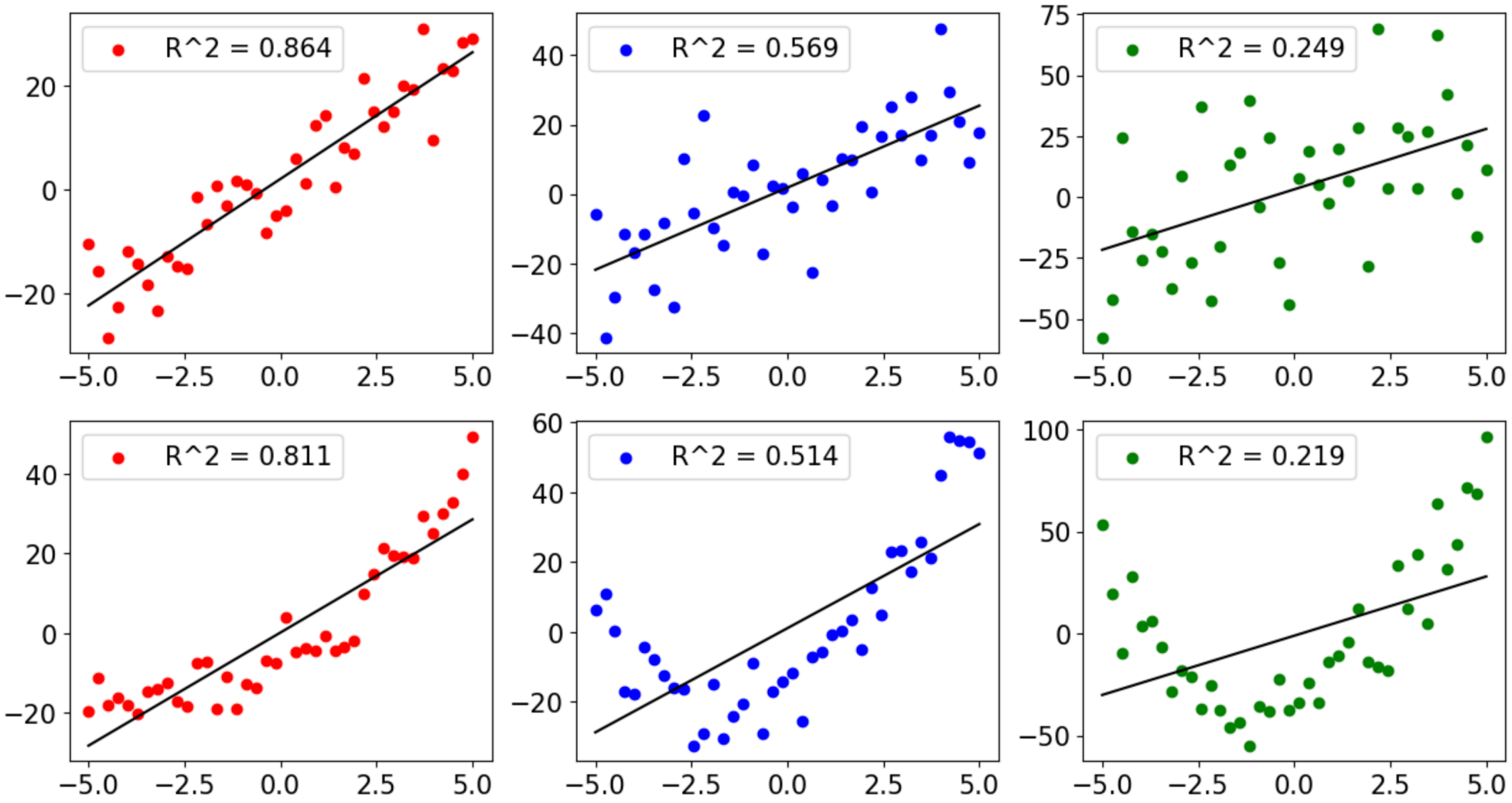
よって、 $\dfrac{\mathbb{V}[\hat{y}]}{\mathbb{V}[y]} = \dfrac{\mathbb{V}[y] - E^2}{\mathbb{V}[y]} = 1 - \dfrac{E^2}{\mathbb{V}[y]}$

$$
\begin{aligned}
\mathsf{Cov}[y, \hat{y}] &:= \frac{1}{D}(\boldsymbol{y} - \mathbb{E}[y]\boldsymbol{e})^T(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e}) \\
&= \frac{1}{D}(\hat{\boldsymbol{y}} + \boldsymbol{\epsilon} - \mathbb{E}[\hat{y}]\boldsymbol{e})^T(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e}) \\
&= \frac{1}{D}(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e})^T(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e}) + \frac{1}{D}\boldsymbol{\epsilon}^T(\hat{\boldsymbol{y}} - \mathbb{E}[\hat{y}]\boldsymbol{e}) \\
&= \mathbb{V}[\hat{y}] \qquad (\because (1)(2))
\end{aligned}
$$

