数学概論（今野・2025/05/14）

最小2乗法と統計的回帰分析

講義の内容

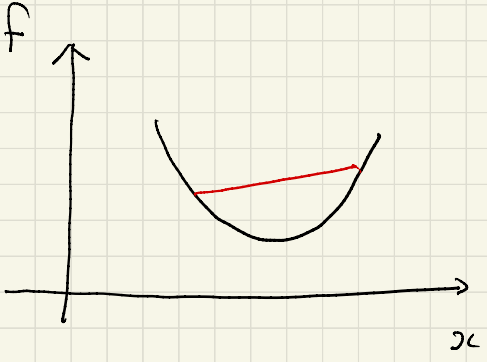
- 凸関数について
- 特異値分解と一般化逆行列
- 最小二乗法
- 最小二乗法の正則化：ridge 解と lasso 解

1. 凸関数の性質（$\mathbb{R}$上の関数）

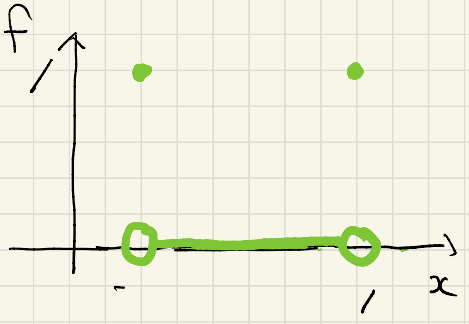
定義C.1 $I \subset \mathbb{R}$を区間とする.

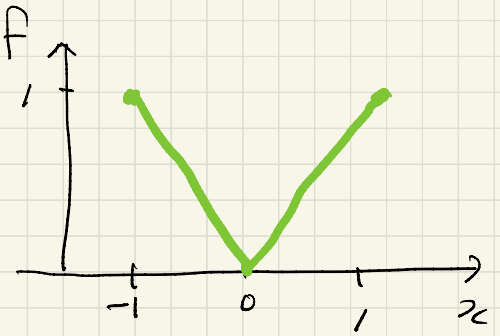
関数$f: I \to \mathbb{R}$は凸 $\overset{\text{def}}{\iff}$ $\forall x, y \in I$ と $0 \leq \forall a \leq 1$ に対して

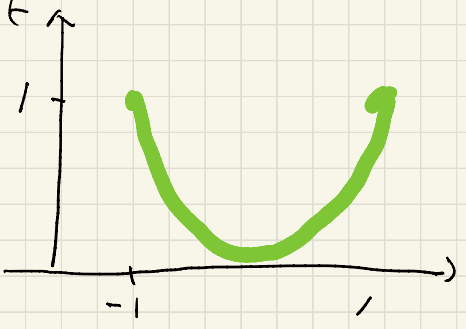
$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y).$$



例1　$I=[-1,1]$ とする.







