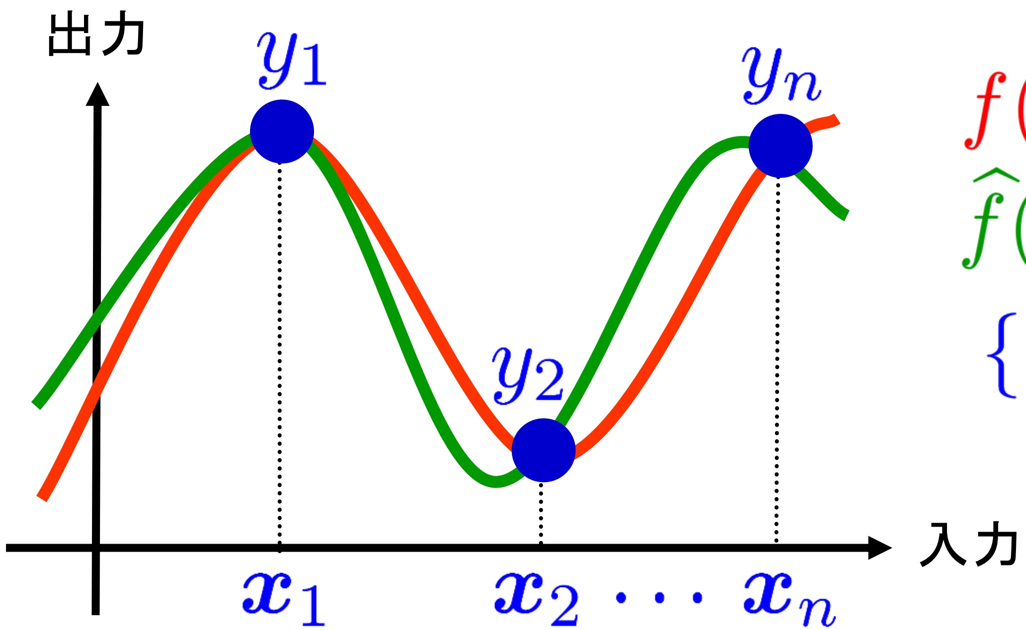
最小二乗回帰

杉山将・本多淳也
sugi@k.u-tokyo.ac.jp, jhonda@k.u-tokyo.ac.jp
http://www.ms.k.u-tokyo.ac.jp

回帰 = 関数近似



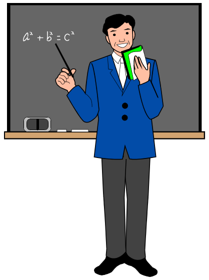
$f(\boldsymbol{x})$:学習したい真の関数

$\widehat{f}(\boldsymbol{x})$:学習結果の関数

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$:訓練標本

$$y_i = f(\boldsymbol{x}_i) \ (+\text{noise})$$

訓練標本から真の関数にできるだけ近い関数を求める



講義の流れ

1. 学習モデル(2章)
   - A) 線形モデル
   - B) カーネルモデル
   - C) 非線形モデル
2. 最小二乗回帰(3章)
3. 正則化回帰(4章)

線形／非線形モデル

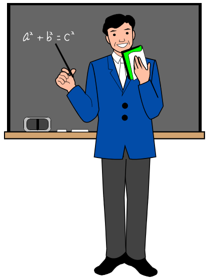
- モデル:学習結果の関数を探す候補集合
  - パラメータ $\boldsymbol{\theta}$ の値を指定すると関数が一つ決まる

$$\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \mid \boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)^\top\}$$

- 線形モデル: $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ が $\boldsymbol{\theta}$ に関して線形

（注:入力 $\boldsymbol{x}$ に関して線形である必要はない）

- 非線形モデル:それ以外



講義の流れ

1. 学習モデル(2章)
   - A) 線形モデル
   - B) カーネルモデル
   - C) 非線形モデル
2. 最小二乗回帰(3章)
3. 正則化回帰(4章)

線形モデル

- 一般形：
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$$
$$\boldsymbol{x} \in \mathbb{R}^d$$
- $\{\phi_j(\boldsymbol{x})\}_{j=1}^{b}$：線形独立な（既知の）基底関数
- 入力次元 $d=1$ のときの基底関数の例：
  - 多項式基底：
  $$1, x, x^2, \ldots, x^{b-1}$$
  - 三角多項式基底：
  $$1, \sin x, \cos x, \ldots, \sin kx, \cos kx$$
  $$b = 2k+1$$

多次元の線形モデル

- 加法モデル:

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{d} \sum_{j=1}^{b'} \theta_{k,j} \phi_j(x^{(k)})$$

- パラメータ数は $b = b'd$:
  - 入力次元 d に対して線形にしか増加しない（後述の乗法モデルでは指数関数的に増加）
- しかし，一次元の基底関数の和しか考えないため，交互作用のある複雑な関数が表現できない

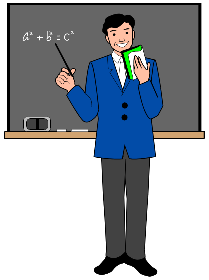
多次元の線形モデル

- 多次元入力に対する乗法モデル:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j_1=1}^{b'} \cdots \sum_{j_d=1}^{b'} \theta_{j_1,\ldots,j_d} \phi_{j_1}(x^{(1)}) \cdots \phi_{j_d}(x^{(d)})$$

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

- パラメータ数は $b = (b')^d$:
  - 入力次元 d に対して指数的に増加する
  - $b' = 10, \ d = 100$ の時には $b = 10^{100}$
  - 計算量的にも統計的にも非現実的
- d が小さい場合しか使えない
（ただしスパース性が使える場合を除く）



講義の流れ

1. 学習モデル(2章)
   A) 線形モデル
   B) カーネルモデル
   C) 非線形モデル
2. 最小二乗回帰(3章)
3. 正則化回帰(4章)

カーネルモデル

線形モデル:

- 基底関数$\{\phi_j(\boldsymbol{x})\}_{j=1}^b$は訓練標本$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$に依存しない

カーネルモデル:

- 基底関数の設計に訓練入力標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ を用いる

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^n \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