よって、 $\rho(y, \hat{y})^2 := \dfrac{\mathsf{Cov}[y, \hat{y}]^2}{\mathbb{V}[y]\mathbb{V}[\hat{y}]} = \dfrac{\mathbb{V}[\hat{y}]^2}{\mathbb{V}[y]\mathbb{V}[\hat{y}]} = \dfrac{\mathbb{V}[\hat{y}]}{\mathbb{V}[y]}$

決定係数

上段は一次関数、下段は二次関数の予測。R^2 だけでは全てを判断できない。



正則化項の導入

線形回帰を行う上での問題

- 多重共線性：
  $\mathrm{rank}(\boldsymbol{X}) < n$ のとき、$\boldsymbol{w}^*$ が計算できない。
- 数値的不安定：
  $\boldsymbol{X}^T\boldsymbol{X}$ に 0 に近い固有値があると、$\boldsymbol{w}^*$ の分散が大きくなる。
  $$\mathbb{V}[\boldsymbol{w}^*] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
- 過学習：
  テストデータで予測精度が悪化することもある

これらの問題を解決するために、正則化項を加えることが有効

正則化付き学習

正則化付き学習

- Ridge 回帰

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|^2$$

- LASSO 回帰

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|_1$$

- Elastic Net 回帰

$$\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|^2$$

$\|\boldsymbol{w}\|_1 := \sum_{i=1}^n |w_i|$

- ハイパーパラメータ $\lambda \geq 0$ は正則化の強さを表す定数
- LASSO 回帰は最適解 $\boldsymbol{w}^*$ の各要素が 0 になりやすい。このため、予測器の解釈がし易くなる。
- Elastic net 回帰は、Ridge 回帰と LASSO 回帰の中間の性質がある。

学習アルゴリズム

Ridge 回帰： $\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$

解析解が得られ、$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

LASSO 回帰： $\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1$

この問題は $\boldsymbol{w}, \boldsymbol{s} \in \mathbb{R}^n$ を変数とする次のような 2 次計画問題となる。

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{s}} \quad & \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{e}^T\boldsymbol{s} \\ \text{s.t.} \quad & -\boldsymbol{s} \le \boldsymbol{w} \le \boldsymbol{s}. \end{aligned}$$

Elastic Net 回帰： $\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|^2$

この問題は $\boldsymbol{w}, \boldsymbol{s} \in \mathbb{R}^n$ を変数とする次のような 2 次計画問題となる。

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{s}} \quad & \boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda_2\boldsymbol{I})\boldsymbol{w} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \lambda_1\boldsymbol{e}^T\boldsymbol{s} \\ \text{s.t.} \quad & -\boldsymbol{s} \le \boldsymbol{w} \le \boldsymbol{s}. \end{aligned}$$

2 次計画問題の汎用解法で解くことも可能だが、専用解法の方が高速

LASSO の学習アルゴリズム

反復解法で解く。k 回目の反復点を $\boldsymbol{w}_k$ とする。

目的関数 $f(\boldsymbol{w}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1$ に対し、次を満たす関数 $g_k(\boldsymbol{x})$ を考える。

- $f(\boldsymbol{w}_k) = g_k(\boldsymbol{w}_k)$
- $\nabla f(\boldsymbol{w}_k) = \nabla g_k(\boldsymbol{w}_k)$
- $f(\boldsymbol{w}) \leq g_k(\boldsymbol{w}) \quad (\forall \boldsymbol{w} \in \mathbb{R}^n)$

そして、$f(\boldsymbol{w})$ を最小化する代わりに、次のように反復点を更新する。

$$\boldsymbol{w}_{k+1} := \operatorname*{argmin}_{\boldsymbol{w}} g_k(\boldsymbol{w})$$

- $f(\boldsymbol{w}_k)$ $(k = 1, 2, \cdots)$ は非増加（減少 or 変わらない）

$$\begin{aligned} f(\boldsymbol{w}_k) &= g_k(\boldsymbol{w}_k) \\ &\geq g(\boldsymbol{w}_{k+1}) \\ &\geq f(\boldsymbol{w}_{k+1}) \end{aligned}$$