定理 C.3　$I \subset \mathbb{R}$ を区間とする．I 上の凸関数は I の内部で連続．

定理 C.5　区間 (a, b) 上で凸関数 f は次のように表現される．

$$f(x) = f(c) + \int_c^x g(t)\,dt \qquad (a < x < b) \quad \cdots (a)$$

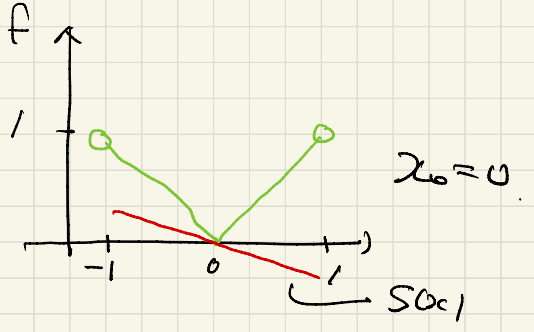
ただし，$c \in (a, b)$ で，g は単調増加で右連続．

注意　(a) の右辺の関数は凸．

定義 C.8 f を区間 (a,b) 上の関数とする．点 $x_0 \in (a,b)$ において，適当に $m \in \mathbb{R}$ を定めると，

$$S(x) := m(x-x_0) + f(x_0) \leq f(x) \quad (\forall x \in (a,b))$$

が成立するとき，$y = S(x)$ を点 x_0 において f を下から支える直線という



定理 C.9 $f:(a,b)\to\mathbb{R}$ とする. このとき

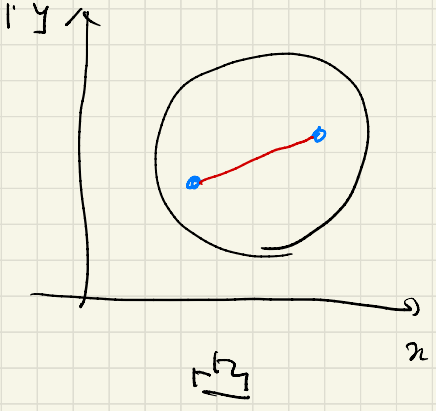
f は凸 $\Leftrightarrow$ 下から支える直線が x_0 で存在 $(\forall x_0\in(a,b))$.

2. 共役凸関数と Young の不等式

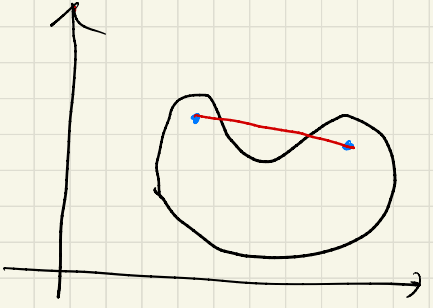
定義 C.10 $C \subset \mathbb{R}^n$ の空でない部分集合とする.

C は凸集合 $\overset{\text{def}}{\Leftrightarrow}$ $\forall x, y \in C$ と $0 \le \forall a \le 1$ について

$$ax + (1-a)y \in C$$



凸



凸でない

定義 C.22 $C \subset \mathbb{R}^n$ を凸部分集合とする.

関数 $f: C \to \mathbb{R}$ は凸 $\overset{\text{def}}{\Leftrightarrow}$ $\forall x, y \in \mathbb{R}^n$ と $0 \leq \forall a \leq 1$ について

$$f(ax + (1-a)y) \leq a f(x) + (1-a) f(y).$$

注意 C.23 $f: \mathbb{R}^n \to \mathbb{R}$ を関数とし.

$$\mathrm{epi}(f) := \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : y \geq f(x) \quad (x \in \mathbb{R}^n) \right\} \subset \mathbb{R}^{n+1}$$

を f の エピグラフという.

すると

f が凸関数 $\Longleftrightarrow$ $\mathrm{epi}(f)$ は凸集合. □

定義C.25 $f:\mathbb{R}^n\to\mathbb{R}$ を関数とする. 各 $x\in\mathbb{R}^n$ に対して

$$\partial f(x) := \{u\in\mathbb{R}^n;\ f(x)+\langle u, y-x\rangle_{2,n}\le f(y)\ (\forall y\in\mathbb{R}^n)\}$$

とおく. ただし, $\langle\cdot,\cdot\rangle_{2,n}$ は $\mathbb{R}^n$ 上の自然な内積とする.

この集合を点 x における関数 f の劣微分という

注意 ・fは$\mathbb{R}^n$上の凸関数 $\Leftrightarrow$ $\partial f(x) \neq \phi$ $(\forall x \in \mathbb{R}^n)$.

特に，凸関数fが点$x_0 \in \mathbb{R}^n$で微分可能ならば

$$\partial f(x_0) = \{\nabla f(x_0)\}.$$

ただし $\partial f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)^T$, $x = (x_1, \cdots, x_n)^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$.

例 $f(x) = |x|$ $(x \in \mathbb{R})$とする．すると

$$\partial f(x) = \begin{cases} \{1\} & (x > 0) \\ [-1, 1] & (x = 0) \\ \{-1\} & (x < 0) \end{cases}$$

fは凸 $\Leftrightarrow$

定理C.28 $f: \mathbb{R}^n \to \mathbb{R}$とする．各$x \in \mathbb{R}^n$において，$\partial f(x) \neq \emptyset$

注意 (1) $f: \mathbb{R}^n \to \mathbb{R}$ を凸関数とし，$x_0 \in \mathbb{R}^n$ とする．

$$0_n \in \partial f(x_0) \iff y = f(x) \text{ は } x = x_0 \text{ で最小}.$$

(2) $f: \mathbb{R}^n \to \mathbb{R}$ を凸とし，$x_0 \in \mathbb{R}$ とする．このとき

ある $\varepsilon > 0$ が存在して

$$f(x_0) \leq f(x) \quad (\forall x \in B(x_0, \varepsilon) = \{y \in \mathbb{R}^n ; |x_0 - y|_{2,n} \leq \varepsilon\})$$

