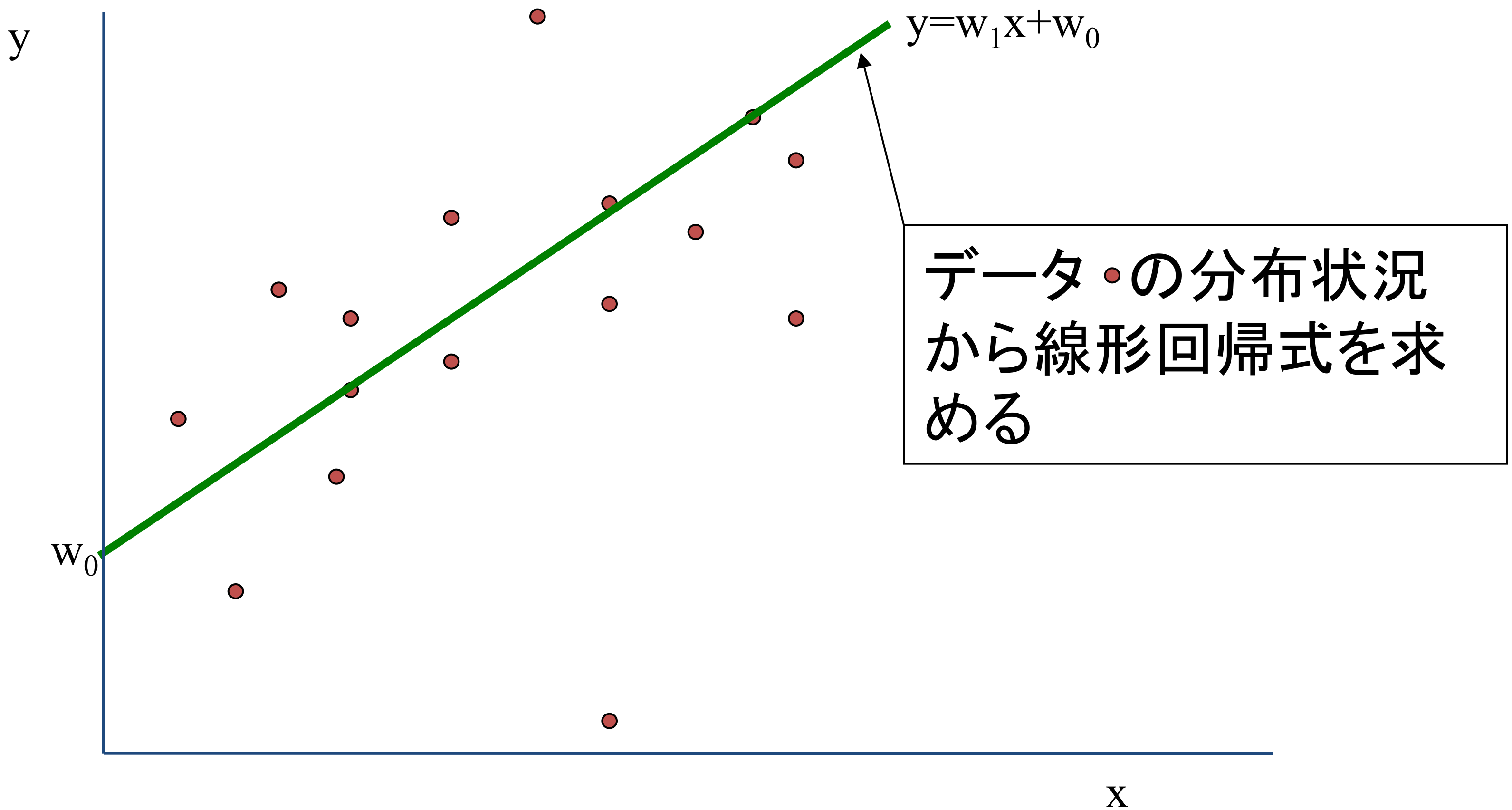
クラシックな機械学習の入門

3. 線形回帰および識別

線形回帰のモデル
正則化項の導入
L2正則化
L1正則化
正則化項のBayes的解釈
線形識別
生成モデルを利用した識別
2乗誤差最小化の線形識別の問題点

by 中川裕志（東京大学）

線形モデル



線形モデル

- 入力ベクトル : $\mathbf{x}$　から出力 : y　を得る関数がxの線形関数（wとxの内積）

$$y = \langle \mathbf{x}, \mathbf{w} \rangle = \sum_{i=0}^{K} w_i x_i \qquad \text{ただし、} \mathbf{x} = [1, x_1, \ldots, x_K]^T, \mathbf{w} = [w_0, w_1, \ldots, w_K]^{\mathrm{T}}$$

- 一般に観測データはノイズを含んでいる。つまり

$$y = \langle \mathbf{x}, \mathbf{w} \rangle + \varepsilon \qquad \varepsilon \text{はノイズで} N(0, \sigma^2) \text{と考える。}$$

- 得られたN個の観測データ の組（$\mathbf{y}, \mathbf{X}$）に対して最適な$\mathbf{w}$を推定する。
- そこで、$\mathbf{y}$と　$\mathbf{X}\mathbf{w}$　の2乗誤差を最小化するように$\mathbf{w}$を選ぶ。

2乗誤差の最小化

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \vdots \\ \mathbf{x_N}^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & & x_{1K} \\ & & \ddots & \\ 1 & x_{N1} & & x_{NK} \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$$

$\mathbf{w}$の推定値$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$

$\dfrac{\partial (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})}{\partial \mathbf{w}} = 0$ を解くと$\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathbf{T}}\mathbf{y}$

$$\boxed{\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

- 正規方程式　と呼ばれる基本式

補遺：正規方程式の導出

$$(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) = \left(\mathbf{y}^T - \mathbf{w}^T\mathbf{X}^T\right)(\mathbf{y}-\mathbf{X}\mathbf{w}) = \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$$

$$\frac{\partial(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})}{\partial\mathbf{w}} = -\frac{\partial\mathbf{w}^T\mathbf{X}^T\mathbf{y}}{\partial\mathbf{w}} - \frac{\partial\mathbf{y}^T\mathbf{X}\mathbf{w}}{\partial\mathbf{w}} + \frac{\partial\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}}{\partial\mathbf{w}} = 0 \quad (1)$$

$\dfrac{\partial\mathbf{x}^T\mathbf{a}}{\partial\mathbf{x}} = \mathbf{a}$ より $\dfrac{\partial\mathbf{w}^T\mathbf{X}^T\mathbf{y}}{\partial\mathbf{w}} = \mathbf{X}^T\mathbf{y}$ $\quad$ $\dfrac{\partial\mathbf{a}^T\mathbf{x}}{\partial\mathbf{x}} = \mathbf{a}$ より $\dfrac{\partial\mathbf{y}^T\mathbf{X}\mathbf{w}}{\partial\mathbf{w}} = \left(\mathbf{y}^T\mathbf{X}\right)^T = \mathbf{X}^T\mathbf{y}$

$$\frac{\partial\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}}{\partial\mathbf{w}} = \frac{\partial\mathbf{w}^T\left(\mathbf{X}^T\mathbf{X}\mathbf{w}\right)}{\partial\mathbf{w}} + \frac{\partial\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\right)\mathbf{w}}{\partial\mathbf{w}} = \mathbf{X}^T\mathbf{X}\mathbf{w} + \left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\right)^T = 2\mathbf{X}^T\mathbf{X}\mathbf{w}$$

$$\Rightarrow \quad (1) = -2\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \quad \Rightarrow \quad \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{w}$$

$$\Rightarrow \quad \mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

---

cf 行列で微分する場合のchain rule $\dfrac{\partial f(g(\mathbf{x}))}{\partial\mathbf{x}} = \dfrac{\partial g(\mathbf{x})}{\partial\mathbf{x}}\dfrac{\partial f(g(\mathbf{x}))}{\partial g(\mathbf{x})}$ を使えば

$$\frac{\partial(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})}{\partial\mathbf{w}} = \frac{\partial(\mathbf{y}..)}{\partial\mathbf{w}}\frac{\partial(\mathbf{y}..)^T(\mathbf{y}..)}{\partial(\mathbf{y}..)} + \frac{\partial(\mathbf{y}..)^T}{\partial\mathbf{w}}\frac{\partial(\mathbf{y}..)^T(\mathbf{y}..)}{\partial(\mathbf{y}..)^T}$$

$$= -\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w}) - \mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w}) = -2\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})$$

正規方程式を解く簡単な例

$$\mathbf{X}=\begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_N\end{bmatrix}\qquad \mathbf{y}=\begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}\qquad \mathbf{w}=\begin{bmatrix}w_0\\ w_1\end{bmatrix}$$

正規方程式 $\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{w}=\mathbf{X}^{T}\mathbf{y}$は

$$\begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_N\end{bmatrix}\begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_N\end{bmatrix}\begin{bmatrix}w_0\\ w_1\end{bmatrix}=\begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_N\end{bmatrix}\begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}$$

$$\Rightarrow\quad \begin{bmatrix}N & \sum_{i=1}^{N}x_i\\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i^2\end{bmatrix}\begin{bmatrix}w_0\\ w_1\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{N}y_i\\ \sum_{i=1}^{N}x_iy_i\end{bmatrix}\quad\Leftrightarrow\quad \mathbf{X}^{T}\mathbf{X}\quad\mathbf{w}\quad=\quad\mathbf{X}^{T}\mathbf{y}$$

$$\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}=\frac{1}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}\begin{bmatrix}\sum_{i=1}^{N}x_i^2 & -\sum_{i=1}^{N}x_i\\ -\sum_{i=1}^{N}x_i & N\end{bmatrix}\quad\Rightarrow\quad \boxed{w_1=\frac{N\sum_{i=1}^{N}x_iy_i-\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}}$$

$$w_0=\frac{\sum_{i=1}^{N}x_i^{2}\sum_{i=1}^{N}y_i-\sum_{i=1}^{N}x_i\sum_{i=1}^{N}x_iy_i}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}=\frac{\sum_{i=1}^{N}y_i}{N}-\frac{\sum_{i=1}^{N}x_i}{N}\boxed{\frac{N\sum_{i=1}^{N}x_iy_i-\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}}=w_0=\boxed{\frac{1}{N}\sum_{i=1}^{N}y_i-\frac{1}{N}w_1\sum_{i=1}^{N}x_i}$$

用語：誤差、損失、目的関数

- 線形モデルで最小化したかったのは2乗誤差
- 真のモデルにおける値(２乗誤差におけるy)と

  予測値(２乗誤差におけるXw)の差異を表す関数を損失関数（単に損失）あるいは*Loss*と呼び、*L*で表すことが多い。
- 上記のような最適化問題において最小化（一般的には最適化）したい関数を目的関数と呼ぶ。
- 線形モデルの２乗誤差最小化では

  ２乗誤差＝損失＝目的関数

線形モデルの一般化

$$y = \langle \varphi(\mathbf{x}), \mathbf{w} \rangle \qquad \varphi(\mathbf{x}) = [1, \phi_1(\mathbf{x}), \cdots, \phi_K(\mathbf{x})]^T$$

基底関数　重み

N個の観測データ(y,x)に対して $\mathrm{y} = \begin{pmatrix} \mathrm{y}_1 \\ \vdots \\ \mathrm{y}_\mathrm{N} \end{pmatrix} \quad \boldsymbol{\varphi}(\mathbf{x}) = \begin{pmatrix} \varphi_1(\mathbf{x})^T \\ \vdots \\ \varphi_\mathrm{N}(\mathbf{x})^T \end{pmatrix}$

(y、φ(x))が得られたとすると、2乗誤差を最小化するwは前を同じく以下の通りだが、少し別の見方で解く。

$$\boxed{\hat{w} = (\boldsymbol{\varphi}(\mathbf{x})^T \boldsymbol{\varphi}(\mathbf{x}))^{-1} \boldsymbol{\varphi}(\mathbf{x})^T \mathbf{y}}$$

基底関数の例

$$\phi_j(x) = x^j \quad \text{: polynominal}$$

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\} \quad \text{: Gaussian}$$

$$\phi_j(x) = \frac{1}{1+\exp(-(x-\mu_j)/s)} \quad \text{: sigmoidal}$$

$$\phi_j(x) = \exp\left(2\pi i \frac{xj}{m}\right) \quad \text{(m : even) : Fast Fourier}$$