LASSO の学習アルゴリズム

- $f(\boldsymbol{w}_k)$ が変わらない時、最適解が得られる

$$
\begin{aligned}
& f(\boldsymbol{w}_{k+1}) = f(\boldsymbol{w}_k) \\
\Longrightarrow\ & \boldsymbol{w}_k \text{ が } \boldsymbol{g}_k(\boldsymbol{w}) \text{ の最適解} \\
\Longrightarrow\ & \nabla \boldsymbol{g}_k(\boldsymbol{w}_k) = \boldsymbol{0} \\
\Longrightarrow\ & \nabla \boldsymbol{f}(\boldsymbol{w}_k) = \boldsymbol{0} \\
\Longrightarrow\ & \boldsymbol{w}_k \text{ は } \boldsymbol{f}(\boldsymbol{w}) \text{ の最適解}
\end{aligned}
$$

$g_k(\boldsymbol{x})$ には次の性質も必要

- $\min_{\boldsymbol{w}} g_k(\boldsymbol{w})$ が効率よく計算できる。
  １反復辺りの計算時間を減らすため
- $g_k(\boldsymbol{w})$ は $f(\boldsymbol{w})$ のタイトなバウンドである。
  最適解を得るまでの反復回数を減らすため

$g_k(\boldsymbol{w})$ の構築

$$
\begin{aligned}
f(\boldsymbol{w}) :=& \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1 \\
=& \|(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) - \boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k)\| + \lambda\|\boldsymbol{w}\|_1 \\
=& (\boldsymbol{w} - \boldsymbol{w}_k)^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) \\
& + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) + \lambda\|\boldsymbol{w}\|_1
\end{aligned}
$$

と変形できる。ここで、次のように $g_k(\boldsymbol{w})$ を定義する（ρ は $\boldsymbol{X}^T\boldsymbol{X}$ の最大固有値）。

$$
\begin{aligned}
g_k(\boldsymbol{w}) :=& \rho(\boldsymbol{w} - \boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) \\
& + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) + \lambda\|\boldsymbol{w}\|_1
\end{aligned}
$$

このとき、次の関係が成り立つ。

- $f(\boldsymbol{w}_k) = g_k(\boldsymbol{w}_k)$
- $\nabla f(\boldsymbol{w}_k) = \nabla g_k(\boldsymbol{w}_k)$
- $f(\boldsymbol{w}) \leq g_k(\boldsymbol{w}) \quad (\forall \boldsymbol{w} \in \mathbb{R}^n)$

証明

$$
\begin{aligned}
&f(\boldsymbol{w}_k) = g_k(\boldsymbol{w}_k) \\
&= -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T \boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) + \lambda \|\boldsymbol{w}\|_1
\end{aligned}
$$

$$
\begin{aligned}
&\nabla f(\boldsymbol{w}_k) \\
&= -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) + \lambda \nabla \|\boldsymbol{w}\|_1 \\
&= \nabla g_k(\boldsymbol{w}_k)
\end{aligned}
$$

$\rho \boldsymbol{I} \succeq \boldsymbol{X}^T \boldsymbol{X}$ より、

$$
\begin{aligned}
&\rho(\boldsymbol{w} - \boldsymbol{w}_k)^T (\boldsymbol{w} - \boldsymbol{w}_k) - (\boldsymbol{w} - \boldsymbol{w}_k)^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) \\
&= (\boldsymbol{w} - \boldsymbol{w}_k)^T (\rho \boldsymbol{I} - \boldsymbol{X}^T \boldsymbol{X})(\boldsymbol{w} - \boldsymbol{w}_k) \\
&\geq 0
\end{aligned}
$$

$g_k(\boldsymbol{w})$ の変形

$\boldsymbol{c} := -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)$ とすると、次のように変形できる。

$$
\begin{aligned}
g_k(\boldsymbol{w}) &= \rho(\boldsymbol{w} - \boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) + \boldsymbol{c}^T(\boldsymbol{w} - \boldsymbol{w}_k) + \lambda\|\boldsymbol{w}\|_1 \quad + \text{定数} \\
&= \sum_{i\in[n]} \rho\left(w_i - [\boldsymbol{w}_k]_i\right)^2 + c_i\left(w_i - [\boldsymbol{w}_k]_i\right) + \lambda|w_i| \quad + \text{定数} \\
&= \sum_{i\in[n]} \rho\left(w_i - [\boldsymbol{w}_k]_i + \frac{1}{\rho}c_i\right)^2 + \lambda|w_i| \quad + \text{定数} \\
&= \sum_{i\in[n]} \rho\left(w_i - \left([\boldsymbol{w}_k]_i - \frac{1}{\rho}c_i\right)\right)^2 + \lambda|w_i| \quad + \text{定数}
\end{aligned}
$$