- 例:ガウスカーネル

$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2h^2}\right)$$

$h(> 0)$:バンド幅

カーネルモデル

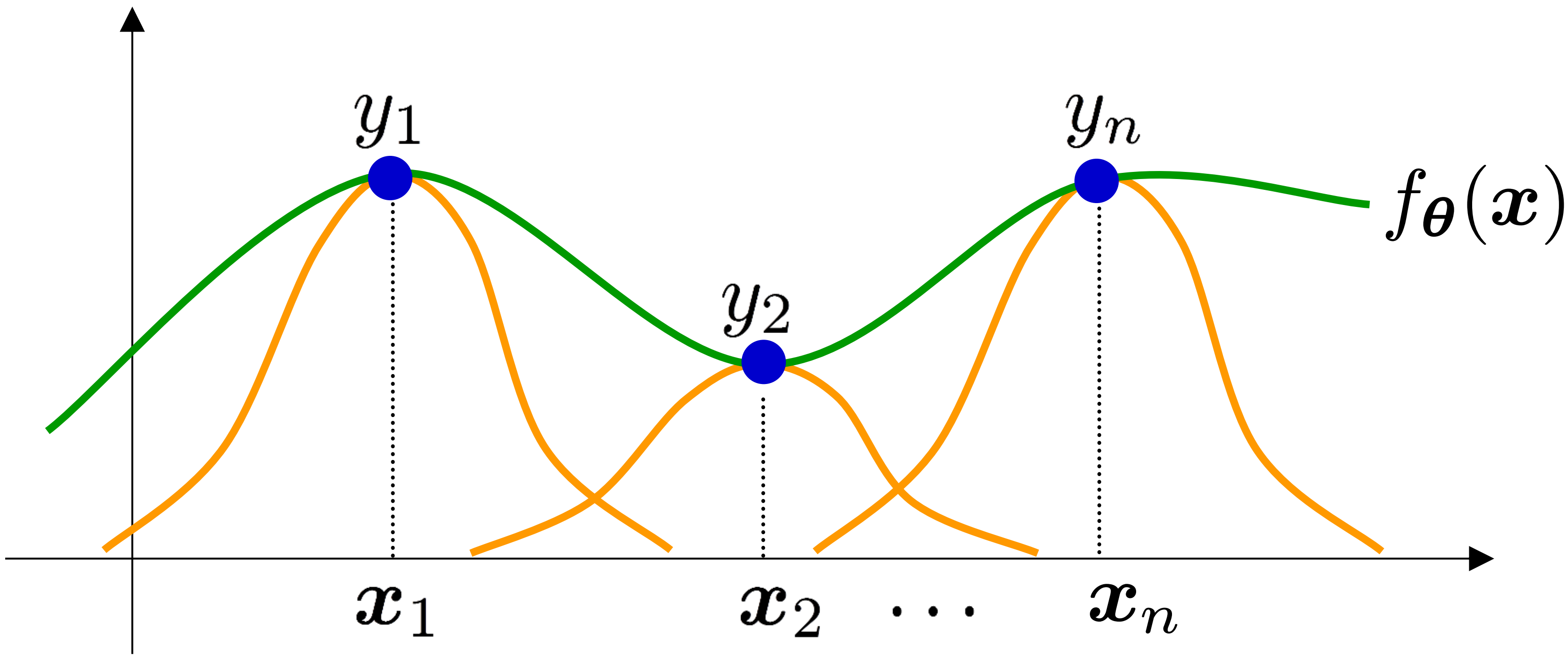
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

- パラメータ数は n であり，入力次元 d に依存しない
- パラメータに関しては線形：
  - カーネルモデルも線形モデルの一種（→計算が容易）
- しかし，パラメータ数が訓練標本数に依存するため，通常の線形モデルとは統計的な性質が異なる
  - そのため，ノンパラメトリックモデルとよばれる
- ただし，本講義の範囲では，カーネルモデルを線形モデルとみなしてもおおよそ問題ない

ガウスカーネルモデル

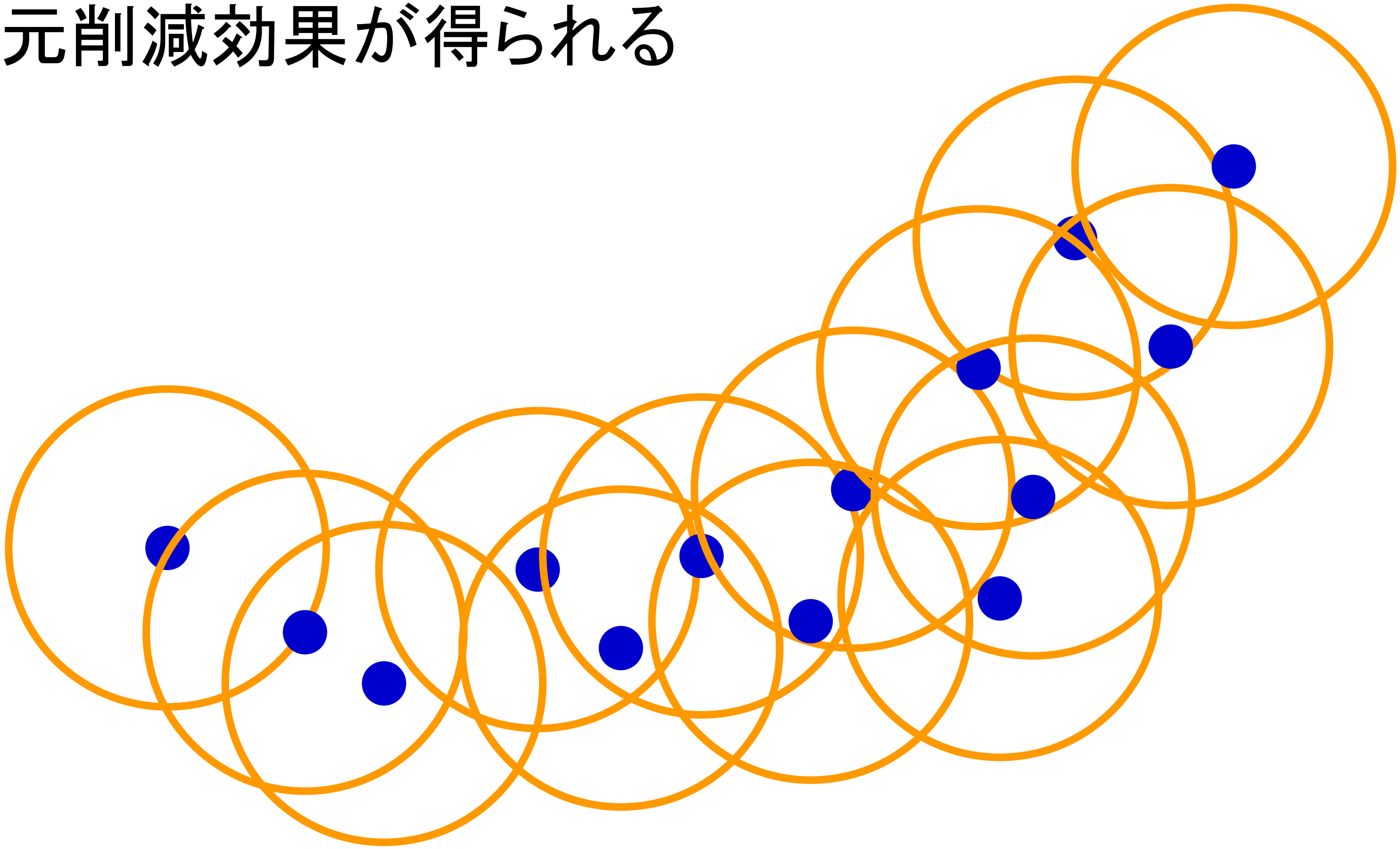
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_j\|^2}{2h^2} \right)$$

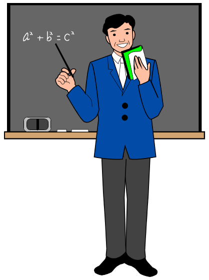
- ガウス関数を各訓練入力標本の場所に配置



ガウスカーネルモデル

- 訓練標本が入力空間上に偏って分布している時，ガウスカーネルモデルは訓練入力標本が存在しない領域を自動的に無視する
  - 次元削減効果が得られる





講義の流れ

1. 学習モデル(2章)
   - A) 線形モデル
   - B) カーネルモデル
   - C) 非線形モデル
2. 最小二乗回帰(3章)
3. 正則化回帰(4章)

非線形モデル

- パラメータに関して線形でないモデル
  - 線形モデルで基底関数がパラメータを含む場合

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \alpha_j \phi(\boldsymbol{x}; \boldsymbol{\beta}_j)$$

