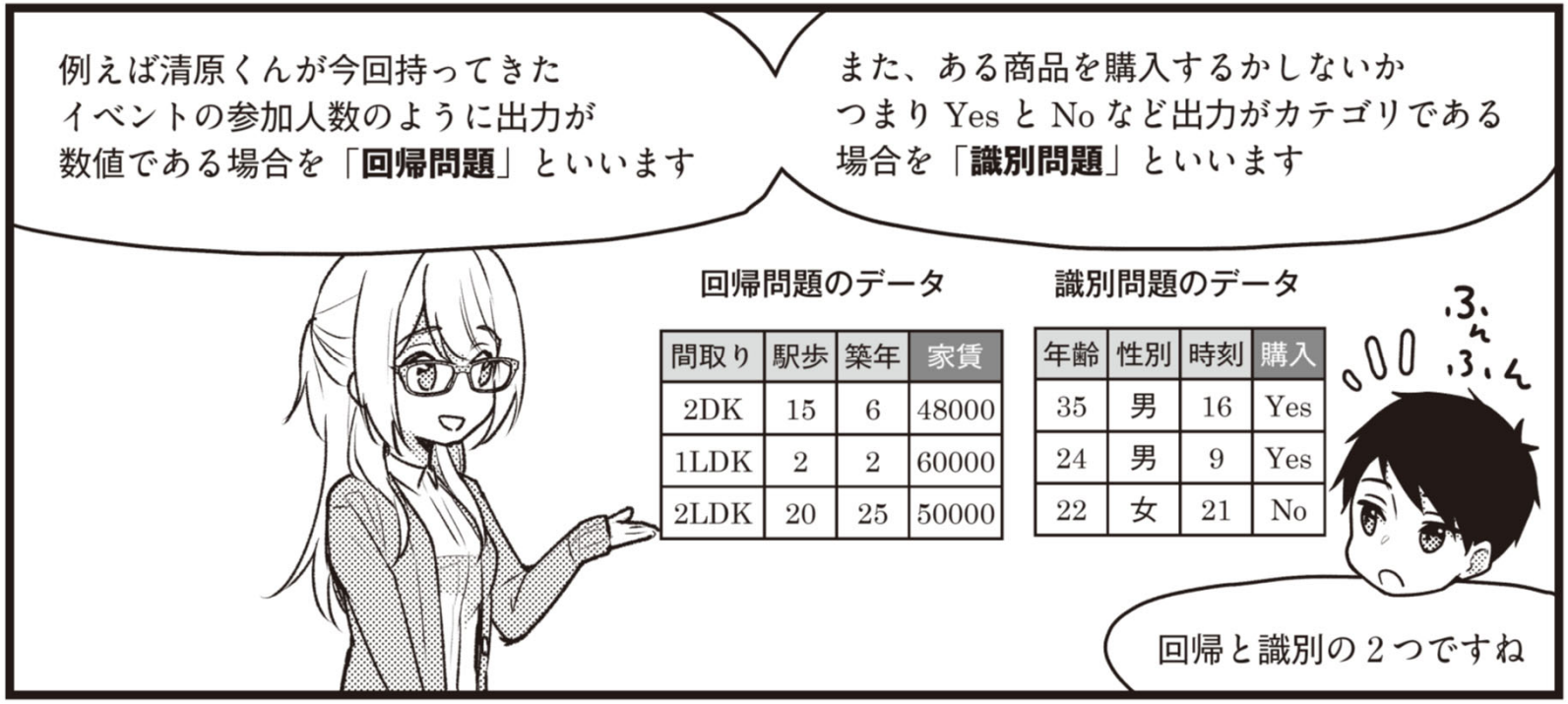
1章のストーリー

- さやかは清原にデータから数値を予測する回帰手法について教える

回帰（1章）