正規方程式を求める別の方法

$$y = \langle \varphi(\mathbf{x}), \mathbf{w} \rangle + \varepsilon \qquad \varepsilon = N(0, \beta^{-1}) \qquad \beta = \sigma^{-2}\text{を精度と呼ぶ.}$$

$$p(y \mid \mathbf{x}, \mathbf{w}, \beta) = N(y \mid \langle \varphi(\mathbf{x}), \mathbf{w} \rangle, \beta^{-1})$$

- {x(ベクトル),y}が観測データ(training data)
- w,βを決定する、即ち (p(y|x,w,β)を最大化）
- N組のi.i.d.観測データすなわち教師データがあるとする。

$$\mathbf{y} = (y_1, \ldots, y_N)^T \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \vdots \\ \mathbf{x_N}^T \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$$

すると次のページのようにp(y|x,w,β）が書ける。

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} N(y_i \mid \langle \varphi(\mathbf{x_i}), \mathbf{w} \rangle, \beta^{-1})$$

両辺のlogをとる

$$\log p(\mathbf{y} \mid \mathbf{w}, w_0, \mathbf{X}, \beta) = \frac{N}{2} \log \beta - \frac{N}{2} \log 2\pi - \beta L(\mathbf{w})$$

$$L(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \langle \varphi(\mathbf{x_i}), \mathbf{w} \rangle \right)^2$$

$\log p(y|\boldsymbol{w},X,\beta)$を$\boldsymbol{w},\beta$について最大化したい。まず、$\boldsymbol{w}$について最大化する。

$$\frac{\partial \log p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}, \beta)}{\partial \mathbf{w}} = \beta \sum_{i=1}^{N} \varphi(\mathbf{x_i}) \left(y_i - \langle \varphi(\mathbf{x_i}), \mathbf{w} \rangle \right) = 0$$

$$\Rightarrow$$

$$\sum_{i=1}^{N} \varphi(\mathbf{x_i}) y_i - \sum_{i=1}^{N} \varphi(\mathbf{x_i}) \varphi(\mathbf{x_i})^T \mathbf{w} = 0$$

$$\Rightarrow$$

$$\boldsymbol{\varphi}(\mathbf{X})^T \cdot \mathbf{y} = (\boldsymbol{\varphi}(\mathbf{X})^T \boldsymbol{\varphi}(\mathbf{X})) \mathbf{w}$$

$$\Rightarrow$$

$$\boxed{\hat{\mathbf{w}} = (\boldsymbol{\varphi}(\mathbf{X})^T \boldsymbol{\varphi}(\mathbf{X}))^{-1} \boldsymbol{\varphi}(\mathbf{X})^T \mathbf{y}}$$

$$\boldsymbol{\varphi}(\mathbf{x}) = \begin{pmatrix} \varphi(\mathbf{x}_1)^T \\ \vdots \\ \varphi(\mathbf{x}_N)^T \end{pmatrix}$$

バイアスw_0の部分だけに注目してみると

- 対数近似関数から最適なw_0を　によって求めると

$$\frac{\partial L(\mathbf{w})}{\partial w_0} = \frac{\partial \sum_{i=1}^{N} \left(y_i - (1, \phi_1(\mathbf{x_i}), .. \phi_K(\mathbf{x_i})) \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{pmatrix} \right)^2}{\partial w_0} = \frac{\partial \sum_{i=1}^{N} \left(y_i - (\phi_1(\mathbf{x_i}), .. \phi_K(\mathbf{x_i})) \begin{pmatrix} w_1 \\ \vdots \\ w_K \end{pmatrix} - w_0 \right)^2}{\partial w_0}$$

$$= -2 \sum_{i=1}^{N} \left(y_i - (1, \phi_1(\mathbf{x_i}), .. \phi_K(\mathbf{x_i})) \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{pmatrix} \right) = -2 \sum_{i=1}^{N} \left(y_i - (\phi_1(\mathbf{x_i}), .. \phi_K(\mathbf{x_i})) \begin{pmatrix} w_1 \\ \vdots \\ w_K \end{pmatrix} - w_0 \right) = 0$$

$$\Rightarrow$$

$$\boxed{w_0 = \frac{1}{N} \sum_{i=1}^{N} y_i \quad - \quad \frac{1}{N} \sum_{j=1}^{K} w_j \left(\sum_{i=1}^{N} \phi_j(\mathbf{x_i}) \right)}$$

yの平均　　　基底関数の学習データの平均のw重み付き和

精度βを求める。

$\log p(y|\boldsymbol{w},X,\beta)$を$\beta$に対して最大化

ただし、wは最適化されたものを用いる

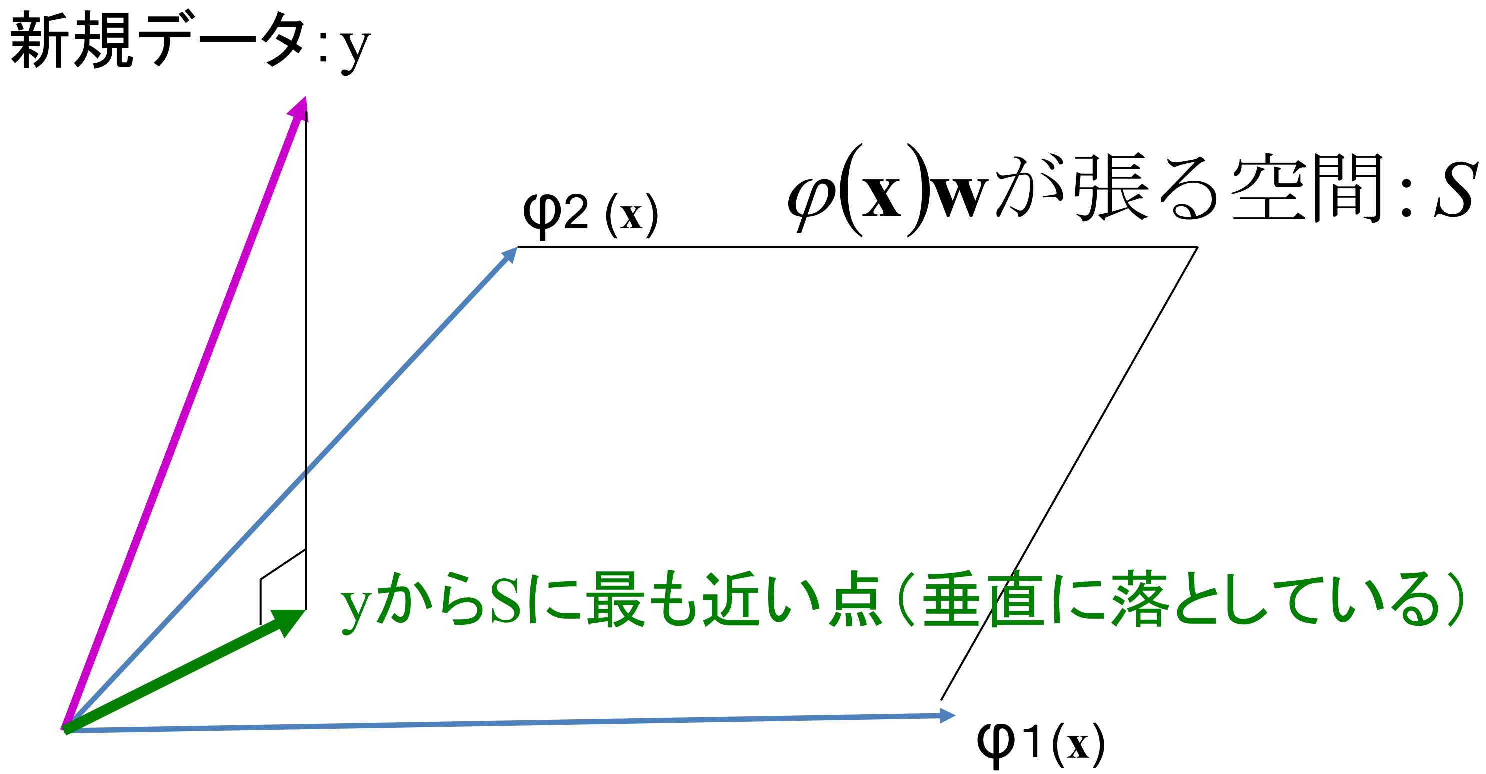
$$\frac{\partial \log p(\mathbf{y} \mid \hat{\mathbf{w}}, \mathbf{X}, \beta)}{\partial \beta} = \frac{N}{2\beta} - L(\hat{\mathbf{w}})$$

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{\varphi}(\mathbf{x_i})\hat{\mathbf{w}}\right)^2$$

$$\boxed{\hat{\beta}^{-1} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \boldsymbol{\varphi}(\mathbf{x_i})\hat{\mathbf{w}}\right)^2}$$

yの予測値と観測された値の差の2乗の平均

幾何学的イメージ


新規データ:y
$\varphi(\mathbf{x})\mathbf{w}$が張る空間: S
φ2 (x)
φ1(x)
yからSに最も近い点（垂直に落としている）

計算の効率化

- 大きなdata setsに対して $\hat{\mathbf{w}} = (\boldsymbol{\varphi}^T\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^T\mathbf{y}$
 の右辺第1項の逆行列計算量が問題
- 特にデータの次元Nに対してO(N^3)なので高次元だと大変
- 定石は、コレスキー分解O(N^2)して上/下半3角行列で表現される連立方程式を2回解く
- ***L***(**w**)を最小化するような**w**の数値計算

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla L(\mathbf{w})$$

$$= \mathbf{w}^{(\tau)} - \eta(y_n - \phi(\mathbf{x}_n)\mathbf{w}^{(\tau)})\phi(\mathbf{x}_n)^T$$

目的関数(すなわち損失L(w))の減る方向へ進む(ー gradientをwに加える)方法をgradient descent は呼ばれ、最適化における基本的数値計算法である。

正則化項の導入

- モデルを複雑にするほど学習データにはよく合致するが、学習データ以外のデータには弱いという過学習を起こす。
- 過学習を抑えるために、損失関数に正則化項を導入。
- 正則化項にはモデルをできるだけ簡単化する方向に作用する。
  - データが高次元の場合には次元削減効果あり。

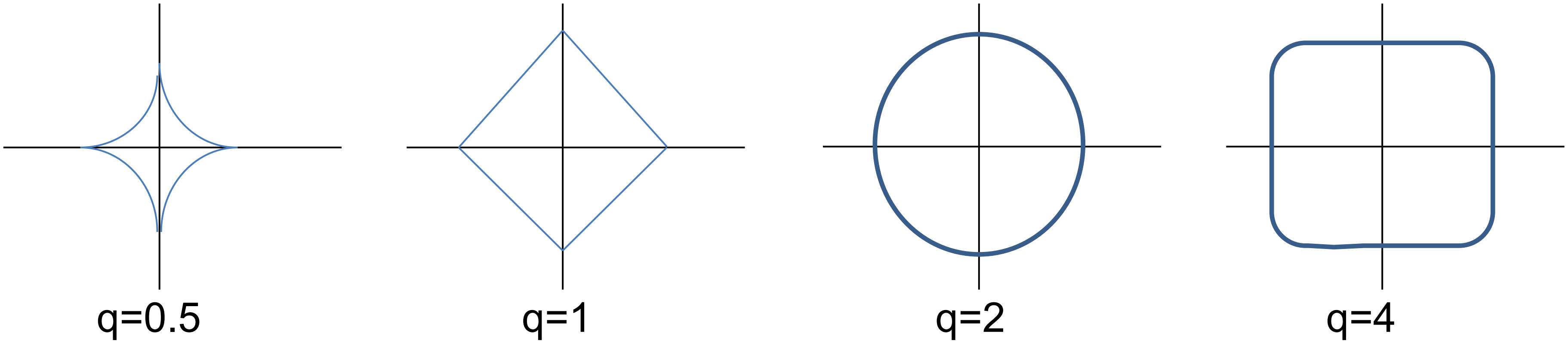
一般的な正則化項

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{K}|w_j|^q}_{\text{正則化項}}$$

- q=2のときがL2正則化
- q=1のときはLASSO:　１ノルムによる正則化なので　L1正則化と呼ぶ
- Least Absolute Shrinkage and Selection Operator
  - λが十分大きいと、w_jのいくつかは0になりやすい　→　スパースなモデル
- q=0のときはL0正則化。解きにくい問題（上記2つと違い凸ではない）