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_b^\top)^\top$$

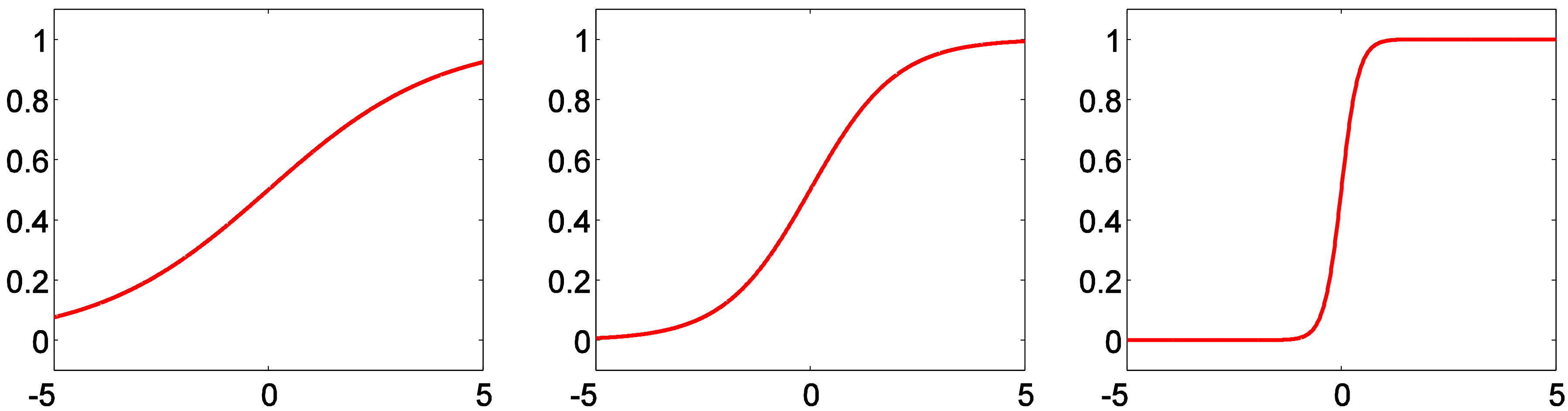
$\phi(\boldsymbol{x}; \boldsymbol{\beta})$: $\boldsymbol{\beta}$ について非線形

パラメータを含む基底関数の例1

■シグモイド関数：

$$\phi(\boldsymbol{x};\boldsymbol{\beta}) = \frac{1}{1+\exp\left(-\boldsymbol{x}^\top \boldsymbol{w} - \gamma\right)}$$

$$\boldsymbol{\beta} = (\boldsymbol{w}^\top, \gamma)^\top$$



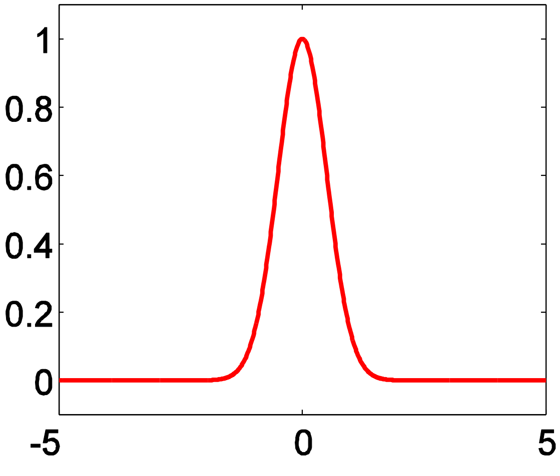
w：小　　w：中　　w：大

パラメータを含む基底関数の例2

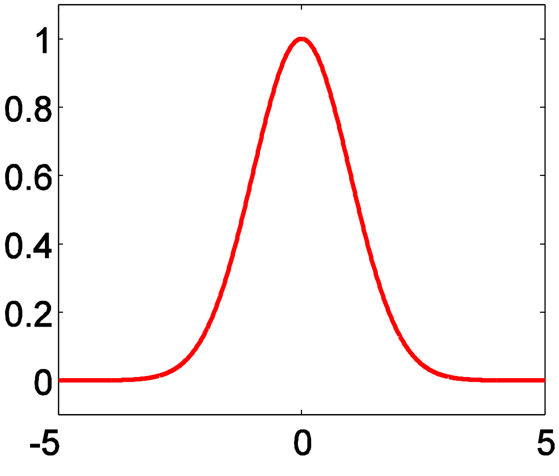
- **ガウスカーネル**: 線形のガウスカーネルモデルと異なり，ガウス関数の中心や幅も学習する

$$\phi(\boldsymbol{x};\boldsymbol{\beta}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{2h^2}\right)$$

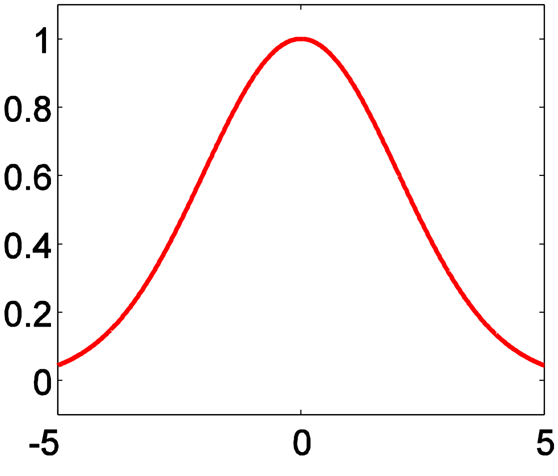
$$\boldsymbol{\beta} = (\boldsymbol{c}^\top, h)^\top$$