要素ごとに分割できるため、簡単に最適解が得られる。

反復点の更新

$$\min_{w_i} \quad \rho\left(w_i - \left([\boldsymbol{w}_k]_i - \frac{1}{\rho}c_i\right)\right)^2 + \lambda|w_i|$$

の最適解は $w_i^* = S_{\lambda/2\rho}\left([\boldsymbol{w}_k]_i - \frac{1}{\rho}c_i\right)$ となる（次スライド）。

ただし、$S_\alpha(x)$ はソフト閾値関数で

$$S_\alpha(x) := \begin{cases} x-\alpha & (\alpha \le x) \\ 0 & (-\alpha < x < \alpha) \\ x+\alpha & (x \le -\alpha) \end{cases}$$

よって、次の式によって反復点を更新する。

$$\forall i \in [n], \quad [\boldsymbol{w}_{k+1}]_i := S_{\lambda/2\rho}\left([\boldsymbol{w}_k]_i - \frac{1}{\rho}c_i\right)$$

この解法を ISTA(iterative shrinkage threshold algorithm) と呼ぶ。

これに改良を加えた FISTA という解法もある。

証明

$y \in \mathbb{R},\ \alpha \geq 0$ を定数としたとき、関数 $f(x) := \dfrac{1}{2}(x-y)^2 + \alpha|x|$ の最小解は、

$$x^* = S_\alpha(y) := \begin{cases} y-\alpha & (\alpha \leq y) \\ 0 & (-\alpha < y < \alpha) \\ y+\alpha & (y \leq -\alpha) \end{cases} \quad \text{となる。}$$

証明：

- $-\alpha < y < \alpha$ の場合、

$$\begin{aligned} f(x) = \frac{1}{2}(x-y)^2 + \alpha|x| &\geq \frac{1}{2}(x^2 - 2|x||y| + y^2) + \alpha|x| \\ &\geq \frac{1}{2}(x^2 - 2\alpha|x| + y^2) + \alpha|x| = \frac{1}{2}(x^2+y^2) \geq \frac{1}{2}y^2 = f(0) \end{aligned}$$

  最後の不等号は $x = 0$ のときのみ等号が成立する。

- $\alpha \leq y$ の場合、$\tilde{f}(x) := \dfrac{1}{2}(x-y)^2 + \alpha x$ とすると、$\forall x,\ f(x) \geq \tilde{f}(x)$ より、

$$f(y-\alpha) \geq \min_x f(x) \geq \min_x \tilde{f}(x) = \tilde{f}(y-\alpha) = f(y-\alpha)$$

  $\min_x \tilde{f}(x)$ は $x^* = y - \alpha$ が唯一の最小点。

- $y \leq \alpha$ の場合も、$\tilde{f}(x) := \dfrac{1}{2}(x-y)^2 - \alpha x$ とすれば、上の証明と同様

Elastic Net 回帰の学習アルゴリズム

$$f(\boldsymbol{w}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|^2$$

次の関数を考える。

$$\begin{aligned} g_k(\boldsymbol{w}) := (\rho + \lambda_2)(\boldsymbol{w} - \boldsymbol{w}_k)^T(\boldsymbol{w} - \boldsymbol{w}_k) - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{w}_k) \\ + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_k) + \lambda_1\|\boldsymbol{w}\|_1 \end{aligned}$$

後は LASSO と同じように考えることができる。

最適解は次の計算による反復で求めることができる。

$$\forall i, \quad [\boldsymbol{w}_{k+1}]_i := S_{\lambda_1/2(\rho+\lambda_2)}\left([\boldsymbol{w}_k]_i - \frac{1}{\rho + \lambda_2}c_i\right)$$

確率モデルに基づいた解析

ベイズの定理を使うと、データ $\mathcal{D} = \{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$ のもとでのパラメタ $\boldsymbol{w}$ の確率は、次のようになる。

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathcal{D})}$$

- $P(\boldsymbol{w})$：パラメタ $\boldsymbol{w}$ の事前確率
- $P(\boldsymbol{w}|\mathcal{D})$：データ $\mathcal{D}$ が与えられた状況でのパラメタ $\boldsymbol{w}$ の事後確率