- 制約 $\sum_{j=1}^{K} \left| w_j \right|^q \leq \eta$

のもとで、$\boldsymbol{L}$(w)を最小化する、と考える。



L2正則化

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w}\rangle\right)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

最小化すると

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} L(\mathbf{w}) = (\lambda\mathbf{I} + \boldsymbol{\varphi}(\mathbf{X})^T\boldsymbol{\varphi}(\mathbf{X}))^{-1}\boldsymbol{\varphi}(\mathbf{X})^T\mathbf{y}$$

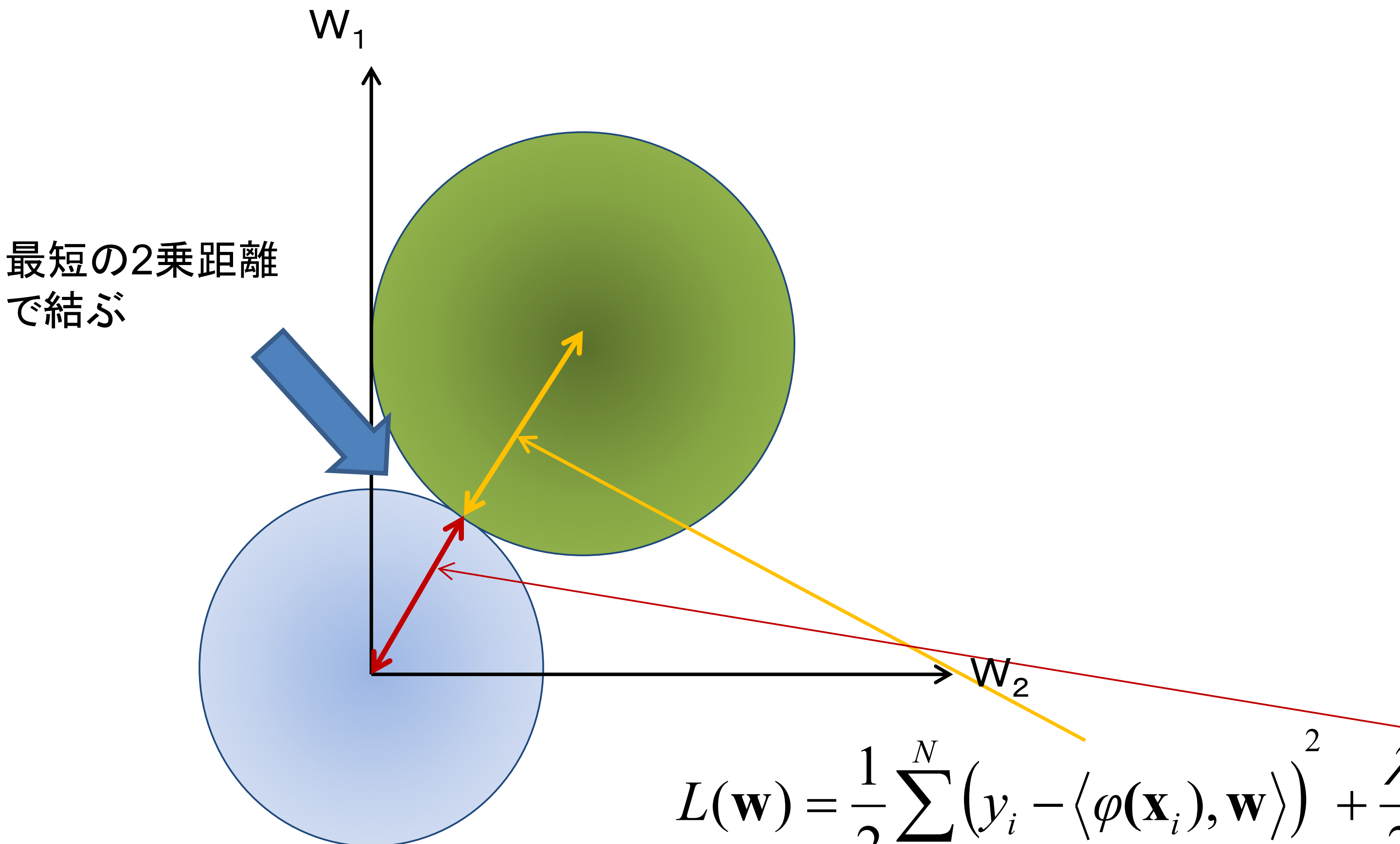
正則化項

(wの影響を小さくする効果)

Wの２ノルムによる正則化であるので、L2正則化と呼ぶ

- 最適なwはL(w)を微分して0とすれば上記のように解析的に閉じた式で求まる。
- これはφ(X)とλの案配よって決まり、どの成分も強制的にゼロにしようという力は働かない

L2正則化のイメージ



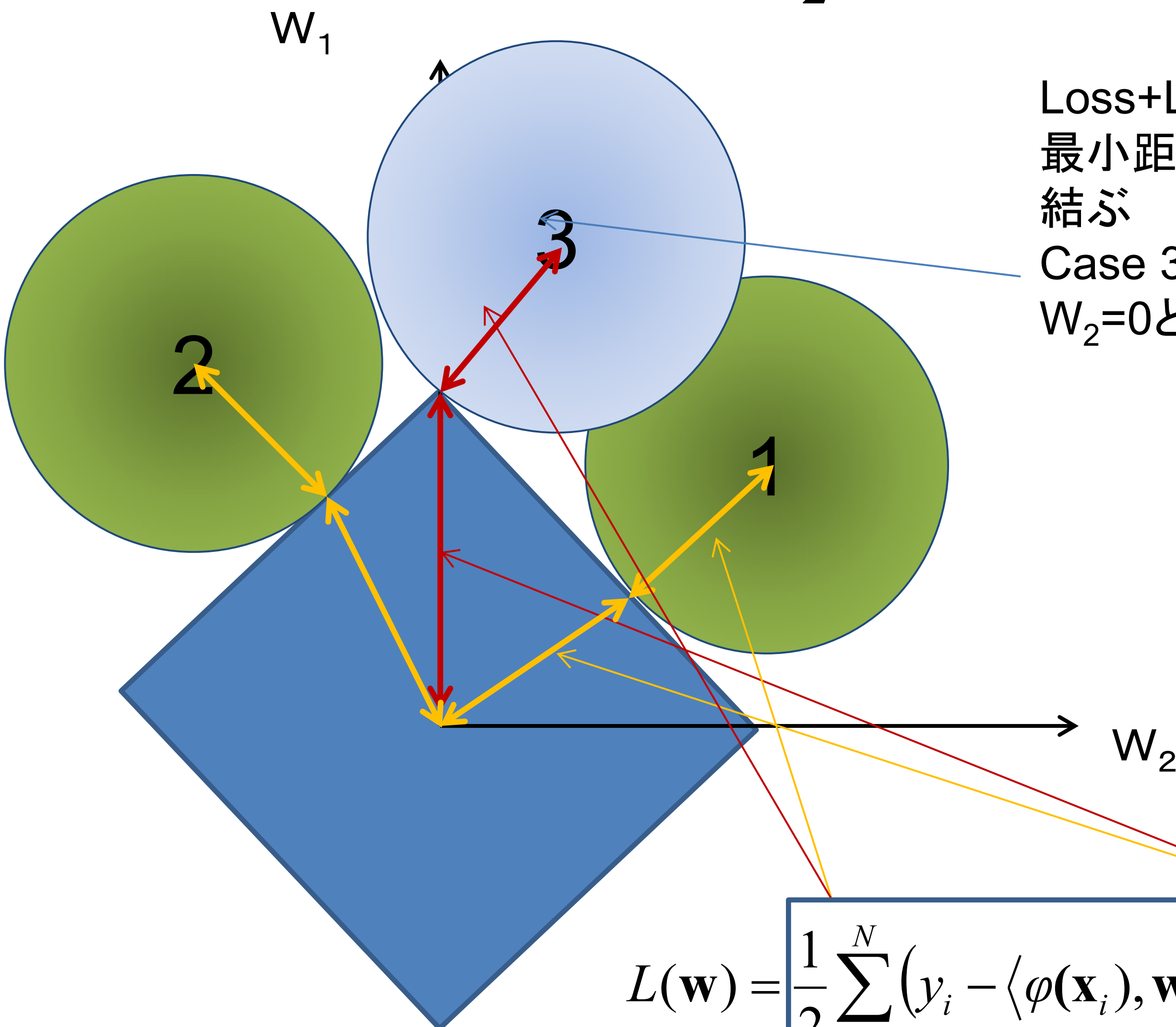
$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \left\langle \varphi(\mathbf{x}_i), \mathbf{w} \right\rangle\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{K} w_j^{\ 2}$$

L1正則化

- L2正則化ではwの最適値 $\hat{\mathbf{w}}$ を損失Lの微分で閉じた式で求められたが、L1正則化では|w|がw＝0で微分できないので、ややこしくなる。
- L1正則化を行う逐次的な方法と

  L1正則化がwの要素の多くをゼロ化する傾向を以下で説明する

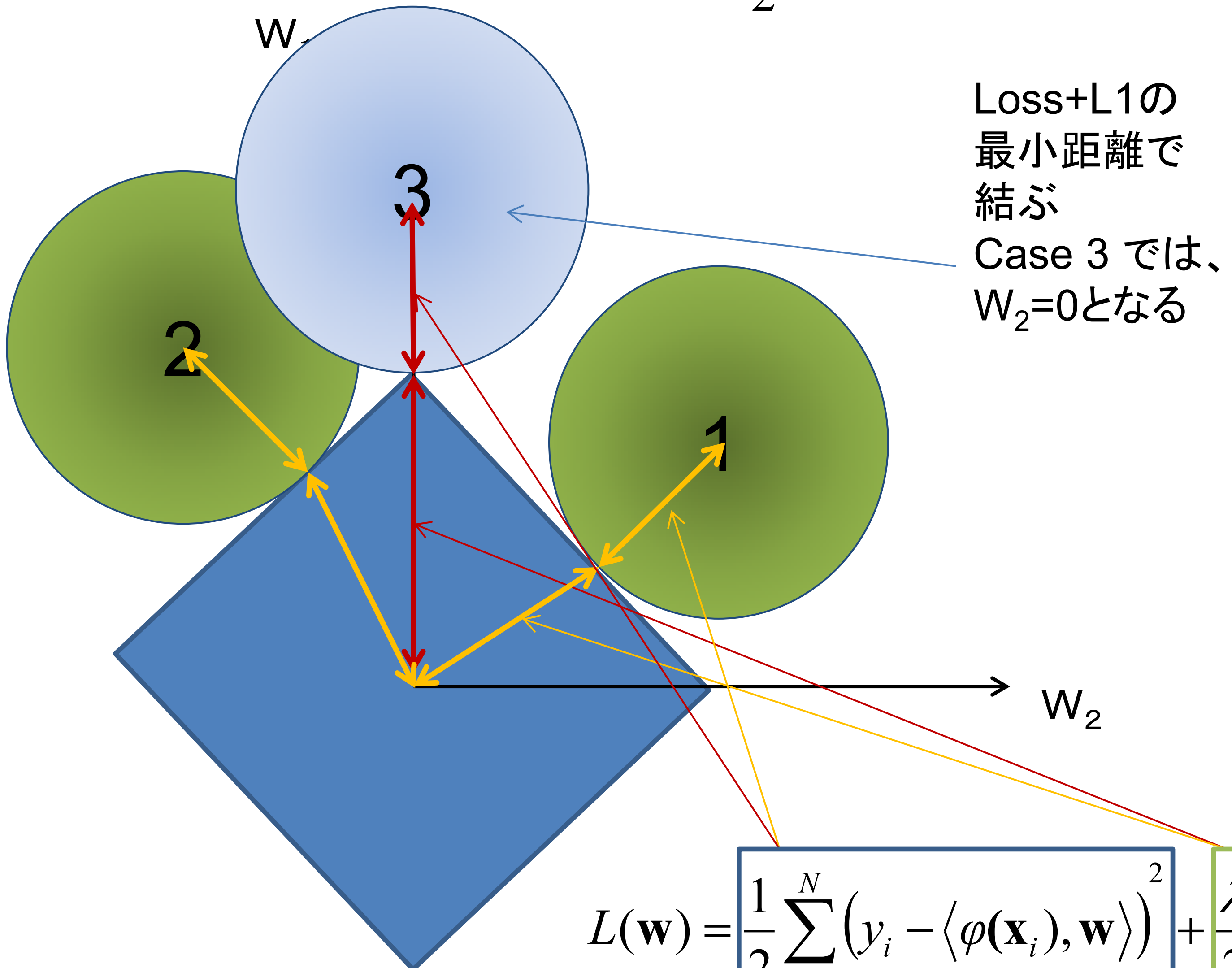
L１正則化イメージ:　（1）
w_2軸でのLossの微分=0として$\tilde{w}_2$を求める