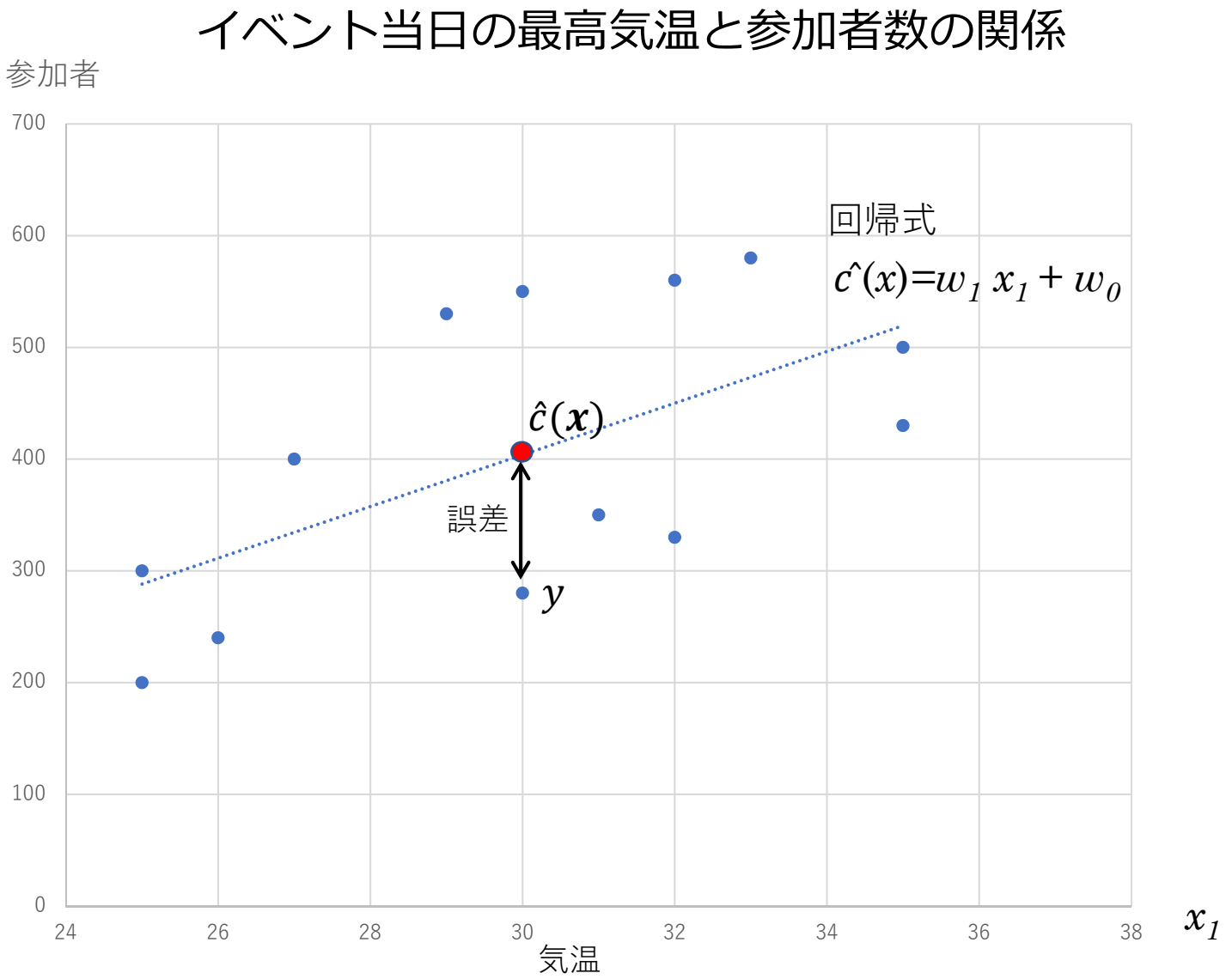
間取り	駅歩	築年	家賃
2DK	15	6	48000
1LDK	2	2	60000
2LDK	20	25	50000