h : 小



h : 中



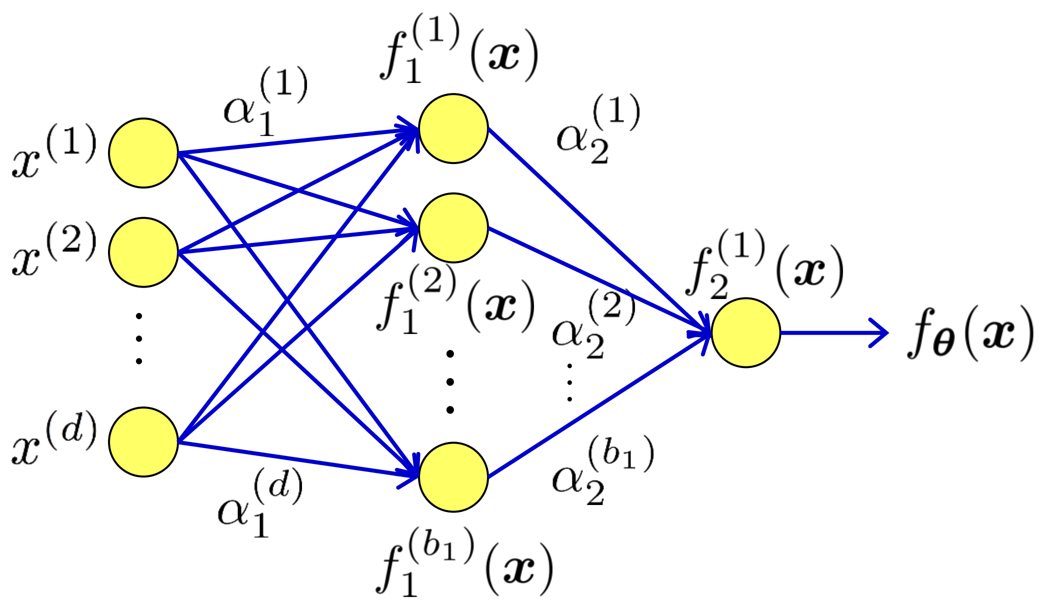
h : 大

非線形モデル

- パラメータに関して線形でないモデル
  - 階層モデル

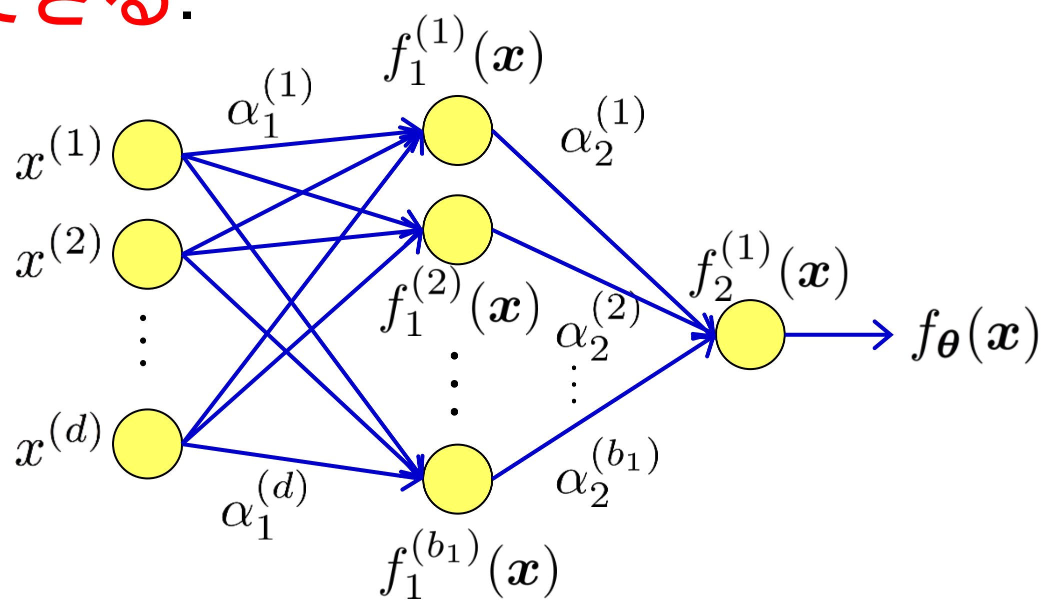
$$f_k^{(i)}(\boldsymbol{x}) = \sum_{j=1}^{b_{k-1}} \alpha_k^{(j)} \phi(f_{k-1}^{(j)}(\boldsymbol{x})), \qquad f_0^{(i)}(\boldsymbol{x}) = x^{(i)}$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_K^{(1)}(\boldsymbol{x}) \qquad \boldsymbol{\theta} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_K^\top)^\top$$



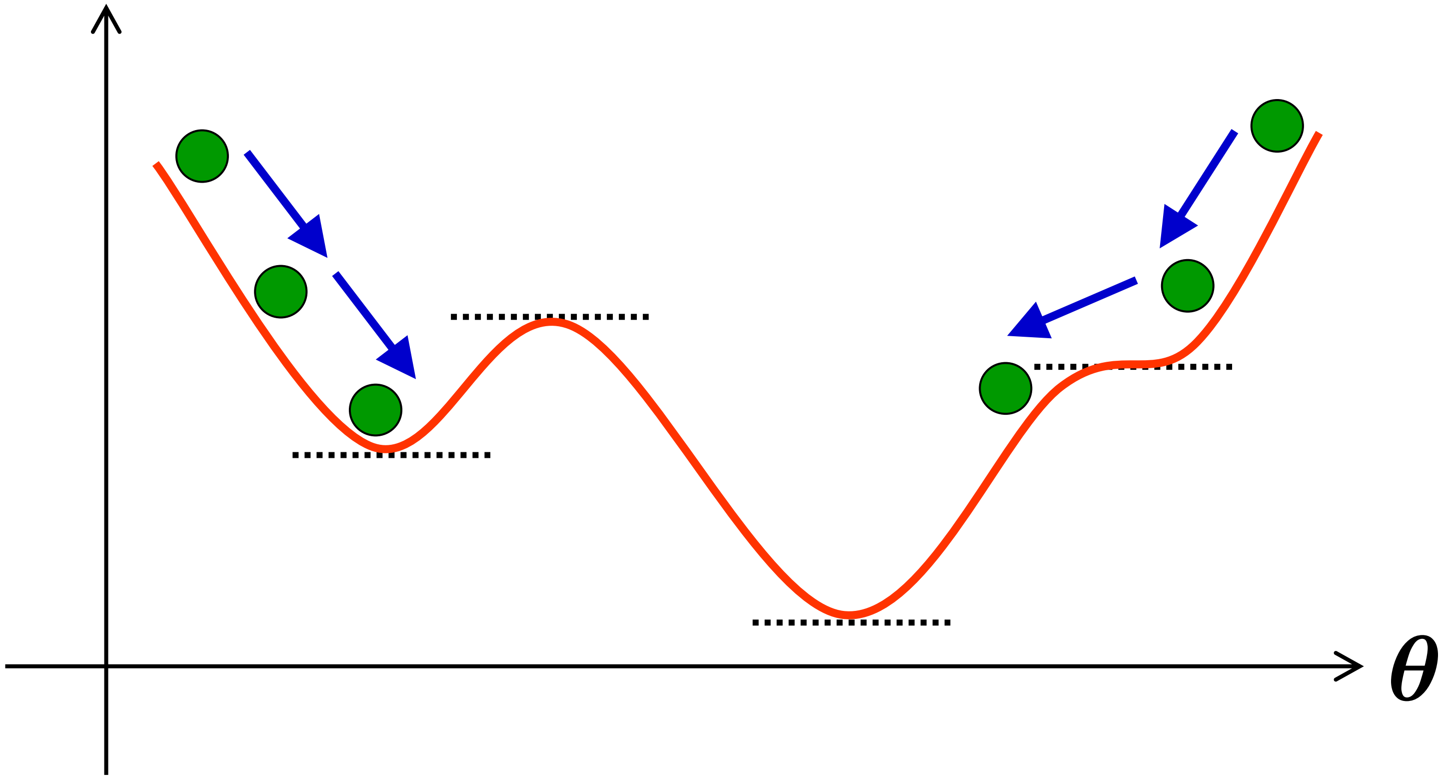
パーセプトロン

- 人間の脳は無数の神経細胞が網目状につながっている.
- シグモイド関数は神経細胞の振る舞いと似ている.
- 人工神経回路網はパーセプトロンとも呼ばれる.
- 数学的には，三層パーセプトロンによって任意の関数を近似できる.