Loss+L1の
最小距離で
結ぶ
Case 3 では、
W_2=0となる

$$L(\mathbf{w}) = \underbrace{\boxed{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2}}_{\text{Loss}} + \underbrace{\boxed{\frac{\lambda}{2}\sum_{j=1}^{K}\left|w_j\right|}}_{\text{L1}}$$

L1正則化イメージ:（2）
w_2軸でのLossの微分=0として$\widetilde{w}_2$を求める

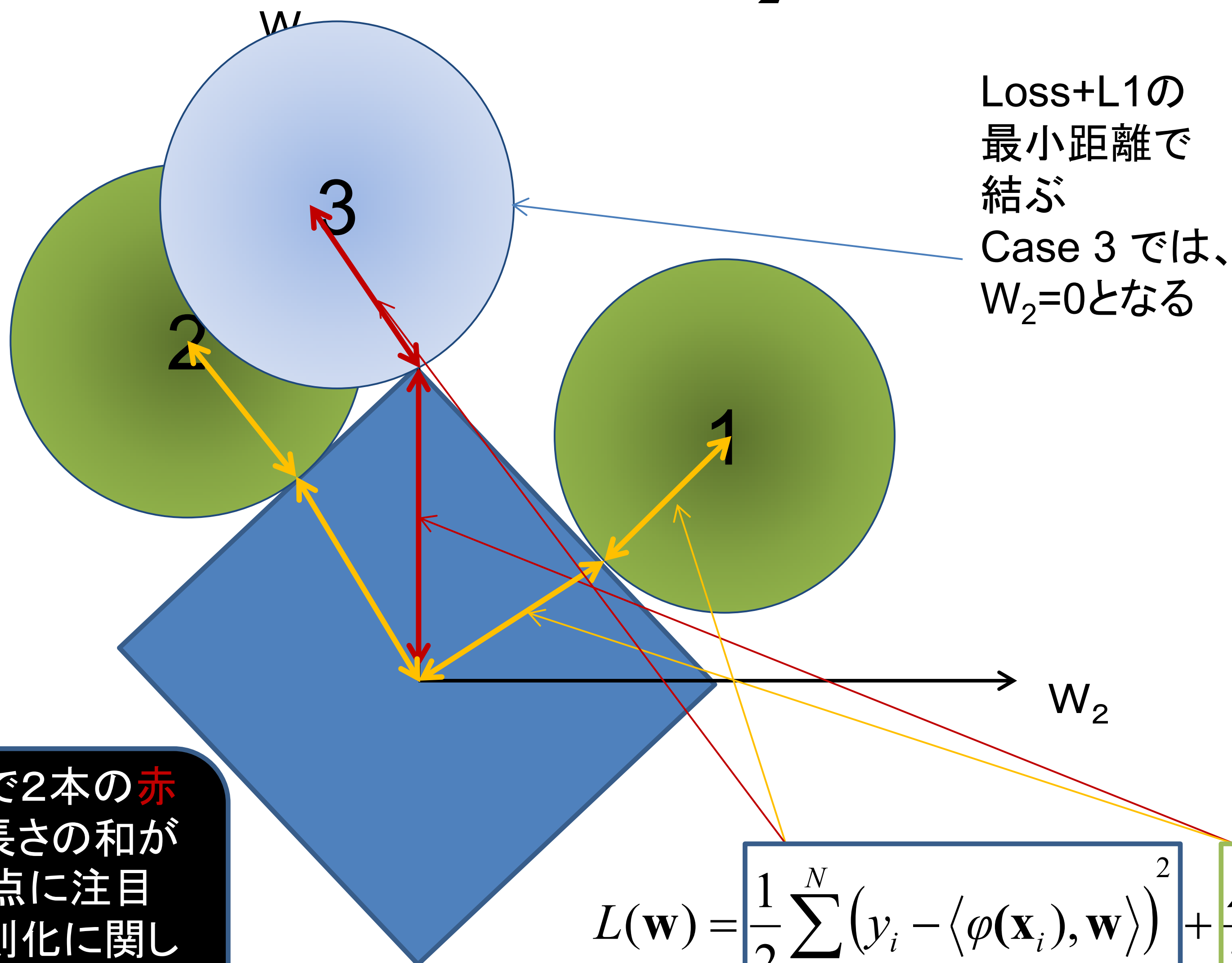


$$L(\mathbf{w}) = \boxed{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w}\rangle\right)^2} + \boxed{\frac{\lambda}{2}\sum_{j=1}^{K}|w_j|}$$

Loss L1

L１正則化イメージ：（３）

w_2軸でのLossの微分=0として$\tilde{w}_2$を求める



$$L(\mathbf{w}) = \underbrace{\boxed{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w}\rangle\right)^2}}_{\text{Loss}} + \underbrace{\boxed{\frac{\lambda}{2}\sum_{j=1}^{K}|w_j|}}_{\text{L1}}$$

（１）（２）（３）で２本の赤い矢印線の長さの和が変わらない点に注目

以下でL1正則化に関してもう少し細かく議論する。

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w}\rangle\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{K}|w_j| \qquad (L1-10)$$

- ある次元dに着目して$L(\mathbf{w})$を最小化するようなw_dを求める。
- これを各次元について繰り返し、$L(\mathbf{w})$の最小化を図る。w_dについて$L(\mathbf{w})$を書き直すと

$$\begin{aligned} L(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \varphi_d(\mathbf{x}_i)w_d - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right)^2 + \frac{\lambda}{2}\left(|w_d| + \sum_{j\neq d}|w_j|\right) \\ &= Loss(\mathbf{w}) \qquad\qquad\qquad\qquad\qquad\qquad\qquad + \quad L1(\mathbf{w}) \qquad (L1-20) \end{aligned}$$

- $\dfrac{\partial L(\mathbf{w})}{\partial w_d} = 0$ とおきw_dの最適値を求めたいが絶対値を含む第２項$L1(w)$が微分できないので、ひとまず$Loss(\mathbf{w})$を微分して0とおくと

$$\frac{\partial Loss(\mathbf{w})}{\partial w_d} = \frac{\partial}{\partial w_d}\left(\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \varphi_d(\mathbf{x}_i)w_d - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right)^2\right)$$

$$= \sum_{i=1}^{N} -\varphi_d(\mathbf{x}_i)\left(y_i - \varphi_d(\mathbf{x}_i)w_d - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right) = 0 \text{の解を}\tilde{w}_d\text{とする}$$

$$\Rightarrow \qquad \tilde{w}_d = \frac{\sum_{i=1}^{N}\varphi_d(\mathbf{x}_i)\left(y_i - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right)}{\sum_{i=1}^{N}\varphi_d(\mathbf{x}_i)^2}$$

これを用いて$L(\mathbf{w})$を書き換える。ただし、$\mathbf{w}_d$に関係しないところは当面定数と見なせるので、無視した。

$$2L(\tilde{w}_d) = \sum_{i=1}^{N}\left(\varphi_d(\mathbf{x}_i)^2 w_d^2 - 2\varphi_d(\mathbf{x}_i)w_d\left(y_i - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right)\right) + \lambda|w_d| + Const$$

$$= w_d^2\sum_{i=1}^{N}\varphi_d(\mathbf{x}_i)^2 - 2w_d\sum_{i=1}^{N}\left(\varphi_d(\mathbf{x}_i)^2\right)\sum_{i=1}^{N}\left(\varphi_d(\mathbf{x}_i)w_d\left(y_i - \sum_{j\neq d}\varphi_j(\mathbf{x}_i)w_j\right)\right)\Bigg/\sum_{i=1}^{N}\left(\varphi_d(\mathbf{x}_i)^2\right) + \lambda|w_d| + Const$$

$$= w_d^2\sum_{i=1}^{N}\varphi_d(\mathbf{x}_i)^2 - 2w_d\tilde{w}_d\sum_{i=1}^{N}\varphi_d(\mathbf{x}_i)^2 + \lambda|w_d| + Const$$

ここで$\gamma = \dfrac{\lambda}{2\sum_{i=1}^{N}\varphi_d(\boldsymbol{x}_i)^2}$とおくと、$L(w_d) = \dfrac{1}{2}\left(w_d^2 - w_d\tilde{w}_d\right) + \gamma|w_d| + const$

$$\frac{\partial L(w_d)}{\partial w_d} = \begin{cases} w_d - \tilde{w}_d + \gamma & w_d > 0 \\ w_d - \tilde{w}_d - \gamma & w_d < 0 \\ \text{undefined} & w_d = 0 \end{cases} \Rightarrow \frac{\partial L(w_d)}{\partial w_d} = 0\text{なる}w_d\text{を探す}$$

case 1　$\tilde{w}_d - \gamma > 0$　なら　$w_d > 0$　なので　$w_d = \tilde{w}_d - \gamma$

case 2　$\tilde{w}_d + \gamma < 0$　なら　$w_d < 0$　なので　$w_d = \tilde{w}_d + \gamma$

case 3　$-\gamma < \tilde{w}_d < \gamma$　なら　$w_d = 0$　なぜなら

$w_d > 0$だと$\tilde{w}_d - \gamma > 0$すなわち$\tilde{w}_d > \gamma$　矛盾

$w_d < 0$だと$\tilde{w}_d + \gamma < 0$すなわち$\tilde{w}_d < -\gamma$　矛盾

case 3により$-\gamma < \tilde{w}_d < \gamma$すなわち$Loss(\boldsymbol{w})$の$w_d$の解$\tilde{w}_d$がゼロに近づくと$w_d = 0$になりゼロ化(スパース化）される力が働く