すなわち，

$$f(x_0) \leq f(x) \quad (\forall x \in \mathbb{R}^n).$$

仮定　$f: \mathbb{R}^n \to \mathbb{R}$ は

$$\lim_{\|x\|_{2,n} \to \infty} \frac{f(x)}{\|x\|_{2,n}} = +\infty$$

をみたすとき，f は超線型成長条件をみたすという

例　$f(x) = |x|$ $(x \in \mathbb{R})$ はダメ．

定義 C.3.1　$f: \mathbb{R}^n \to \mathbb{R}$ は超線型条件をみたす凸関数とする．このとき，f の凸共役関数 f^* を

$$f^*(y) = \max_{x \in \mathbb{R}^n} \{ \langle x, y \rangle_{2,n} - f(x) \} \quad (y \in \mathbb{R}^n)$$

で定める．

例 C.32 (2) $1 < p < \infty$ とし，$f(x) = |x|^p/p$ $(x \in \mathbb{R})$

とする．すると

$$f^*(y) = \frac{|y|^q}{q}$$

となる．ただし，$p^{-1} + q^{-1} = 1$ である．

補題 C.33（Fenchel-Young の不等式）定義 C.31 の設定を

仮定する．

$$\langle x, y \rangle_{2,n} \leq f(x) + f^*(y).$$

定理C.34　定義C.31の設定を仮定する.

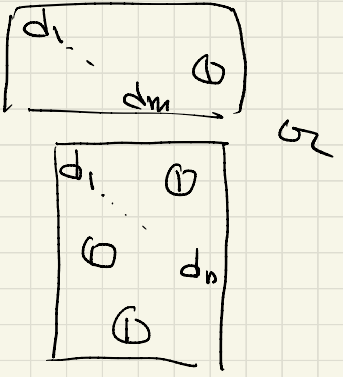
(1) $f^{\vee}: \mathbb{R}^n \to \mathbb{R}$ は凸関数

(2) $\lim_{|y|_{2,n} \to \infty} \dfrac{|f^{\vee}(y)|}{|y|_{2,n}} = +\infty$,

(3) $(f^{\vee})^{\vee} = f$.

A：特異値分解（The singular Value Decomposition = SVD）

定理 A.7 $m, n \in \mathbb{N}$, $A \in Mat(m,n;\mathbb{R})$ は次の分解を持つ.



$$A = U D V^T,$$

$U \in Mat(m;\mathbb{R})$, は直交行列

$V^T \in Mat(n;\mathbb{R})$ は直交行列

$D \in Mat(m,n;\mathbb{R})$ は「対角行列」で,

その対角成分は $d_1 \geq d_2 \geq \cdots \geq d_{\min(m,n)} \geq 0$.

B. Moore-Penrose の一般化逆行列

$A \in \mathrm{Mat}(m, n; \mathbb{R})$ とする. A の Moore-Penrose の一般化逆行列 $A^{\dagger} \in \mathrm{Mat}(n, m; \mathbb{R})$ を次の4条件を満たすとする.

$$A A^{\dagger} A = A, \qquad (B.3)$$

$$A^{\dagger} A A^{\dagger} = A^{\dagger}, \qquad (B.4)$$

$$(A A^{\dagger})^{T} = A A^{\dagger} \qquad (B.5)$$

$$(A^{\dagger} A)^{T} = A^{\dagger} A \qquad (B.6).$$

注意 (1) 一意的に存在する.