MAP(maximum a posteriori) 推定

事後確率 $P(\boldsymbol{w}|\mathcal{D})$ が最大となるパラメタ $\boldsymbol{w}$ を採用する

$$\max_{\boldsymbol{w}} \ \log P(\boldsymbol{w}|\mathcal{D}) \quad \Longleftrightarrow \quad \max_{\boldsymbol{w}} \ \log P(\mathcal{D}|\boldsymbol{w}) + \log P(\boldsymbol{w})$$

Ridge 回帰と MAP 推定

仮定

- パラメタの事前分布：$\boldsymbol{w} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$
- 仮定 3：$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

$$
\begin{aligned}
\log P(\mathcal{D}|\boldsymbol{w}) + \log P(\boldsymbol{w}) &= -\frac{D}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\
&\quad - \frac{n}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\boldsymbol{w}^T\boldsymbol{w} \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) - \frac{1}{2\tau^2}\boldsymbol{w}^T\boldsymbol{w} + \text{定数}
\end{aligned}
$$

この関数の最大化は、λ を適当に定めた $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$ の最小化と等価

つまり、Ridge 回帰は上記の仮定の元での MAP 推定とみなせる

LASSO 回帰と MAP 推定

別の仮定

- パラメタの事前分布：$\boldsymbol{w}$ の各要素が平均 0 のラプラス分布に従う

  ※　平均 μ, 分散 $2b^2$ であるラプラス分布の密度関数

$$f(x; \mu, b) = \frac{1}{2b} \exp\left\{ -\frac{|x - \mu|}{b} \right\}$$

- 仮定 3：$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$,　$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

この仮定の元での、事後確率の最大化は LASSO 回帰における学習と等価

つまり、LASSO 回帰は MAP 推定とみなせる。

Elastic Net 回帰と等価な事前分布を作ることも可能だが、名前が付いているようなありふれた分布ではない

判別・分類

ここでは２クラスの判別・分類を考える。

$$\text{データ：}\ \{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}, \qquad \boldsymbol{x}_d \in \mathbb{R}^n,\ y_d \in \{0,1\}$$

ロジスティック回帰

- ロジスティック回帰では確率的判別モデルを考える

  $y=1$ の確率を計算する。　　$f:\ \mathbb{R}^n \to [0,1]$
- 確率（実数）を予測するので回帰
- 新しいデータ $\boldsymbol{x}$ に対し、予測した確率が閾値（例えば 0.5）以上のとき $y=1$、そうでないとき $y=0$ と予測
- 多クラス判別に拡張した多項ロジスティック回帰というのもある