w全体の正則化

[step 1] w の各要素を適当な値に初期化

[step 2] w の各要素の値w $_k$(k=1,..,K)が収束するまで以下step 3,4,5 を繰り返す

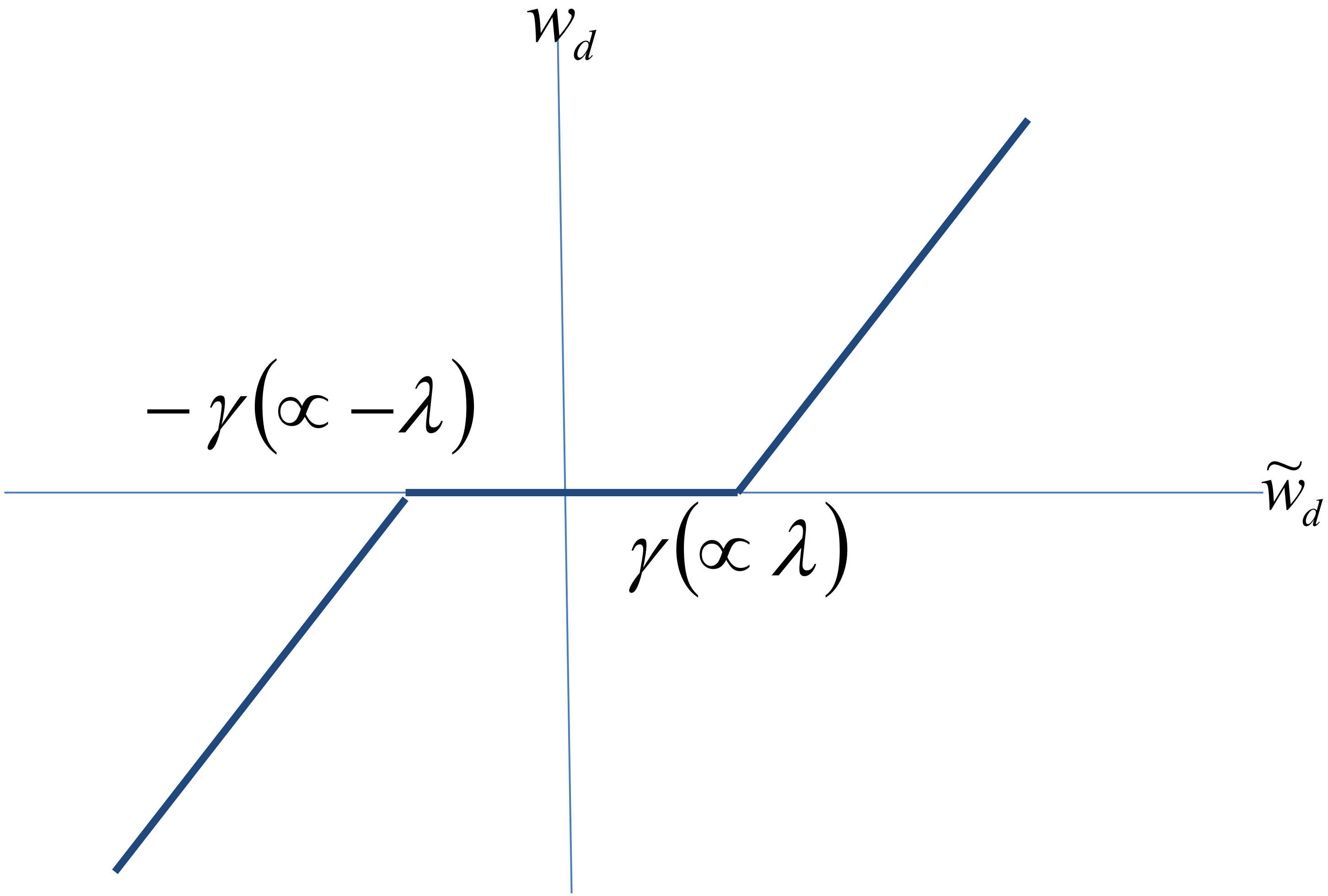
[step 3] k=1,.., Kでstep 4, step 5を繰り返す

[step 4] w_j ($j \neq k$)を用いて case1,2,3にしたがってw_jを計算してゼロ化

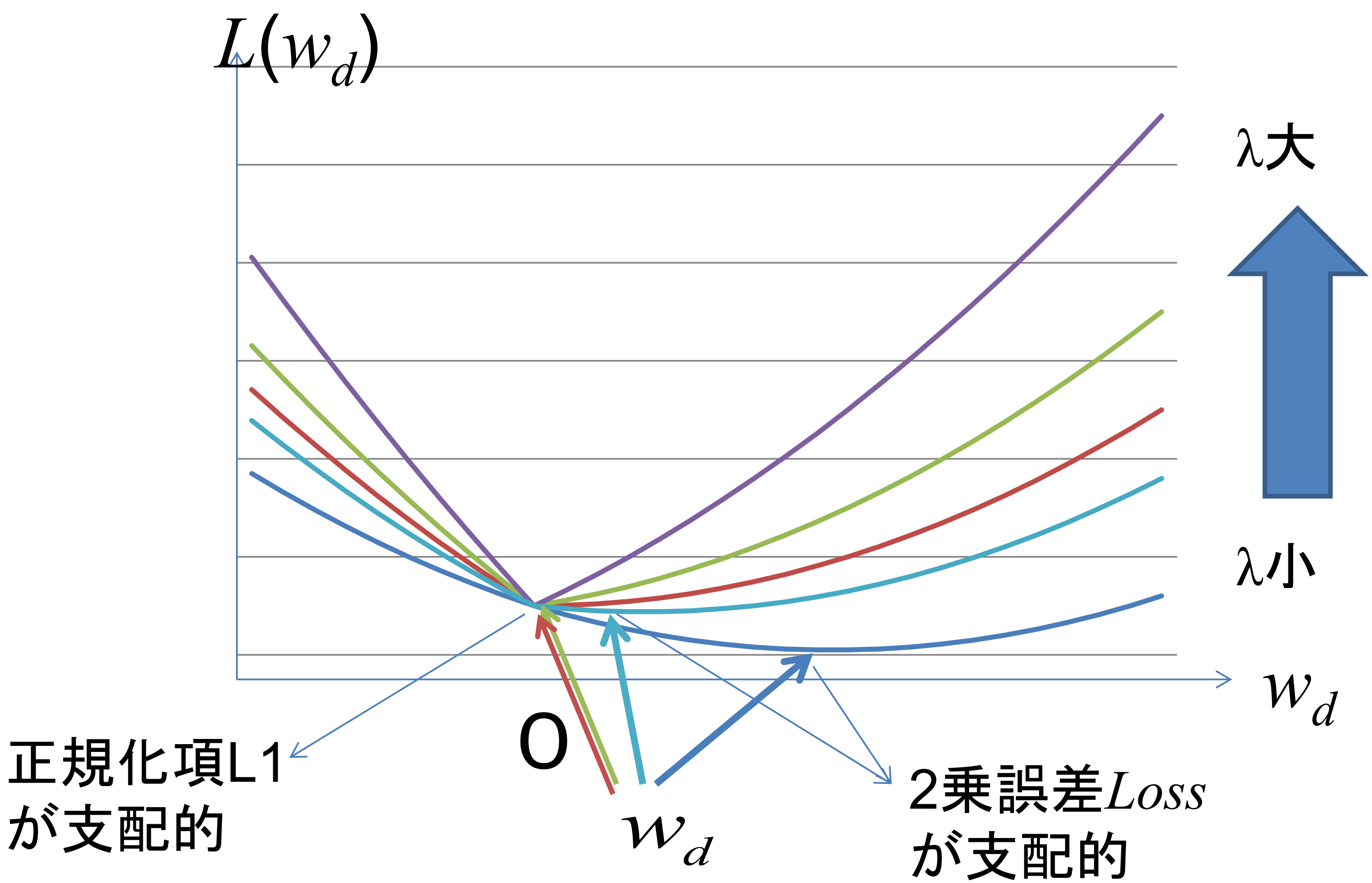
[step 5] w_kを更新

[step 6] 収束したら**w**の最終結果とする

w_dのゼロ化のイメージ



L1正則化が支配的になり$\hat{w}_d$ をゼロ化する様子を下図で例示する



正則化項のBayes的解釈

- Bayesでは事後確率は
  観測データの確率×事前確率
- 事後確率を最大化するパラメタηを求めたい

$$\hat{\eta} = \arg\max_{\eta} P(X \mid \eta) P(\eta \mid \alpha) \qquad \alpha\text{は事前分布のハイパーパラメタ}$$

- ここで対数尤度にしてみると、次のように解釈できる

$$\hat{\eta} = \arg\max_{\eta} \left(\underbrace{\log P(X \mid \eta)}_{\text{損失関数}} + \underbrace{\log P(\eta \mid \alpha)}_{\text{正則化項}} \right)$$

例:事前分布、事後分布とも正規分布

$$\mathbf{y} = (y_1, \ldots, y_N)^T \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \vdots \\ \mathbf{x_N}^T \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$$

$$y = \varphi(\mathbf{x})\mathbf{w} + \varepsilon \qquad \varepsilon = N(0,1)$$

$$\log \prod_i p(y_i \mid \mathbf{x}_i, \mathbf{w}, 1) = \sum_i \log N(y_i \mid \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle, 1) \propto \sum_i -\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 / 2$$

事前分布$p(\mathbf{w} \mid \alpha, \gamma)$も同様にすると

$$\log p(\mathbf{w} \mid \alpha, \gamma) \propto -(\mathbf{w} - \alpha)^T (\mathbf{w} - \alpha) / 2$$

$$\Rightarrow \quad \arg\min_{\mathbf{w}} \left(\log \prod_i p(y_i \mid \mathbf{x}_i, \mathbf{w}, \beta) + \log p(\mathbf{w} \mid \alpha, \gamma) \right)$$

$$\arg\min_{\mathbf{w}} \left(\frac{1}{2} \sum_i \left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 + \frac{1}{2} (\mathbf{w} - \alpha)^T (\mathbf{w} - \alpha) \right)$$

事前分布の$\mathbf{w}$の分散:λ^{-1} とも見える。

ここで、$\alpha = 0,$ 事前分布の重みをλとすると

$$\Rightarrow \quad \arg\max_{\mathbf{w}} \left(\frac{1}{2} \sum_i \left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \right)$$

L2ノルムによる正則化項

例：事前分布がLaplace分布、事後分布が正規分布

$$y = \varphi(\mathbf{x})\mathbf{w} + \varepsilon \qquad \varepsilon = N(0,1)$$

$$\log \prod_i p(y_i \mid \mathbf{x}_i, \mathbf{w}, 1) = \sum_i \log N(y_i \mid \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle, 1) \propto \sum_i -\left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 / 2$$

事前分布は期待値0の*Laplace*分布$p(\mathbf{w} \mid \lambda) = \dfrac{\lambda}{4}\exp\left(-\dfrac{\lambda|\mathbf{w}|}{2}\right)$も同様にすると

$$\log p(\mathbf{w} \mid \lambda) \propto -\frac{\lambda|\mathbf{w}|}{2}$$

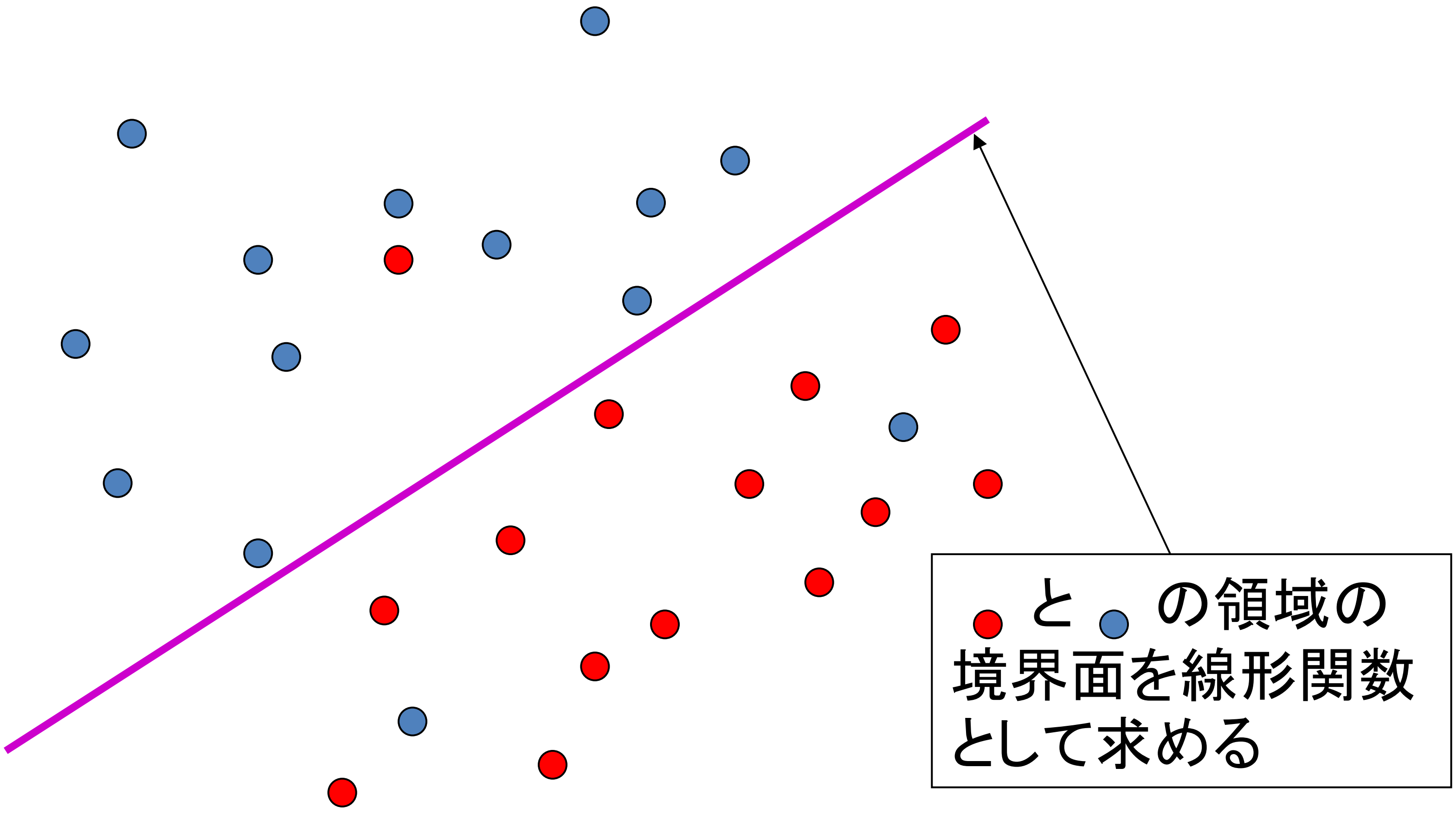
$$\Rightarrow \quad \arg\min_{\mathbf{w}} \left(\log \prod_i p(y_i \mid \mathbf{x}_i, \mathbf{w}, \beta) + \log p(\mathbf{w} \mid \lambda) \right)$$

$$= \arg\min_{\mathbf{w}} \left(\frac{1}{2} \sum_i \left(y_i - \langle \varphi(\mathbf{x}_i), \mathbf{w} \rangle\right)^2 + \frac{\lambda}{2}|\mathbf{w}| \right)$$