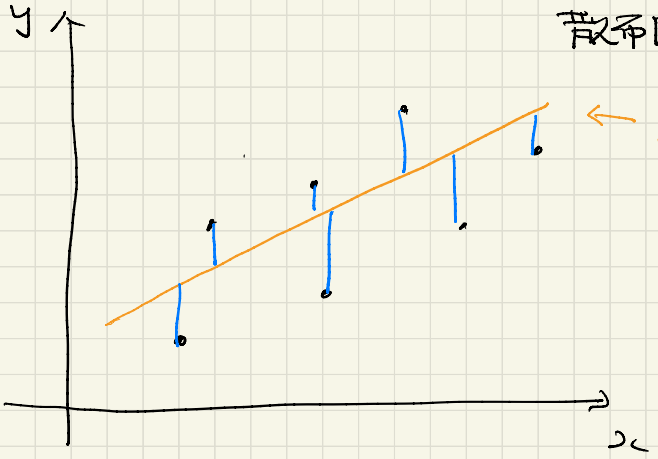
(2) $m \geq n$ かつ A^TA は正則のとき.

$$A^+ = (A^TA)^{-1}A^T$$

となる.

1. 最小2乗法の考え方

$n \geq 2$ とし、$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ を観測値とする



散布図

← $y = \beta_1 + \beta_2 x$ と書く。

↑

x の情報を使って

y の動きを説明しよう！

↓

単回帰

1変量で1変量を説明

$\hat{y}_j := \beta_1 + \beta_2 x_j \quad (j = 1, 2, \cdots, n)$

$e_j := y_j - \hat{y}_j$ ：残差

最小2乗法

$$\min_{\beta_1,\beta_2} \sum_{j=1}^{n} \varepsilon_j^2 \quad \leftarrow \text{残差平方和の最小化}$$

すなわち,

$$\begin{cases} n\beta_1 + \left(\sum_{j=1}^{n} x_j\right)\beta_2 = \sum_{j=1}^{n} y_j \\ \left(\sum_{j=1}^{n} x_j\right)\beta_1 + \left(\sum_{j=1}^{n} x_j^2\right)\beta_2 = \sum_{j=1}^{n} x_j y_j \end{cases}.$$

$x_1 = \cdots = x_n$ 以外は解が一意に定まる.

多次元への拡張（重回帰）

$(x_1, y_1^{data}), \ldots, (x_n, y_d^{data})$ を観測したとする

$x_j \in \mathbb{R}^d$（縦ベクトル），$y_j^{data} \in \mathbb{R}$ $(j = 1, \ldots, n)$

記号

$$\underline{X}^{data} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in Mat(n, d; \mathbb{R}),\ y^{data} = \begin{pmatrix} y_1^{data} \\ \vdots \\ y_n^{data} \end{pmatrix} \in \mathbb{R}^n,\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^d$$

$n \geq d$ かつ $y^{data} \notin span(x_1, \ldots, x_n)$ とする．

すなわち，$y^{data} = \underline{X}^{data}\beta$ をみたす $\beta \in \mathbb{R}^d$ は存在しない！

残差ベクトル

$$\varepsilon := X\beta - y^{data}$$ ← 方程式 $y = X\beta$ を定めること

と定める. 以後は"data"を取る.

最小二乗解 $\hat{\beta}^{ls}$ を

$$|X\hat{\beta}^{ls} - y|_{2,n} \leq |X\beta - y|_{2,n} \quad (\forall \beta \in \mathbb{R}^q)$$

で定める. (存在はするが一意とは限らない)

最小二乗解を求める.　　$\beta = (\beta_1, \ldots, \beta_d)^T$

$$f(\beta) := |X\beta - y|_{2,n}^2, \qquad \nabla := \left(\frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_d}\right)^T$$

すると, $\hat{\beta}^{ols}$は

$$\nabla f(\beta) = 2X^T(X\beta - y) = 0_d \iff \nabla f(\hat{\beta}^{ols}) = 0_d$$

をみたす.　　$\iff X^T X \hat{\beta}^{ols} = X^T y$

$X^T X$ が正則のとき

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

となる.