年齢	性別	時刻	購入
35	男	16	Yes
24	男	9	Yes
22	女	21	No

p.11　3コマ目

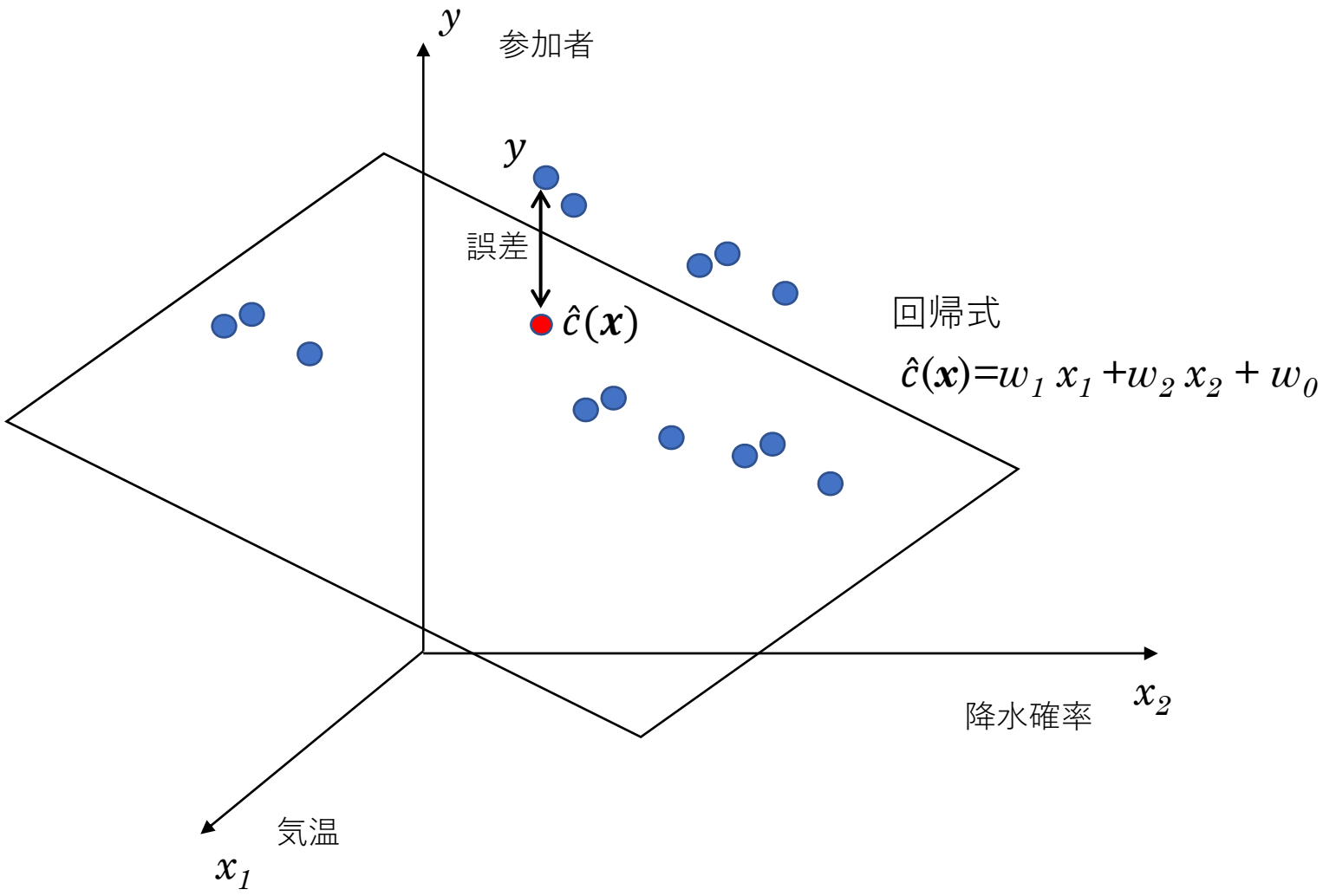
- 回帰とは
  - 教師あり学習のひとつ
  - 特徴の集合から**数値**を予測する

単純な回帰問題


イベント当日の最高気温と参加者数の関係
参加者
700
600
500
400
300
200
100
0
24
26
28
30
32
34
36
38
気温
x_1
回帰式
$\hat{c}(x)=w_1 x_1 + w_0$
$\hat{c}(\boldsymbol{x})$
誤差
y

やや複雑な回帰問題

イベント当日の最高気温・降水確率と参加者数の関係



線形回帰

- 問題の定義
  - 入力 $\boldsymbol{x}$ から出力 $\hat{c}(\boldsymbol{x})$ を求める回帰式を1次式に限定

直線・平面

$$\hat{c}(\boldsymbol{x}) = \sum_{i=0}^{d} w_i x_i = w_0 + w_1 x_1 + \cdots + w_d x_d$$

d : 特徴の次元数

x_0 : 1に固定

  - 学習データに対してなるべく誤差の少ない直線（あるいは平面）の係数 $\boldsymbol{w}$ を求める

最小二乗法による解法

- 推定の基準：誤差の二乗和 E を最小化

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \hat{c}(\boldsymbol{x}_i))^2$$