L1ノルムによる正則化項

- 以上、述べてきた線形回帰のよるモデル化は、生成モデル
- 当然、線形の識別モデルもある。次以降は線形識別モデルの話

線形識別



線形識別

- データ: $\mathbf{x} = [x_1, x_2, \ldots, x_M]^T$
- $\mathbf{x}$がいくつかのクラス(あるいはカテゴリー):C_kのどれかに属する。
  - 例:新聞記事が「政治」「経済」「スポーツ」「芸能」「社会」などのクラスのどれかに属する場合。この場合、データ:xは例えば、記事に現れる単語の集合、など。
- データ:$\mathbf{x}$がK個のクラスの各々に属するかどうかの判定は(−1=属さない, 1=属する)の2値を要素とするK次元ベクトル:y_i=(-1,1,-1,..,1)で表される。
  - ただし、1つのクラスに属するか属さないかだけを識別すの場合は2クラス分類という。当然、$y_i = -1$ or $y_i = 1$
- この属するか否かの判断をする式が線形の場合を線形識別という。

- 線形識別の関数

$$\boxed{y(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle + w_0}$$

あるいは　$\tilde{\mathbf{x}} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix}, \tilde{\mathbf{w}} = \begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix}$ とおくなら $y(\mathbf{x}) = \langle \tilde{\mathbf{x}}, \tilde{\mathbf{w}} \rangle$

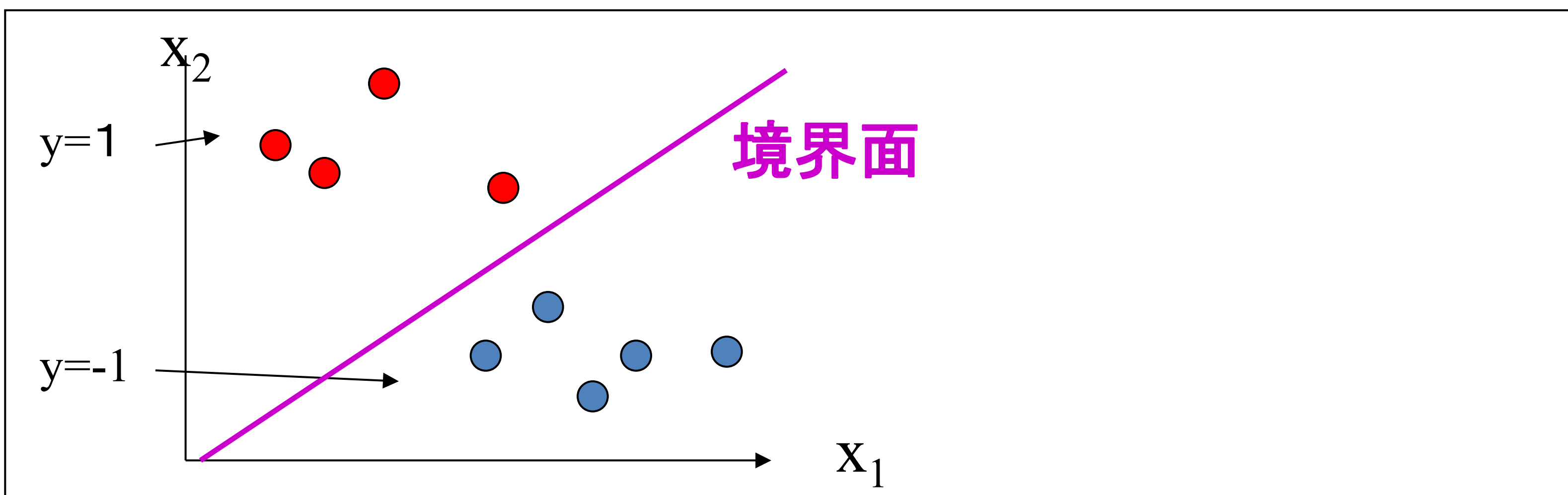
  - 一般化線形識別の関数は以下

$$y(\mathbf{x}) = f(\langle \mathbf{x}, \mathbf{w} \rangle + w_0)$$　f は非線形でもよい

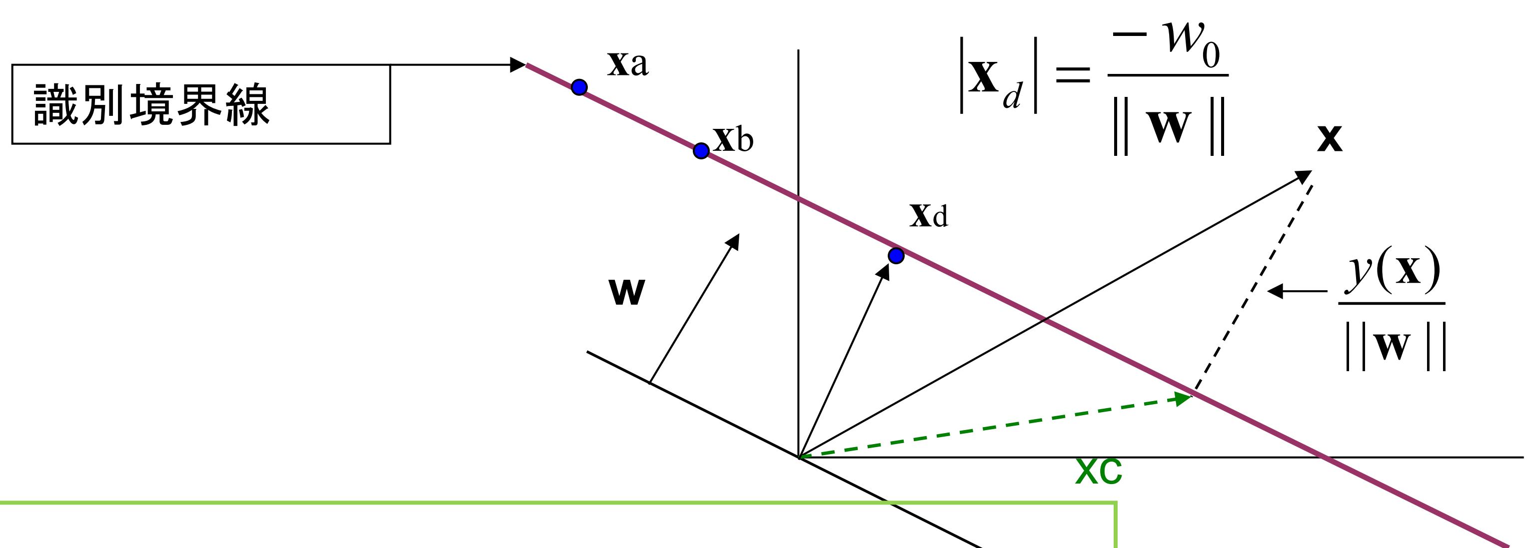
- ２クラス分類
- クラスC_1に属するかC_2（＝notC1）に属するかは、次の通り
- if $y(x) \geq 0$ then データ：xはC_1に属する
  otherwiseデータ：xはC_2に属する
  (すなわちC_1に属さない）

2値分類の直観的説明

- y={-1,1}、xは2次元とする。（下図を参照）
- {y,x}を教師データとして、2乗誤差の最小化を行って正規方程式を求めると、下図の のようなクラスを分類する分離平面が得られる。



線形識別関数の幾何学的解釈



$$y(\mathbf{x}_a) = \langle \mathbf{x}_a, \mathbf{w} \rangle + w_0 = 0, \quad y(\mathbf{x}_b) = \langle \mathbf{x}_b, \mathbf{w} \rangle + w_0 = 0$$

$$0 = y(\mathbf{x}_a) - y(\mathbf{x}_b) = \langle (\mathbf{x}_a - \mathbf{x}_b), \mathbf{w} \rangle$$

$\Rightarrow \quad \mathbf{w}$は$(\mathbf{x}_a - \mathbf{x}_b)$すなわち識別境界線と直交。

原点から識別境界線への垂線の交点を$\mathbf{x}_d$とおく。

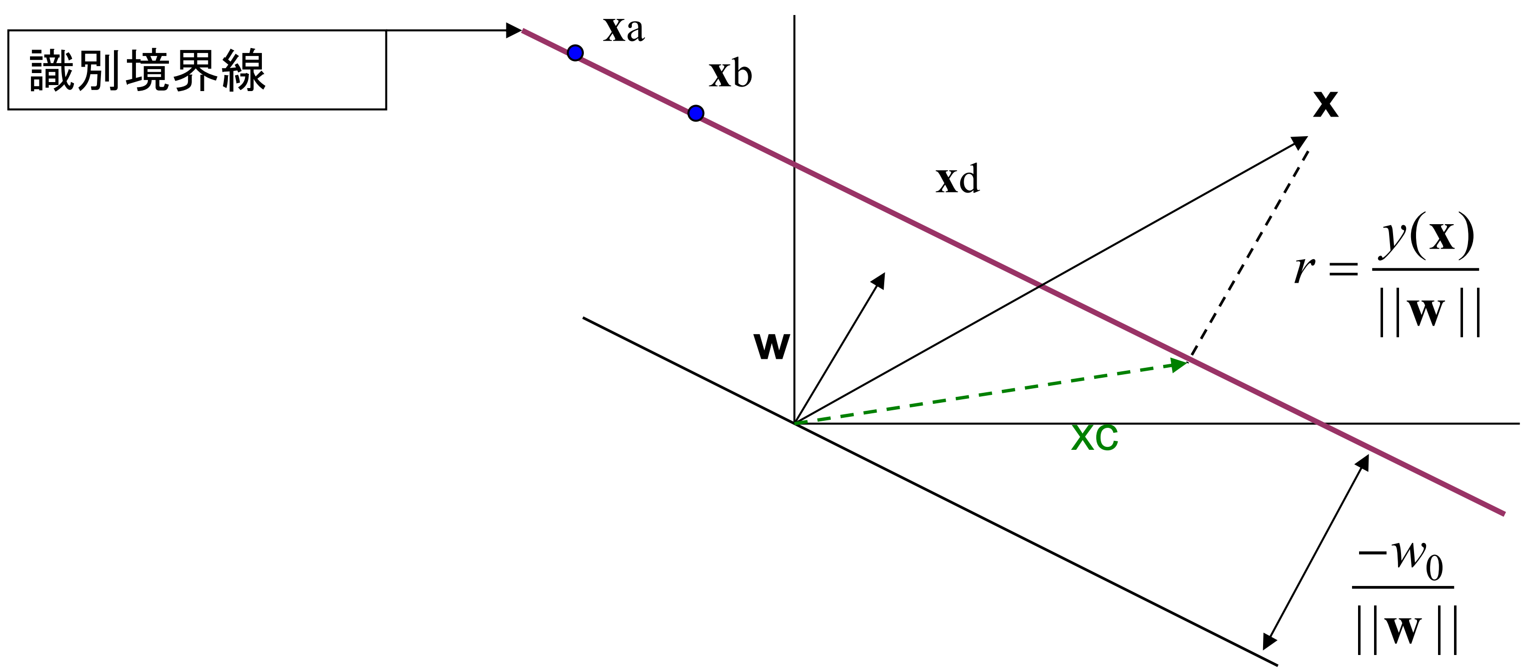
$$0 = y(\mathbf{x}_d) = \langle \mathbf{x}_d, \mathbf{w} \rangle + w_0$$