ロジスティック回帰

予測器

線形関数 $\boldsymbol{w}^T\boldsymbol{x}$ とシグモイド関数を使った回帰。パラメタは $\boldsymbol{w}$

$$\hat{y} = \sigma(\boldsymbol{w}^T\boldsymbol{x})$$

$\hat{y} \in [0, 1]$：予測値（確率）

シグモイド関数：

$$\sigma(a) := \frac{1}{1+\exp(-a)} = \frac{\exp a}{1+\exp a}$$

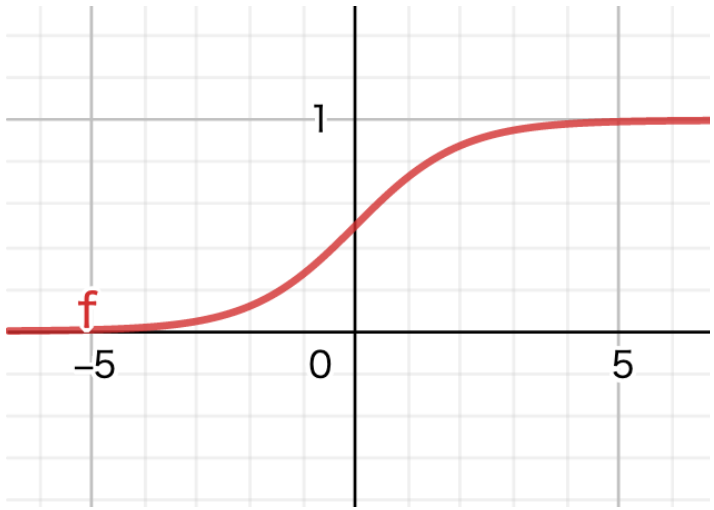
$$0 < \sigma(a) < 1$$

$$\lim_{a\to-\infty} \sigma(a) = 0, \quad \lim_{a\to\infty} \sigma(a) = 1$$

$$a_1 < a_2 \implies \sigma(a_1) < \sigma(a_2)$$

$$\sigma(-a) = 1 - \sigma(a)$$

$$\sigma'(a) = \sigma(a)(1-\sigma(a))$$



判別面

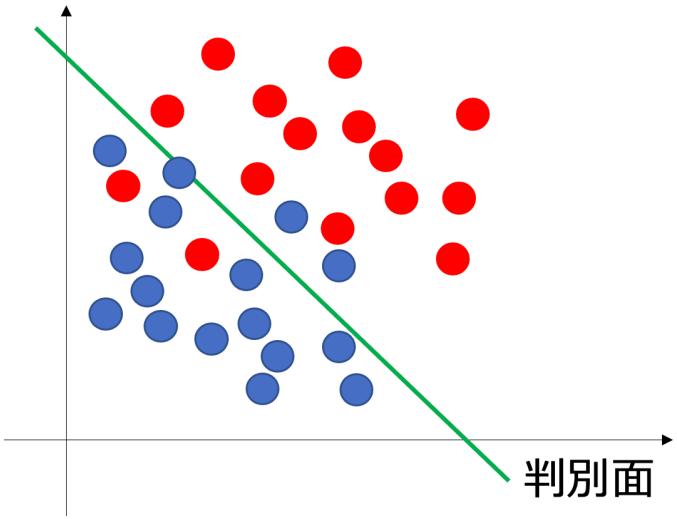
$\boldsymbol{x}$ に対する判別 (α は閾値)

$$\text{判別結果} \quad = \begin{cases} 1 & (\sigma(\boldsymbol{w}^T\boldsymbol{x}) \geq \alpha) \\ 0 & (\sigma(\boldsymbol{w}^T\boldsymbol{x}) \leq \alpha) \end{cases}$$

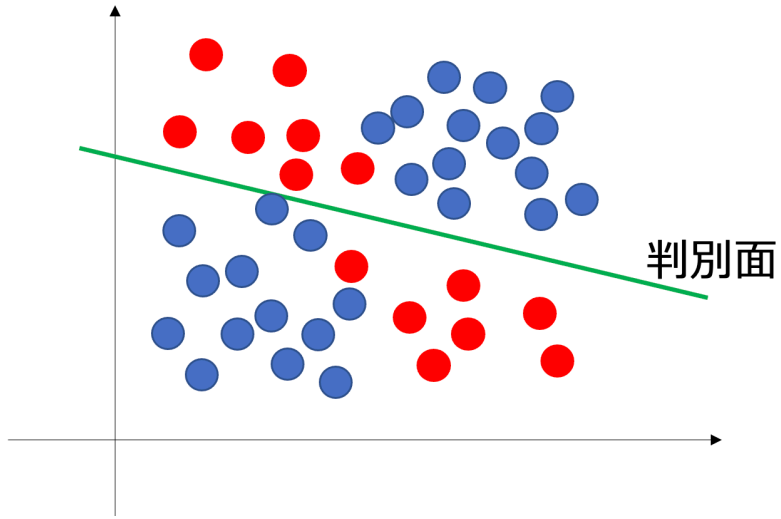
$\sigma(a)$ は単調増加関数なので、判別面は超平面となる。

$$C_1 := \{\boldsymbol{x} \mid \boldsymbol{w}^T\boldsymbol{x} \geq \sigma^{-1}(\alpha)\}, \quad C_0 := \{\boldsymbol{x} \mid \boldsymbol{w}^T\boldsymbol{x} < \sigma^{-1}(\alpha)\}$$

線形分離**可能**



線形分離**不可能**



学習

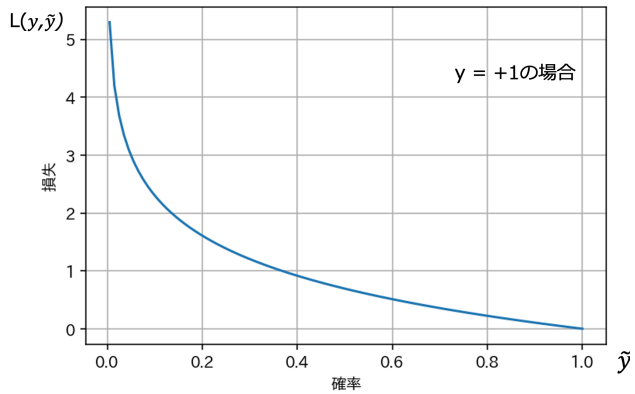
誤差関数

（2 クラス判別における）クロスエントロピー誤差

$$L(y, \hat{y}) := -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$y \in \{0, 1\}$：目標，　$\hat{y} \in [0, 1]$：予測値