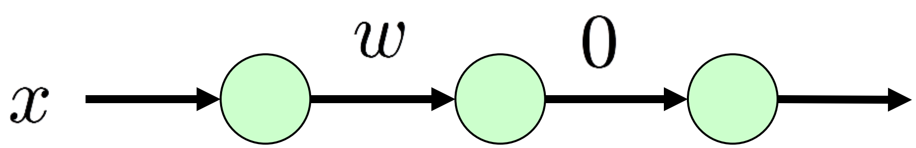
階層モデル学習の困難さ

- 局所的最適解が多数存在するため，
大域的最適解を求めるのが困難

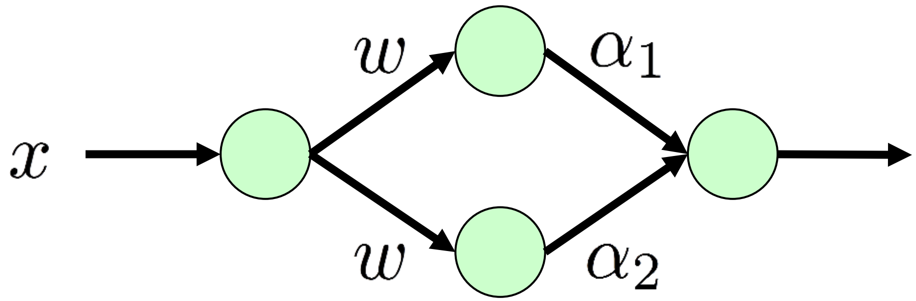


階層モデル学習の困難さ

- パラメータと関数が一対一に対応しないため，学習がより困難になる
  - 2層目の重みがゼロのとき，1層目の重みwを変えても関数は変わらない

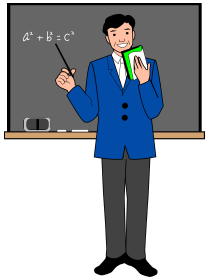


  - 1層目の重みが等しいとき，2層目の重みの和$\alpha_1+\alpha_2$が一定ならば，関数は同じ



学習モデル：まとめ

- 線形モデル：パラメータに関して線形なモデル
  - 加法モデル：パラメータ数は少ないが表現力が劣る
  - 乗法モデル：表現力は豊かだがパラメータが多すぎる
  - カーネルモデル：程よい表現力かつ程よいパラメータ数
- 非線形モデル：パラメータに関して非線形なモデル
  - 階層モデル：特異性のため学習が困難
- どのモデルが良いかは応用事例に依存するため，実際にはデータから適切なモデルを選ぶ必要がある（モデル選択）
- 本講義では主に線形（カーネル）モデルを扱う



講義の流れ

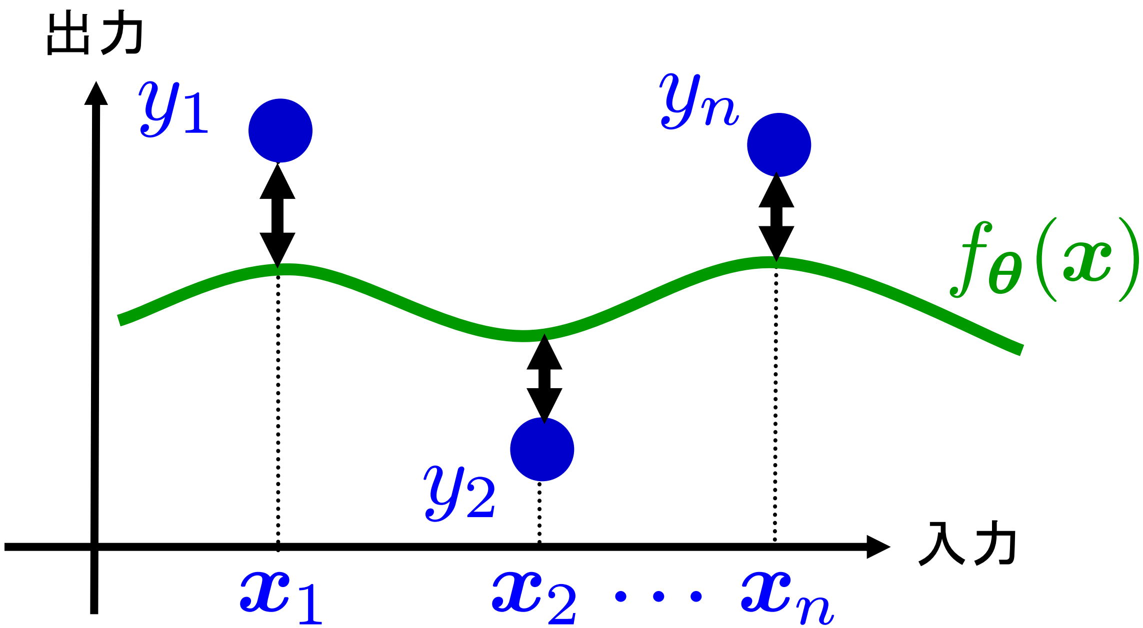
1. 学習モデル（2章）
2. 最小二乗回帰（3章）
3. 正則化回帰（4章）

最小二乗回帰

■ **規準**:訓練出力との二乗誤差を最小にする

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta})$$

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2$$



線形モデルに対する解の求め方

■ 線形モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$ に対する解:

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\right)^2$$

$$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

$$= \frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{y}\|^2 = \frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{y})^\top(\boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{y})$$

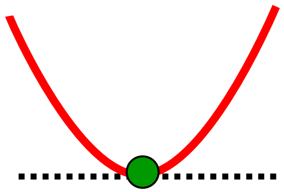
$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \cdots & \phi_b(\boldsymbol{x}_n) \end{pmatrix}$$:計画行列

線形モデルに対する解の求め方(続き)

- 偏微分がゼロの点の満たす式:

$$\nabla_{\boldsymbol{\theta}} J_{\mathrm{LS}} = \left(\frac{\partial J_{\mathrm{LS}}}{\partial \theta_1}, \ldots, \frac{\partial J_{\mathrm{LS}}}{\partial \theta_b} \right)^{\top} = \boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{\Phi}^{\top}\boldsymbol{y} = \mathbf{0}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^{\top} \boldsymbol{A} \boldsymbol{\theta} = 2\boldsymbol{A}\boldsymbol{\theta} \qquad \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{b}^{\top} \boldsymbol{\theta} = \boldsymbol{b}$$



- 解が満たす方程式: $\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\boldsymbol{\theta} = \boldsymbol{\Phi}^{\top}\boldsymbol{y}$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y}$$

注: $(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}$が存在すると仮定

- 最適解が解析的に求められる!

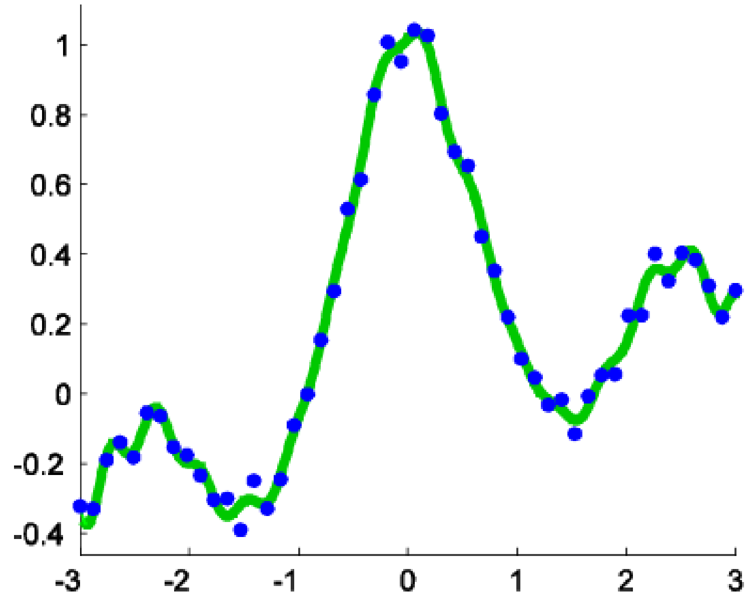
実行例

- 三角多項式モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$

$$\phi_i(\boldsymbol{x}) = 1, \sin x/2, \cos x/2, \ldots, \sin 15x/2, \cos 15x/2$$