$\mathbf{x}_d$は$\mathbf{w}$に並行で横ベクトルだから、$\langle \mathbf{x}_d, \mathbf{w} \rangle = |\mathbf{x}_d\| \cdot \|\mathbf{w}\|$

これを上式に代入して整理すると

$$\langle \mathbf{x}_d, \mathbf{w} \rangle + w_0 = \|\mathbf{x}_d\| \cdot \|\mathbf{w}\| + w_0 = 0 \qquad \Rightarrow \quad |\mathbf{x}_d| = -\frac{w_0}{\|\mathbf{w}\|}$$

線形識別関数の幾何学的解釈



$\mathbf{x} = \mathbf{x}_c + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$　　両辺と$\mathbf{w}$の内積をとり、w_0を足すと

$$y(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle + w_0 = \langle \mathbf{x}_c, \mathbf{w} \rangle + w_0 + r\frac{\langle \mathbf{w}, \mathbf{w} \rangle}{\|\mathbf{w}\|} = y(\mathbf{x}_c) + r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|}$$

$y(\mathbf{x}_c) = 0$　　だから　$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$

wの計算方法:2クラス分類の場合

- . クラスC_1, C_2の境界が$y(\mathbf{x}) = \langle \tilde{\mathbf{x}}, \tilde{\mathbf{w}} \rangle$で書けるとする
- すると新規のデータ: $\mathbf{x}$は $y(\tilde{\mathbf{x}})$ が正ならクラス C_1に,負ならC_2属する

N個の教師データ$\{\tilde{\mathbf{x}}_n, y_n\}(n = 1, N)$があったとき

ただしクラス1なら$y_n = 1$,　0なら$y_n = -1$

- .

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{\mathbf{x}}_1{}^T \\ \vdots \\ \tilde{\mathbf{x}}_N{}^T \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \tilde{\mathbf{X}}\tilde{\mathbf{W}} = \begin{pmatrix} \langle \tilde{\mathbf{x}}_1, \tilde{\mathbf{w}} \rangle \\ \vdots \\ \langle \tilde{\mathbf{x}}_N, \tilde{\mathbf{w}} \rangle \end{pmatrix}$$

- すると、観測データ（教師データ）において個々のクラスに分類されたか否かの観点からの2乗誤差は次式となる

$$\boxed{E(\widetilde{\mathbf{W}}) = \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)}$$

- もう少し詳しく書くと

$$\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right) =$$

$$= \begin{pmatrix} \langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}} \rangle - y_1 & \cdots & \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}} \rangle - y_N \end{pmatrix} \begin{pmatrix} \langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}} \rangle - y_1 \\ \vdots \\ \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}} \rangle - y_N \end{pmatrix}$$

$$= \left(\langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}} \rangle - y_1\right)^2 + \cdots + \left(\langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}} \rangle - y_N\right)^2$$

$$E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{Y}\right)$$

- これを最小化する $\tilde{\mathbf{W}}$ は $\tilde{\mathbf{W}}$ で微分して0とおけば、線形回帰のときと同様の計算により求まる。
- 微分は次式：

$$\frac{\partial \mathbf{A}^T\mathbf{A}}{\partial \mathbf{W}} = 2\frac{\partial \mathbf{A}^T}{\partial \mathbf{W}}\mathbf{A} \rightarrow \mathbf{A} = \left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{Y}\right) \rightarrow \quad = 2\tilde{\mathbf{X}}^T\left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{Y}\right)$$

$$\frac{\partial E\left(\tilde{\mathbf{W}}\right)}{\partial \tilde{\mathbf{W}}} = \tilde{\mathbf{X}}^T\left(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{Y}\right) = 0$$

$$\Rightarrow \qquad \tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$$

- 新規のデータ$\mathrm{x}_{\mathrm{new}}$に対する予測を行う$\mathrm{y}(\mathrm{x}_{\mathrm{new}})$も求まる。

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$$

$$\mathbf{y}(\tilde{\mathbf{x}}_{new}) = \begin{bmatrix} y_1(\tilde{\mathbf{x}}_{new}) \\ \vdots \\ y_K(\tilde{\mathbf{x}}_{new}) \end{bmatrix} = \tilde{\mathbf{x}}_{new}\tilde{\mathbf{W}} = \tilde{\mathbf{x}}_{new}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$$

- $\mathbf{y}(\mathbf{x}_{\mathbf{new}})$が大きいほどクラス　C_1　に属する可能性が高い。

wの計算方法：多クラス分類の場合

- クラスC_kが線形識別モデル$y_k(\mathbf{x}) = \tilde{\mathbf{x}}\tilde{\mathbf{w}}_k$で書けるとする。
- すると新規のデータ：$\mathbf{x}$は $y_k(\tilde{\mathbf{x}})$ が最大のkのクラスC_kに属する

$$y_k(\mathbf{x})\text{を}K\text{個並べたベクトル}\mathbf{y} = [y_1(\mathbf{x}) \quad \cdots \quad y_K(\mathbf{x})]^T$$
$$= \left(\langle \tilde{\mathbf{x}}, \tilde{\mathbf{w}}_1 \rangle \quad \cdots \quad \langle \tilde{\mathbf{x}}, \tilde{\mathbf{w}}_K \rangle \right) = \tilde{\mathbf{x}}\tilde{\mathbf{W}}$$

- N個の教師データ$\{\tilde{\mathbf{x}}_n, \mathbf{y}_n\}(n = 1,.., N)$があったとき

注 $\tilde{\mathbf{x}}_n$はK個のクラス内の複数個に属することもあるなら
$\mathbf{y}_n$はK次元ベクトル$(-1,1,-1,...,1)$のような形。

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{\mathbf{x}}_1{}^T \\ \vdots \\ \tilde{\mathbf{x}}_N{}^T \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix} \qquad \tilde{\mathbf{X}}\tilde{\mathbf{W}} = \begin{pmatrix} \langle \tilde{\mathbf{x}}_1, \tilde{\mathbf{w}}_1 \rangle & \cdots & \langle \tilde{\mathbf{x}}_1, \tilde{\mathbf{w}}_K \rangle \\ \vdots & \ddots & \vdots \\ \langle \tilde{\mathbf{x}}_N, \tilde{\mathbf{w}}_1 \rangle & \cdots & \langle \tilde{\mathbf{x}}_N, \tilde{\mathbf{w}}_K \rangle \end{pmatrix}$$

- すると、観測データ(教師データ)において個々のクラスに分類されたか否かの観点からの2乗誤差は次式となる

$$\boxed{E(\widetilde{\mathbf{W}}) = Tr\left\{\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)\right\}}$$

- もう少し詳しく書くと

$$\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right) =$$

$$= \begin{pmatrix} \langle \widetilde{\mathbf{x}}_1 \widetilde{\mathbf{w}}_1 \rangle - y_{11} & & \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_1 \rangle - y_{N1} \\ \vdots & \ddots & \\ \langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}}_K \rangle - y_{1K} & & \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_K \rangle - y_{NK} \end{pmatrix} \begin{pmatrix} \langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}}_1 \rangle - y_{11} & & \langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}}_K \rangle - y_{1K} \\ & \ddots & \\ \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_1 \rangle - y_{N1} & & \langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_K \rangle - y_{NK} \end{pmatrix}$$

$$\therefore Tr\left(\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)\right)$$

$$= \left(\langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}}_1 \rangle - y_{11}\right)^2 + \cdots + \left(\langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_1 \rangle - y_{N1}\right)^2$$

$$+ \cdots + \quad \left(\langle \widetilde{\mathbf{x}}_1, \widetilde{\mathbf{w}}_K \rangle - y_{1K}\right)^2 + \cdots + \left(\langle \widetilde{\mathbf{x}}_N, \widetilde{\mathbf{w}}_K \rangle - y_{NK}\right)^2$$

$$E(\widetilde{\mathbf{W}}) = Tr\left\{\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)\right\}$$

- これを最小化する $\widetilde{\mathbf{W}}$ は $\widetilde{\mathbf{W}}$ で微分して0とおけば、線形回帰のときと同様の計算により求まる。
- Trの微分は次式:

$$\frac{\partial Tr(\mathbf{A}^T\mathbf{A})}{\partial \mathbf{W}} = 2\frac{\partial \mathbf{A}^T}{\partial \mathbf{W}}\mathbf{A} \rightarrow \mathbf{A} = \left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right) \rightarrow \quad = 2\widetilde{\mathbf{X}}^T\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right)$$

$$\frac{\partial E\left(\widetilde{\mathbf{W}}\right)}{\partial \widetilde{\mathbf{W}}} = \widetilde{\mathbf{X}}^T\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{Y}\right) = 0$$

$$\Rightarrow \qquad \widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{Y}$$

- 新規のデータ$\mathrm{x}_{\mathrm{new}}$に対する予測を行う$\mathrm{y}(\mathrm{x}_{\mathrm{new}})$も求まる。

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}$$

$$\mathbf{y}(\tilde{\mathbf{x}}_{new}) = \begin{bmatrix} y_1(\tilde{\mathbf{x}}_{new}) \\ \vdots \\ y_K(\tilde{\mathbf{x}}_{new}) \end{bmatrix} = \tilde{\mathbf{x}}_{new} \tilde{\mathbf{W}} = \tilde{\mathbf{x}}_{new} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}$$

- $\mathbf{y}_i(\mathbf{x}_{\mathbf{new}})$が大きいほどそのクラス i に属する可能性が高い。

もちろん、$\mathbf{y}_i(\mathbf{x}_{\mathbf{new}})$が最大となる$i$ のクラスに属すると考えるのが自然。だが。。。

生成モデルを利用した識別

- 識別はベイズ統計的には次式

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)p(C_k)}{p(\mathbf{x})}$$

- N個のデータ：x_k (k=1,..,N)があるクラスに属するかどうかの判定は（0＝属さない，１＝属する）の２値を要素とするN個のK次元ベクトル：y＝(0,1,0,..,1)で表される。
  - 以下のベイズ統計による分類では、属さない場合を-1ではなく0とすることに注意。
- 以下ではベイズ統計による2クラス分類をする場合に事後確率について考える。

Logistic sigmoid function

- クラスC_1の事後分布は次式(s-1)

$$p(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_1)p(C_1) + p(\mathbf{x} \mid C_2)p(C_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a) \qquad \text{-(s-1)} \qquad \text{logistc sigmoid function}$$

$$where \qquad a = \log\frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_2)p(C_2)}$$

$$\sigma(-a) = 1 - \sigma(a) \qquad a = \log\frac{\sigma}{1-\sigma}$$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \frac{\exp(-a)}{(1+\exp(-a))^2} = \frac{1}{1+\exp(-a)} \cdot \frac{\exp(-a)}{1+\exp(-a)} = \sigma(1-\sigma)$$

クラスC_1,C_2が共分散∑が等しい2つの正規分布の場合の事後確率 $p(C_1|x)$

- 式(s-1)によって以下のように導ける。

$$p(\mathbf{x}\mid C_i)=\frac{1}{(2\pi)^{K/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma^{-1}(\mathbf{x}-\mu_i)\right\}$$

$$\log\frac{p(\mathbf{x}\mid C_1)p(C_1)}{p(\mathbf{x}\mid C_2)p(C_2)}$$

$$=\frac{\log\left((2\pi)^{K/2}|\Sigma|^{1/2}\right)}{\log\left((2\pi)^{K/2}|\Sigma|^{1/2}\right)}\cdot\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\mu_2)\right)+\log\frac{p(C_1)}{p(C_2)}$$

$$=\left(\frac{1}{2}\left(\mathbf{x}^T\Sigma^{-1}\mu_1+\mu_1{}^T\Sigma^{-1}\mathbf{x}-\mu_1{}^T\Sigma^{-1}\mu_1\right)-\frac{1}{2}\left(\mathbf{x}^T\Sigma^{-1}\mu_2+\mu_2{}^T\Sigma^{-1}\mathbf{x}-\mu_2{}^T\Sigma^{-1}\mu_2\right)\right)+\log\frac{p(C_1)}{p(C_2)}$$

$$=\left(\mathbf{x}^T\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}\mu_1{}^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_2{}^T\Sigma^{-1}\mu_2\right)+\log\frac{p(C_1)}{p(C_2)}$$