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

N : 全データ数

y_i : 正解

$\boldsymbol{X}$: 学習データを並べた行列

$\boldsymbol{y}$: 正解を並べたベクトル

$\boldsymbol{w}$: 係数を並べたベクトル

- Eが最小となるのは $\boldsymbol{w}$ で偏微分したものが0となるとき

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$$

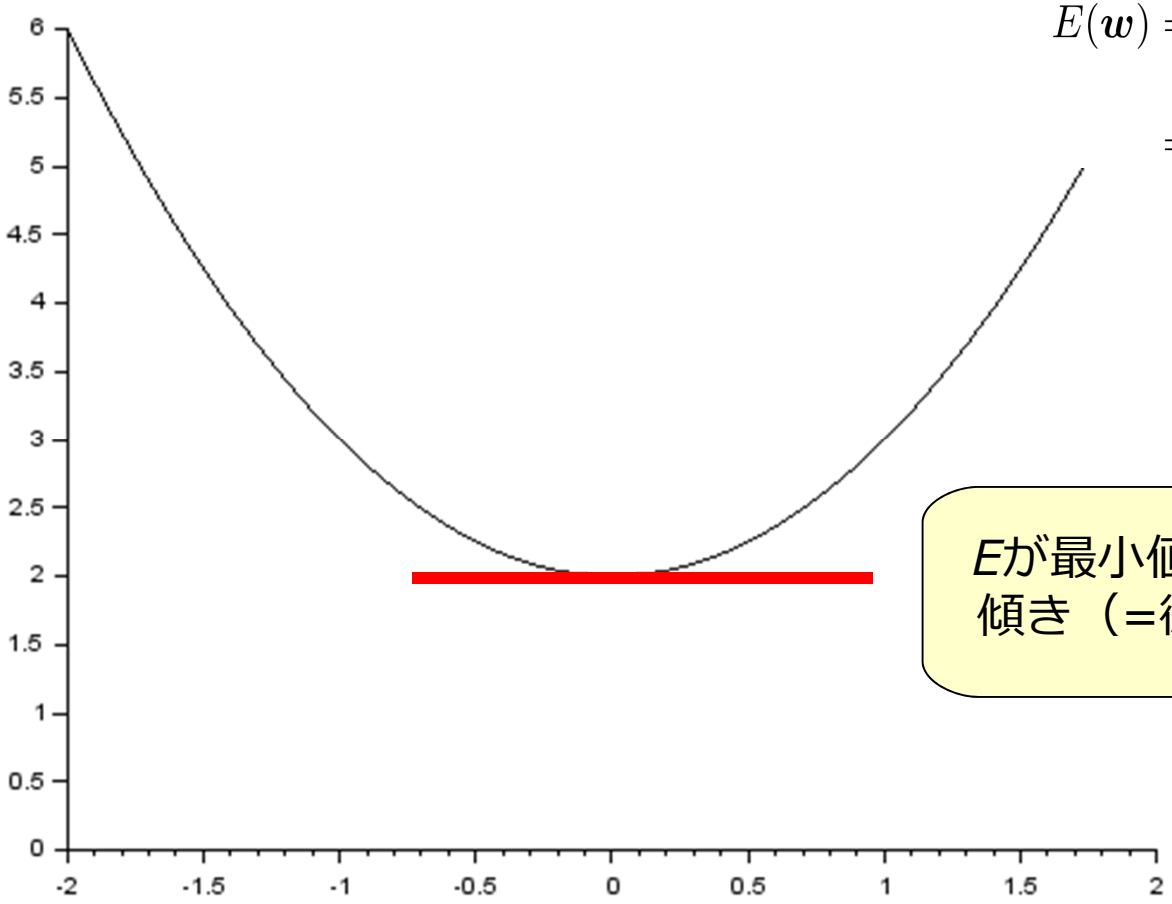
$$\Leftrightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$\boldsymbol{w}$ が行列の計算のみで求まる