$y = 1$ のとき



ロジスティック回帰における学習

$$\min_{\boldsymbol{w}} \sum_{d \in [D]} L(y_d, \sigma(\boldsymbol{w}^T \boldsymbol{x}_d))$$

確率モデルに基づいた解析 1

確率 $\hat{y}$ で $y = 1$、それ以外（確率 $1 - \hat{y}$）で $y = 0$ となるベルヌーイ試行で、データが生成されたとする。

ただし $\hat{y} = \sigma(\boldsymbol{w}^T \boldsymbol{x})$

尤度関数

$$l(\boldsymbol{y}; \boldsymbol{w}) := \prod_{d \in [D]} (\hat{y}_d)^{y_d} \times (1 - \hat{y}_d)^{1 - y_d}$$

対数尤度関数

$$\begin{aligned} \log l(\boldsymbol{y}; \boldsymbol{w}) &:= \sum_{d \in [D]} \{ y_d \log \hat{y}_d + (1 - y_d) \log(1 - \hat{y}_d) \} \\ &= - \sum_{d \in [D]} L(y_d, \hat{y}_d) \end{aligned}$$

クロスエントロピー誤差を用いた学習は最尤推定に相当する

$$\max_{\boldsymbol{w}} \ \log l(\boldsymbol{y}; \boldsymbol{w}) \quad \Longleftrightarrow \quad \min_{\boldsymbol{w}} \ \sum_{d \in [D]} L(y_d, \sigma(\boldsymbol{w}^T \boldsymbol{x}_d))$$

確率モデルに基づいた解析 2

データ y_d, 予測値 $\hat{y}_d$ とする

経験分布 P

$$p(y) := \begin{cases} 1 & (y = y_d) \\ 0 & (y \neq y_d) \end{cases}$$

予測した分布 Q

$$q(y) := \begin{cases} \hat{y}_d & (y = 1) \\ 1 - \hat{y}_d & (y = 0) \end{cases}$$

とすると、

$$\begin{aligned} KL(P||Q) := & \sum_{x_i \in X} p(x_i) \log \frac{p(x_i)}{q(x_i)} \\ = & - y_d \log \hat{y}_d - (1 - y_d) \log(1 - \hat{y}_d) \\ = & L(y_d, \hat{y}_d) \end{aligned}$$

クロスエントロピー誤差は、経験分布と予測分布の距離（KL ダイバージェンス）

別の見方

予測器

線形関数 $\boldsymbol{w}^T\boldsymbol{x}$ を使った回帰。パラメタは $\boldsymbol{w}$

$$\hat{y} := \boldsymbol{w}^T\boldsymbol{x}$$

誤差関数

ロジスティック損失：　$y \in \{1, -1\}$　とする

$$L(y, \hat{y}) := \log(1 + \exp(-y\hat{y}))$$

この予測器と誤差関数を用いた学習は次のようになる。

学習

$$\min_{\boldsymbol{w}} \sum_{d \in [D]} L(y_d, \boldsymbol{w}^T\boldsymbol{x}_d)$$

これは、ロジスティック回帰における学習と一致する。

等価性の証明

$$\hat{y} := \boldsymbol{w}^T \boldsymbol{x}, \qquad L(y, \hat{y}) := \log(1 + \exp(-y\hat{y}))$$

より、次の関係が成り立つ。

$$L(y_d, \boldsymbol{w}^T \boldsymbol{x}_d) = \begin{cases} \log\left\{1 + \exp(+\boldsymbol{w}^T \boldsymbol{x}_d)\right\} & (y_d = +1) \\ \log\left\{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x}_d)\right\} & (y_d = -1) \end{cases}$$

一方、

$$\hat{y} := \sigma(\boldsymbol{w}^T \boldsymbol{x}_d), \qquad \sigma(a) := \frac{1}{1 + \exp(-a)},$$
$$L(y, \hat{y}) := -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

より、次の関係が成り立つ。

$$L(y_d, \sigma(\boldsymbol{w}^T \boldsymbol{x}_d)) = -y_d \log \frac{1}{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x}_d)} - (1 - y_d) \log \frac{1}{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_d)}$$
$$= \begin{cases} \log\left\{1 + \exp(+\boldsymbol{w}^T \boldsymbol{x}_d)\right\} & (y_d = 1) \\ \log\left\{1 + \exp(-\boldsymbol{w}^T \boldsymbol{x}_d)\right\} & (y_d = 0) \end{cases}$$