に対する最小二乗回帰の実行例

(真の関数: $f(x) = \sin(\pi x)/(\pi x) + 0.1x$)



MATLABでの実装例

```
clear all; rand('state',0); randn('state',0);
n=50; N=1000;
x=linspace(-3,3,n)'; X=linspace(-3,3,N)';
pix=pi*x; y=sin(pix)./(pix)+0.1*x+0.05*randn(n,1);

p(:,1)=ones(n,1); P(:,1)=ones(N,1);
for j=1:15
  p(:,2*j)=sin(j/2*x); p(:,2*j+1)=cos(j/2*x);
  P(:,2*j)=sin(j/2*X); P(:,2*j+1)=cos(j/2*X);
end
t=(p'*p)\(p'*y); F=P*t;

figure(1); clf; hold on; axis([-2.8 2.8 -0.5 1.2]);
plot(X,F,'g-'); plot(x,y,'bo');
```

Pythonコードは添付資料参照(今後も同様)

数学演習

■ 訓練出力の雑音が正規分布に独立に従う時，最小二乗回帰はガウスモデルの最尤推定法と一致することを証明せよ

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \qquad y_i = f(\boldsymbol{x}_i) + \epsilon_i \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- 確率モデル: $$p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2}{2\sigma^2}\right)$$
- 対数尤度: $$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$$
- 最尤推定: $$\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2$$

解答例

- $$\log p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2}{2\sigma^2}$$

- $$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$$
$$= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2}{2\sigma^2}$$

- $$\operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta})$$

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2$$

最小二乗回帰の正当化

- 出力の雑音が正規分布に独立に従う時:
- 一致性:最小二乗解は最適な解に確率収束

⇒ 標本が無限にたくさんあれば,
最適なパラメータが求まる

- 漸近有効性:漸近正規推定量の中で漸近分散が最小

⇒ 標本が十分にたくさんあるとき,
推定結果は安定している

カーネルモデルに対する最小二乗回帰

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2h^2}\right)$$

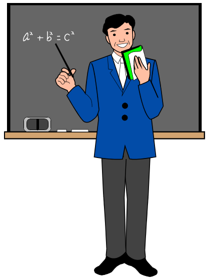
- 二乗誤差規準:

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{K}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \Longrightarrow \widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \boldsymbol{K}^{-1}\boldsymbol{y}$$

$$\boldsymbol{K} = \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}$$:カーネル行列

最小二乗回帰：まとめ

- 訓練出力との二乗誤差を最小にする
- 回帰法の最も基礎的な手法
- 出力の雑音が正規分布のとき，
最尤推定と解釈できる
- 線形モデル・カーネルモデルに対して，
解を解析的に求められる



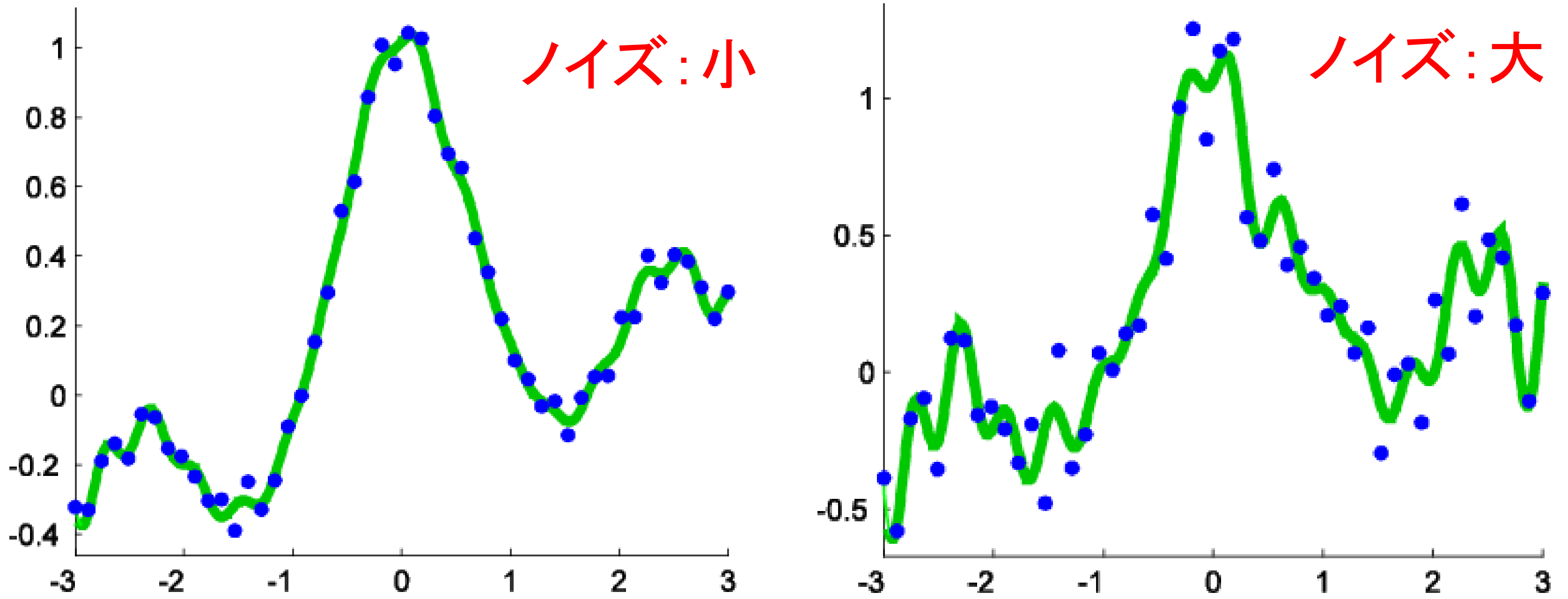
講義の流れ

1. 学習モデル（2章）
2. 最小二乗回帰（3章）
3. 正則化回帰（4章）
   - A) 制約付き最小二乗回帰
   - B) モデル選択

最小二乗回帰の問題点

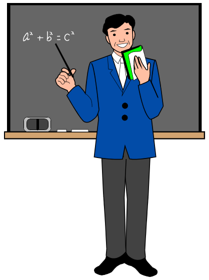
- 過適合:ノイズを含む訓練標本に適合し過ぎる

$$f(x) = \sin(\pi x)/(\pi x) + 0.1x$$



$$\phi_i(\boldsymbol{x}) = 1, \sin x/2, \cos x/2, \ldots, \sin 15x/2, \cos 15x/2$$