最小二乗法による解法

- 誤差 E は二次関数

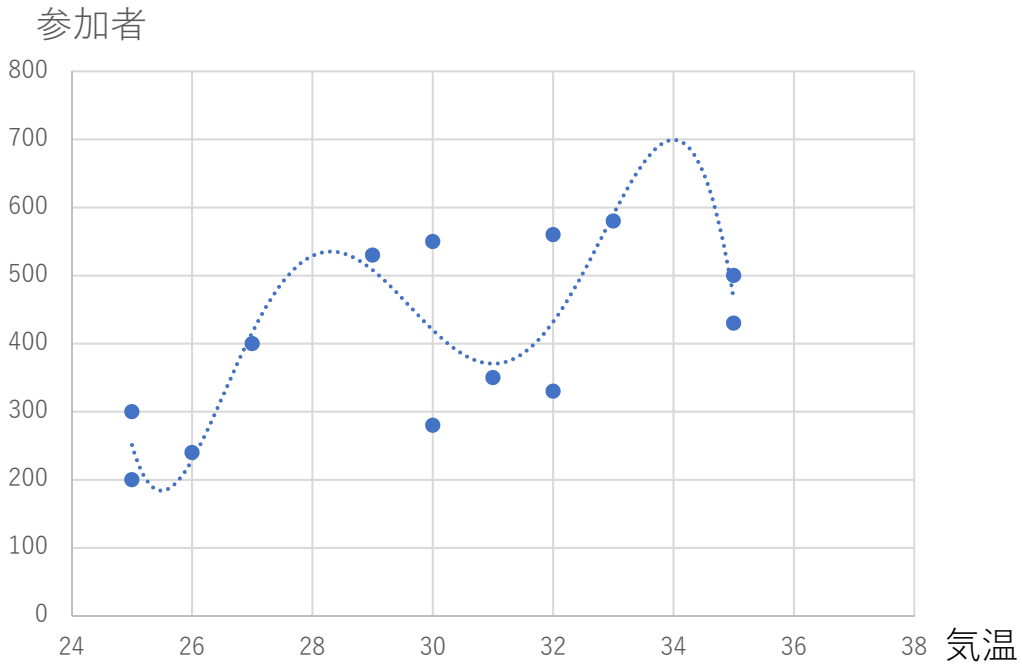
$$E(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - \hat{c}(\boldsymbol{x}_i))^2$$
$$= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$



*E*が最小値となるときは
傾き（=微分係数）が0

正則化

- 過学習
  - 最小二乗法は係数が線形であれば高次式でも適用可
  - 特徴の次数を上げたり、特徴の次元数を増やしたりすると、複雑な回帰式で解を近似することになる