よって、両者は等しい。

学習アルゴリズム

目的関数を $F(\boldsymbol{w})$ とする。このとき、

$$\nabla F(\boldsymbol{w}) = \boldsymbol{X}^T(\hat{\boldsymbol{y}} - \boldsymbol{y})$$
$$\nabla^2 F(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X}$$

ここで、$\hat{y}_d = \sigma(\boldsymbol{w}^T \boldsymbol{x}_d)$, $D_{dd} = \hat{y}_d(1 - \hat{y}_d) > 0 \quad (d \in [D])$

式変形は次のスライド

性質など

- $\nabla^2 F(\boldsymbol{w})$ は半正定値行列より、$F(\boldsymbol{w})$ は凸関数
- $\nabla F(\boldsymbol{w}) = \boldsymbol{0}$ が最適解の必要十分条件だが、解析的に解くことはできない
- $F(\boldsymbol{w})$ の最小点を Newton 法で求める

式変形

$$
\begin{aligned}
\hat{y}_d &= \sigma(\boldsymbol{w}^T \boldsymbol{x}_d) \\
\frac{\partial \hat{y}_d}{\partial \boldsymbol{w}} &= \hat{y}_d(1-\hat{y}_d)\boldsymbol{x}_d^T \\
F(\boldsymbol{w}) &= \sum_{d\in[D]} \{-y_d \log \hat{y}_d - (1-y_d)\log(1-\hat{y}_d)\} \\
\nabla F(\boldsymbol{w}) = \left(\frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{w}}\right)^T &= \sum_{d\in[D]} \left\{ -\frac{y_d}{\hat{y}_d}\hat{y}_d(1-\hat{y}_d)\boldsymbol{x}_d + \frac{1-y_d}{(1-\hat{y}_d)}\hat{y}_d(1-\hat{y}_d)\boldsymbol{x}_d \right\} \\
&= \sum_{d\in[D]} (-y_d + \hat{y}_d)\boldsymbol{x}_d \\
&= \boldsymbol{X}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) \\
\nabla^2 F(\boldsymbol{w}) &= \sum_{d\in[D]} \hat{y}_d(1-\hat{y}_d)\boldsymbol{x}_d\boldsymbol{x}_d^T \\
&= \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}
\end{aligned}
$$

Newton 法

Newton 法の適用

$$\begin{aligned}\boldsymbol{w}_{k+1} :=& \boldsymbol{w}_k - \left(\nabla^2 F(\boldsymbol{w}_k)\right)^{-1} \nabla F(\boldsymbol{w}_k) \\ =& \boldsymbol{w}_k - (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\hat{\boldsymbol{y}} - \boldsymbol{y}) \\ =& (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} \left(\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X} \boldsymbol{w}_k - \boldsymbol{X}^T (\hat{\boldsymbol{y}} - \boldsymbol{y}) \right) \\ =& (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{D} \left(\boldsymbol{X} \boldsymbol{w}_k - \boldsymbol{D}^{-1} (\hat{\boldsymbol{y}} - \boldsymbol{y}) \right)\end{aligned}$$

$\boldsymbol{z} := \boldsymbol{X} \boldsymbol{w}_k - \boldsymbol{D}^{-1} (\hat{\boldsymbol{y}} - \boldsymbol{y})$ とおくと、

$$\boldsymbol{w}_{k+1} := (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{z}.$$

これは、内積が $(\boldsymbol{x} \cdot \boldsymbol{y}) := \boldsymbol{x}^T \boldsymbol{D} \boldsymbol{y}$ で定義される内積空間上での直交射影

学習アルゴリズム

学習アルゴリズム

反復重み付き最小２乗法　(iterative reweighted least squares method; IRLS)

1. $k := 1$ として、$\boldsymbol{w}_1$ の初期値を定める。
2. $\hat{y}_d := \sigma(\boldsymbol{w}_k^T \boldsymbol{x}_d) \qquad d \in [D]$
   $D_{dd} := \hat{y}_d(1 - \hat{y}_d) \quad d \in [D]$
3. $\boldsymbol{z} := \boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{D}^{-1}(\hat{\boldsymbol{y}} - \boldsymbol{y})$
4. $\boldsymbol{w}_{k+1} := (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{z}$
5. 終了条件を満たしていれば終了。
   そうでなければ、$k := k + 1$ として (2) に戻る