➡ モデルを適切に制限する必要がある



講義の流れ

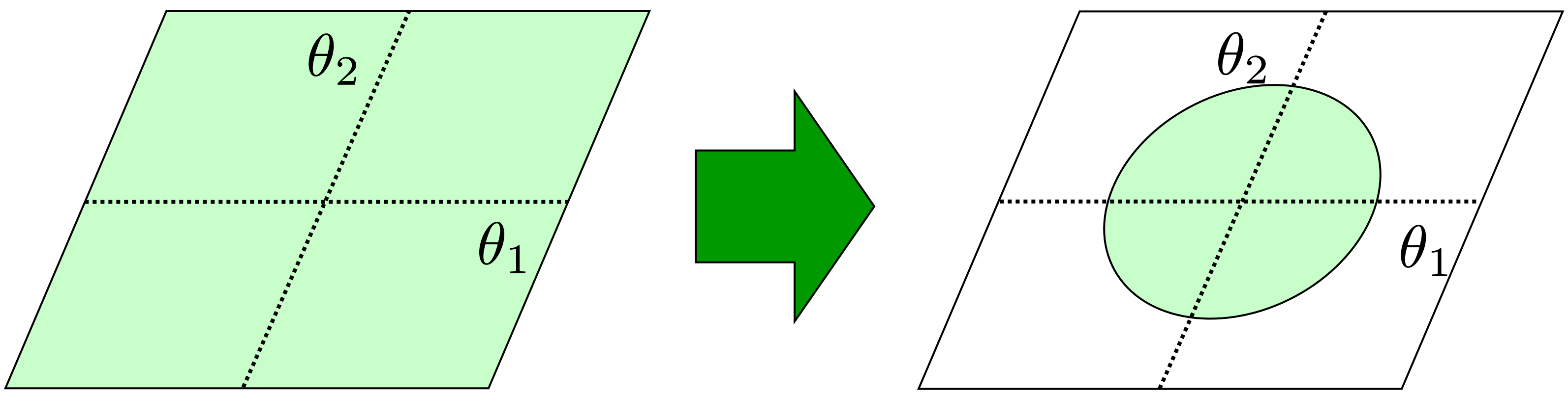
1. 学習モデル（2章）
2. 最小二乗回帰（3章）
3. 正則化回帰（4章）
   - A) 制約付き最小二乗回帰
   - B) モデル選択

ℓ_2-制約付き最小二乗回帰

- モデルを超球に限定することにより過適合を回避

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2 \quad \text{subject to} \ \|\boldsymbol{\theta}\|^2 \leq R$$

$$R \geq 0$$



通常の最小二乗回帰

ℓ_2-制約付き最小二乗回帰

解の求め方

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2 \quad \text{subject to } \|\boldsymbol{\theta}\|^2 \leq R$$

- 等価な表現:

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{2}\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2\right]$$

$\lambda\,(\geq 0)$: R から決まる定数

- 実際には R でなく λ を直接指定すればよい.

解釈

$$\min_{\boldsymbol{\theta}} \left[\underbrace{\frac{1}{2}\sum_{i=1}^{n}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2}_{\text{訓練出力に対する適合の良さ}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2}_{\text{パラメータの値が大きくなり過ぎることに対する罰則（正則化, regularization）}} \right]$$

- 訓練出力に対する適合のよさとパラメータの値の大きさをバランスよく小さくしている.
- ℓ_2-正則化回帰とも呼ばれる.

数学演習

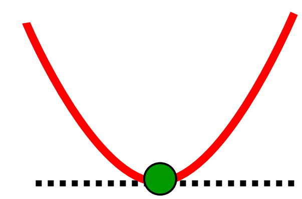
$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}} \left[\frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \right]$$

- 線形モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$ に対する解を求めよ
- ヒント: $\displaystyle \frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2 = \frac{1}{2} \|\boldsymbol{\Phi\theta} - \boldsymbol{y}\|^2$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \cdots & \phi_b(\boldsymbol{x}_n) \end{pmatrix} \qquad \boldsymbol{y} = (y_1, \ldots, y_n)^\top$$

解答例

- 偏微分をゼロとおく:



$$\nabla_{\boldsymbol{\theta}}\left(\frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{\theta}-\boldsymbol{y}\|^2+\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2\right)$$

$$=\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\theta}-\boldsymbol{\Phi}^\top\boldsymbol{y}+\lambda\boldsymbol{\theta}=\boldsymbol{0}$$

- 解が満たす方程式: $(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\lambda\boldsymbol{I})\boldsymbol{\theta}=\boldsymbol{\Phi}^\top\boldsymbol{y}$

$$\widehat{\boldsymbol{\theta}}=(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}$$

$\boldsymbol{I}$:単位行列

- 最小二乗回帰と同様に，最適解を解析的に求めることができる！

実行例

- (ガウス)カーネルモデル

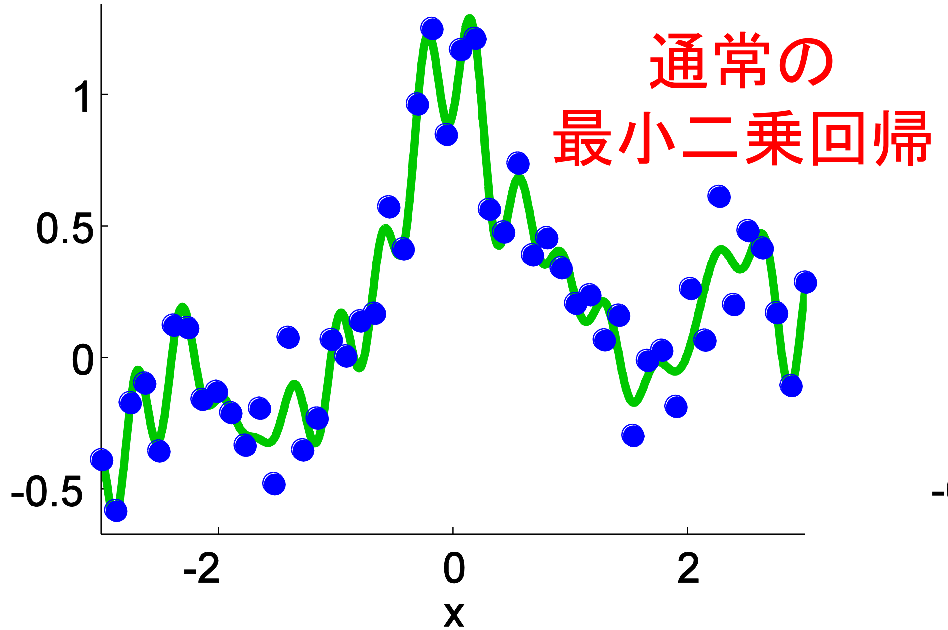
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

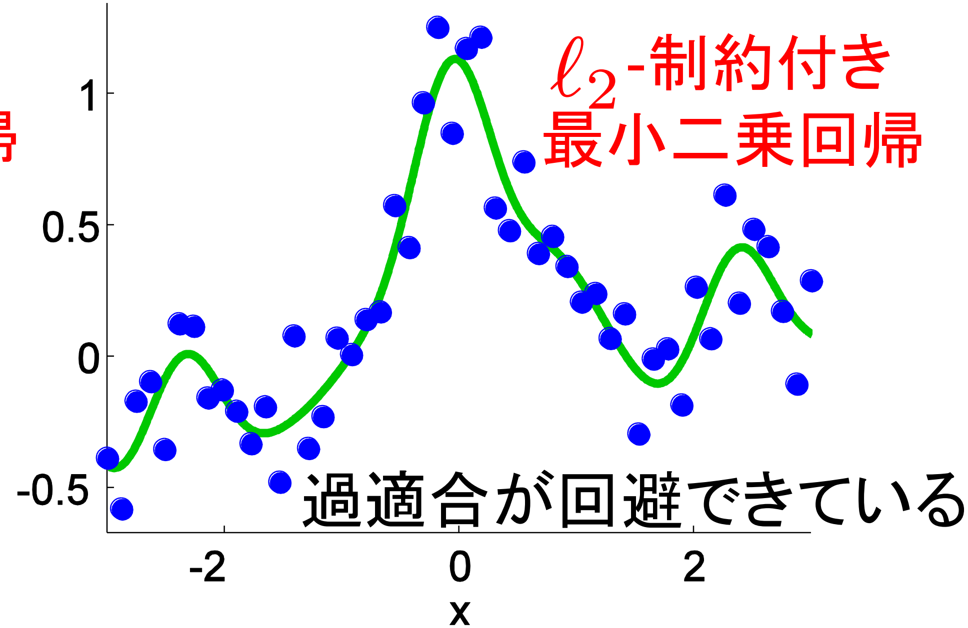
$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{2h^2}\right)$$