学習データだけに
当てはまる不自然な
回帰式が求まって
しまう

正則化

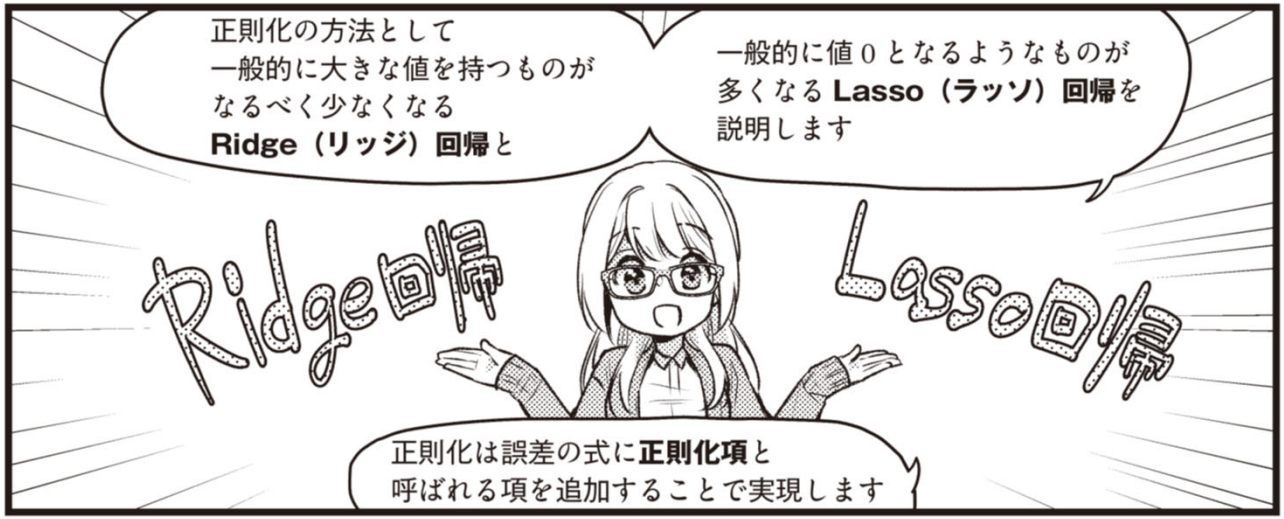
- 過学習への対処
  - 過学習した回帰式とは

    ⇒　入力が少し動いただけで出力が大きく動く

    ⇒　回帰式の係数 $\boldsymbol{w}$ が大きい
  - 正則化

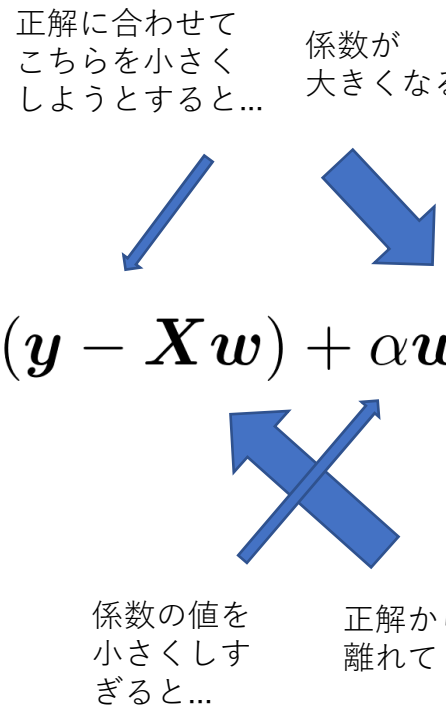
誤差が多少増えることと
引き換えに$\boldsymbol{w}$を小さくする

p.23　3コマ目



Ridge回帰

- 係数$\boldsymbol{w}$の2乗を正則化項として誤差の式に加える
  - 全体的に係数が小さくなり、極端な値の変動がなくなる



$$E(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \alpha \boldsymbol{w}^T\boldsymbol{w}$$

α：誤差と正則化項のバランス

Lasso回帰

- 係数 $\boldsymbol{w}$ の絶対値を正則化項として誤差の式に加える
  - 値が0となる係数が多くなり、出力に影響を与えている特徴を絞り込むことができる

$$E(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \alpha \sum_{j=1}^{d} |w_j|$$

回帰式の具体例

- Bostonデータ
  - 犯罪発生率、部屋数、立地など13の条件から不動産価格を推定

線形回帰

```
CRIM    :   -0.11
ZN      :    0.05
INDUS   :    0.02
CHAS    :    2.69
NOX     :  -17.80
RM      :    3.80
AGE     :    0.00
DIS     :   -1.48
RAD     :    0.31
TAX     :   -0.01
PTRATIO:   -0.95
B       :    0.01
LSTAT   :   -0.53
```

Ridge

```
CRIM    :   -0.10
ZN      :    0.05
INDUS   :   -0.04
CHAS    :    1.95
NOX     :   -2.37
RM      :    3.70
AGE     :   -0.01
DIS     :   -1.25
RAD     :    0.28
TAX     :   -0.01
PTRATIO:   -0.80
B       :    0.01
LSTAT   :   -0.56
```

Lasso

```
CRIM    :   -0.02
ZN      :    0.04
INDUS   :   -0.00
CHAS    :    0.00
NOX     :   -0.00
RM      :    0.00
AGE     :    0.04
DIS     :   -0.07
RAD     :    0.17
TAX     :   -0.01
PTRATIO:   -0.56
B       :    0.01
LSTAT   :   -0.82
```

回帰の実用化事例

- NEC
  - 日配品（主に冷蔵が必要なもの）の需要予測

    https://jpn.nec.com/ai/solution/value.html
  - 健診結果予測シミュレーション

    https://wisdom.nec.com/ja/technology/2018031501/index.html
- 心疾患リスクスコアの推定 [Ganz et al. 16]
  - 血液中の1130種類のタンパク質から心疾患に関連する9種のタンパク質を特定

    prognostic index
    $= 16.61 - 1.55 \times \mathrm{ANGPT2} + 1.22 \times \mathrm{GDF8/11} - 2.12 \times \mathrm{C7} + 2.64 \times \mathrm{SERPINF2} - 0.57 \times \mathrm{CCL18} - 1.02 \times \mathrm{ANGPTL4} - 1.43 \times \mathrm{SERPINA3} - 0.72 \times \mathrm{MMP12} - 0.59 \times \mathrm{TNN13}$