最小二乗法によるデータのあてはめ

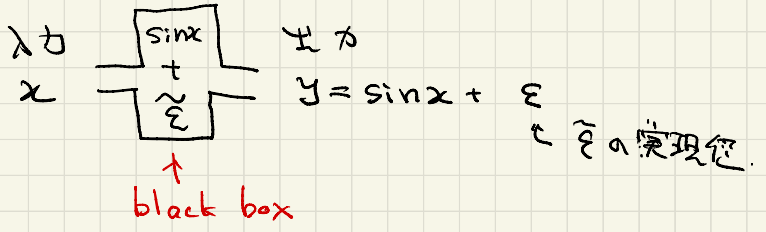
回帰モデル

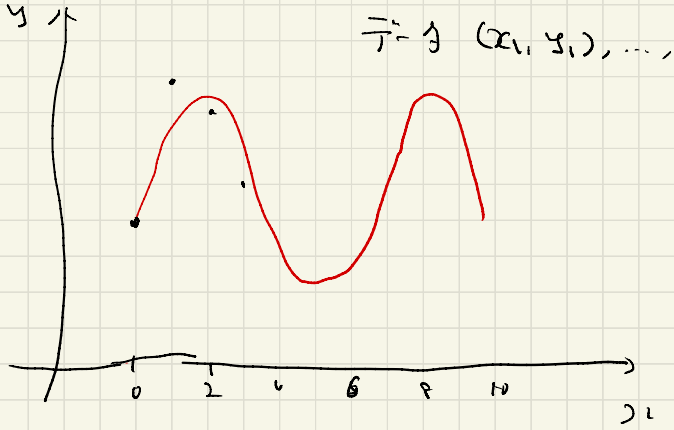
$$\tilde{y} = \sin x + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim N(0, 0.04)$$

（平均0，分散(0.2)²の正規分布）

人間には未知

を考える．xは固定した値，$\tilde{y}$と$\tilde{\varepsilon}$は確率変数





データ $(x_1, y_1), \cdots, (x_{11}, y_{11})$

多項式のあてはめ

$$f(x) = \beta_1 + \beta_2 x + \cdots + \beta_{11} x^{10}$$

$y = X\beta +$ 誤差ベクトル

$y = (y_1, \cdots, y_{11})^T$, $\beta = (\beta_1, \cdots, \beta_{11})^T$

$$\underline{X}_d = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{d-1} \\ 1 & x_2 & \cdots & x_2^{d-1} \\ & \vdots & & \\ 1 & x_{11} & \cdots & x_{11}^{d-1} \end{pmatrix} \in \mathrm{Mat}(11, d; \mathbb{R}).$$

$\beta_d := (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ $\quad (d = 2, 3, \ldots, 11)$

$X_{11} = X$, $\beta_{11} = \beta$ とおく.

$\hat{\beta}_d^{(d)} = (\underline{X}_d^T \underline{X}_d)^{-1} \underline{X}_d^T y$ とし, $\hat{\beta}_d^{(d)}$ を係数とする

5次の多項式を考えるとうまくいく.

↑ 答えを知っているからわかる.

$\min_{\beta \in \mathbb{R}^d} |X\beta - y|_{2,n}^2$ ← 曲線のデータへの適合度

を考えてもだめ．すなわち，残差平方和を最小化しても無駄．

変数選択（AIC, BIC） ⇒ β のいくつかの成分は「0」と見う．

$\hat{\beta}$ ：データ $(x_1, y_1), \ldots, (x_n, y_n)$ から作った多項式の係数

(y^{future}, x^{future}) ：将来のデータ．

$x^{future} \Rightarrow x^{future} \Rightarrow |y^{future} - (x^{future})^T \hat{\beta}|^2$

$x \mapsto \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{10} \end{bmatrix}$

を小さくする $\hat{\beta}$ がよいもの．

d が大きいとき，brutal search，

・正則化法

(曲線のデータの適合度) + (βの複雑さへの罰則)

βが複雑 ↑ ←こえん

最適化

βに対する罰則　　$\beta = (\beta_1, \ldots, \beta_M)$

(a) $|\beta|_{0,M} := (\beta$ の零でない成分の個数$)$

(b) $|\beta|_{1,M} := \sum_{j=1}^{M} |\beta_j|$

(c) $|\beta|_{2,M}^2 := \sum_{j=1}^{M} \beta_j^2$

(a)は β に関する非凸問題だが、(b)と(c)は凸問題となる

$$\min_{\beta \in \mathbb{R}^n} |X\beta - y|_{2,n}^2 + r|\beta|_{2,n}^2 \longrightarrow \hat{\beta}^{ridge}$$

$$\min_{\beta \in \mathbb{R}^n} |X\beta - y|_{2,n}^2 + 2r|\beta|_{1,n} \longrightarrow \hat{\beta}^{lasso}$$

ただし、$r > 0$ は 正則化変数 ⇐ 人間が決める