∑が2つのクラスで等しいことにとってキャンセルしていることに注意。等しくないともう少し複雑。

クラスC_1,C_2が共分散∑が等しい2つの正規分布の場合の事後確率 $p(C_1|x)$

$$p(\mathbf{x}\,|\,C_i)=\frac{1}{(2\pi)^{K/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma^{-1}(\mathbf{x}-\mu_i)\right\}$$

$$\log\frac{p(\mathbf{x}\,|\,C_1)p(C_1)}{p(\mathbf{x}\,|\,C_2)p(C_2)}=\left(\mathbf{x}^T\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}\mu_1{}^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_2{}^T\Sigma^{-1}\mu_2\right)+\log\frac{p(C_1)}{p(C_2)}$$

$$\Rightarrow \text{事後確率}：p(C_1\,|\,\mathbf{x})=\sigma\left(\mathbf{w}^T\mathbf{x}+w_0\right)=\frac{1}{1+\exp\left(-\left(\mathbf{w}^T\mathbf{x}+w_0\right)\right)}$$

$$where\qquad \mathbf{w}=\Sigma^{-1}(\mu_1-\mu_2)$$

$$w_0=-\frac{1}{2}\mu_1{}^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_2{}^T\Sigma^{-1}\mu_2+\log\frac{p(C_1)}{p(C_2)}$$

∑が2つのクラスで等しいことにとってキャンセルしていることに注意。等しくないともう少し複雑。

次に Maximum likelihood solution （つまり$\mathbf{w}, w_0$)を求める。これによって、各クラスの事後確率が求まる

- ここで各クラスの事前確率が以下だったとする

$$p(C_1) = \pi \qquad p(C_2) = 1 - \pi$$

このとき観測データ$\mathbf{x}_n$がC_1に属するとき$t_n = 1$とし

$$p(\mathbf{x}_n, C_1) = p(C_1)p(\mathbf{x}_n \mid C_1) = \pi N\left(\mathbf{x}_n \mid \mu_1, \Sigma\right)$$

観測データ$\mathbf{x}_n$がC_2に属するとき$t_n = 0$とし

$$p(\mathbf{x}_n, C_2) = p(C_2)p(\mathbf{x}_n \mid C_2) = (1-\pi)N\left(\mathbf{x}_n \mid \mu_2, \Sigma\right)$$

ここで*likelihood*は次式　　観測データはN個あることを思い出そう

$$p(\mathbf{t} \mid \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[\pi N\left(\mathbf{x}_n \mid \mu_1, \Sigma\right)\right]^{t_n} \left[(1-\pi)N\left(\mathbf{x}_n \mid \mu_2, \Sigma\right)\right]^{1-t_n}$$

$$where \quad \mathbf{t} = \left(t_1, \ldots, t_N\right)^T \qquad \cdots(s-10)$$

- (s-10)のlogすなわち log likelihood function を最大化することが目標
- まず、最大化するπを求める。
- (s-10)のlogのπに関する部分は次式(s-20)　log$p$ (π)

$$\log p(\pi) = \sum_{n=1}^{N} \left\{ t_n \log \pi + (1 - t_n) \log(1 - \pi) \right\}$$

$$\frac{\partial \log p(\pi)}{\partial \pi} = 0 \qquad \boxed{\Rightarrow \quad \pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}}$$

where N_1 はクラス　C_1に属するデータ数。

N_2 はクラス　C_2に属するデータ数。

- 次に (s-10)の log を最大化する μ_1 を求める。
- (s-10)のlogのμ_2に関する部分は次式(s-30) $\log p\ (\mu_1)$

$$\log p(\mu_2) = \sum_{n=1}^{N}(1-t_n)\log N(\mathbf{x}_n \mid \mu_2, \Sigma) = -\frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n-\mu_2)^T\Sigma^{-1}(\mathbf{x}_n-\mu_2) + \mathrm{const}$$

$$\frac{\partial \log p(\mu_2)}{\partial \mu_2} = 0 \quad \Rightarrow \quad \boxed{\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathbf{x}_n}$$

- 同様にしてμ_1も求めると

$$\log p(\mu_1) = \sum_{n=1}^{N} t_n \log N(\mathbf{x}_n \mid \mu_1, \Sigma) = -\frac{1}{2}\sum_{n=1}^{N} t_n(\mathbf{x}_n-\mu_1)^T\Sigma^{-1}(\mathbf{x}_n-\mu_1) + \mathrm{const}$$

$$\frac{\partial \log p(\mu_1)}{\partial \mu_1} = 0 \quad \Rightarrow \quad \boxed{\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n\mathbf{x}_n}$$

- 最後に (s-10)の　log を最大化する精度行列 $\Lambda = \Sigma^{-1}$（C_1 とC_2共分散） を求める。

- (s-10)のlogのΣ に関する部分は次式(s-40)　$\log p\,(\Sigma)$

$$\log p(\Lambda) = \frac{1}{2}\sum_{n=1}^{N} t_n \log|\Lambda| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \mu_1)^T \Lambda (\mathbf{x}_n - \mu_1)$$
$$+ \frac{1}{2}\sum_{n=1}^{N} (1-t_n) \log|\Lambda| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n)(\mathbf{x}_n - \mu_2)^T \Lambda (\mathbf{x}_n - \mu_2)$$
$$= \frac{N}{2}\log|\Lambda| - \frac{N}{2} Tr(\Lambda S) \qquad \cdots (s-40)$$

- $\log p\,(\Lambda)$をΛ で微分して0とおき、(s-10)の　log を最大化するΛ $=\Sigma^{-1}$　を求める。

- まず第1項の微分は線形代数学の公式より

$$\frac{\partial\left(\frac{N}{2}\log|\Lambda|\right)}{\partial \Lambda} = \frac{N}{2}\left(\Lambda^{-1}\right)^T = \frac{N}{2}\left(\Lambda^{-1}\right) \quad \cdots (s-50)$$

$$\because \Lambda \text{が対称} \Rightarrow \Lambda^{-1} \text{が対称}$$

$(s-40)$のSは次式

$$S = \frac{1}{N}\sum_{n \in C_1}(\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^T + \frac{1}{N}\sum_{n \in C_2}(\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^T$$

- 次はTr(ΛS)をΛで微分して0とおき、$\log p$(Λ) を最大化するΛを求める。

$$\frac{\partial Tr(\Lambda S)}{\partial \Lambda} = -S^T = -S \qquad \cdots (s-60)$$

$$\therefore \frac{\partial \log p(\Lambda)}{\partial \Lambda} = \frac{N}{2}\Lambda^{-1} - \frac{N}{2}S = 0$$

$$\Rightarrow \quad \boxed{\Lambda^{-1} = \Sigma = S = \frac{1}{N}\sum_{n \in C_1}(\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^T + \frac{1}{N}\sum_{n \in C_2}(\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^T}$$

- このようにして、教師データ集合 $\{(\mathbf{x}_n, t_n) n=1,..N\}$ から μ_1, μ_2, Σ^{-1}(＝Λ), π が求まったので、これらを用いて定義される $\mathbf{w}, w_0$ も求まる。
- 未知データ $\mathbf{x}$ がクラス C_1 に属する確率は

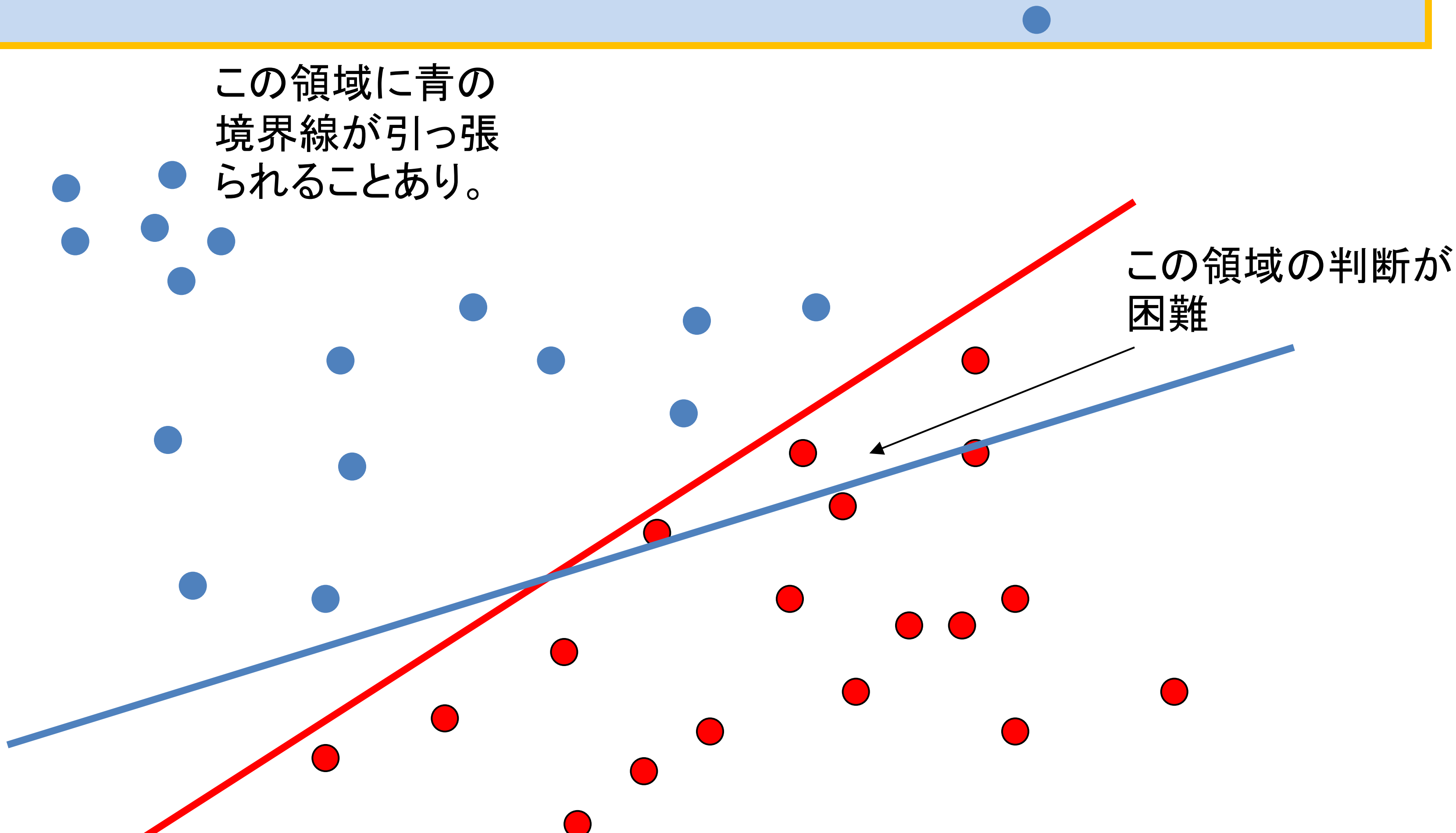
$$p(C_1 \mid \mathbf{x}) = \sigma\left(\mathbf{w}^T \mathbf{x} + w_0\right) = \frac{1}{1 + \exp\left(-\left(\mathbf{w}^T \mathbf{x} + w_0\right)\right)}$$

$$\text{where} \qquad \mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$$

$$w_0 = -\frac{1}{2}\mu_1{}^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2{}^T \Sigma^{-1} \mu_2 + \log\frac{\pi}{1-\pi}$$

なので、この分布を教師データから学習できた。

2乗誤差最小化の線形識別の問題点



そもそも、Yの値は正規分布を想定した理論なのに、{0、1}の2値しかとらないとして2乗誤差最小化を当てはめたところに無理がある。