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{K}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{K}^\top \boldsymbol{y}$$

$$\boldsymbol{K} = \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}$$

$$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$$



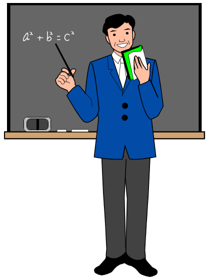


MATLABでの実装例


```
clear all; rand('state',0); randn('state',0);
n=50; N=1000;
x=linspace(-3,3,n)'; X=linspace(-3,3,N)';
pix=pi*x; y=sin(pix)./(pix)+0.1*x+0.2*randn(n,1);

x2=x.^2; X2=X.^2; hh=2*0.3^2; l=0.1;
k=exp(-(repmat(x2,1,n)+repmat(x2',n,1)-2*x*x')/hh);
K=exp(-(repmat(X2,1,n)+repmat(x2',N,1)-2*X*x')/hh);
t=(k^2+l*eye(n))\(k*y); F=K*t;

figure(1); clf; hold on; axis([-2.8 2.8 -1 1.5]);
plot(X,F,'g-'); plot(x,y,'bo');
```

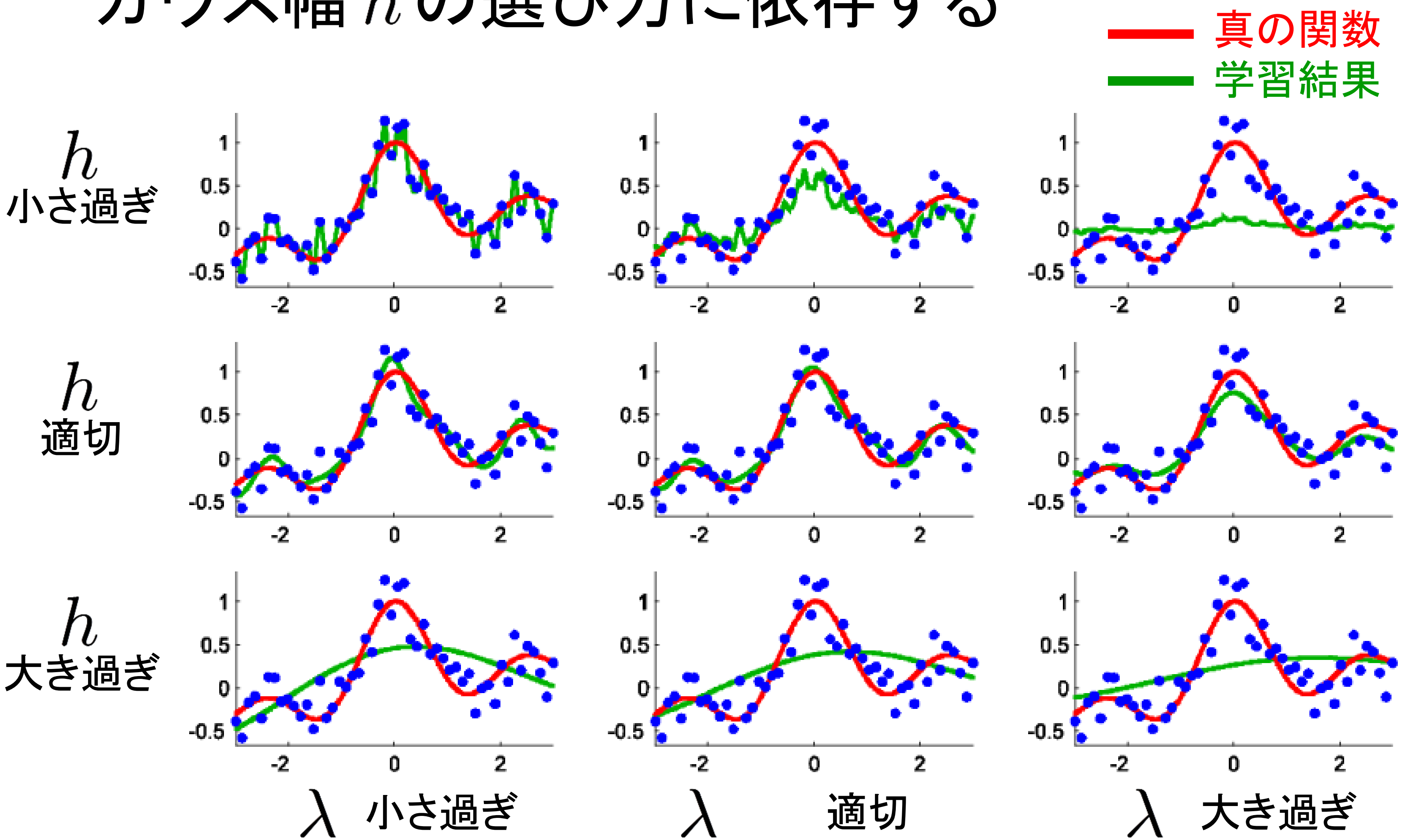


講義の流れ

1. 学習モデル（2章）
2. 最小二乗回帰（3章）
3. 正則化回帰（4章）
   A) 制約付き最小二乗回帰
   B) モデル選択

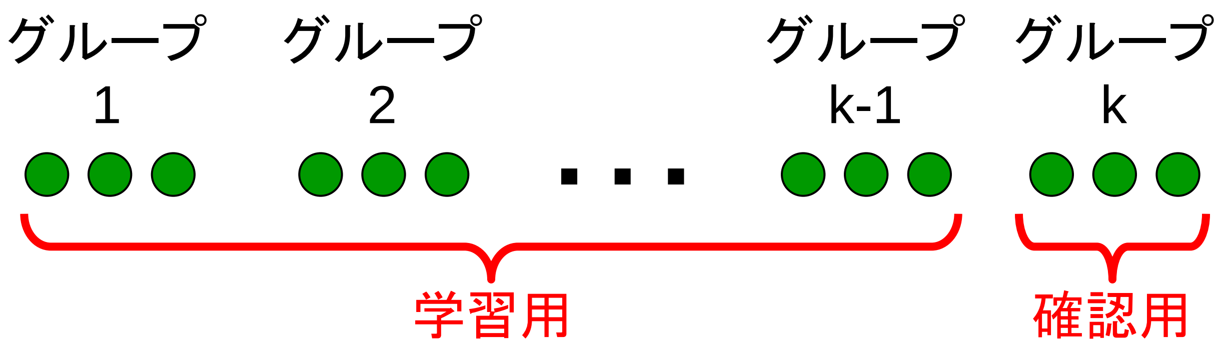
正則化回帰のモデル選択

- 正則化回帰の結果は，正則化パラメータλと ガウス幅hの選び方に依存する



交差確認法 (Cross-validation)

- 訓練標本 $\mathcal{Z}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$ をk分割する: $\{\mathcal{Z}_i\}_{i=1}^{k}$
- $\mathcal{Z}_i$ 以外を使って(固定したλ,hに対し) θ を学習
- 残った $\mathcal{Z}_i$ を使ってテスト誤差を確認する
- これを全ての組み合わせに対して繰り返し，平均を出力する

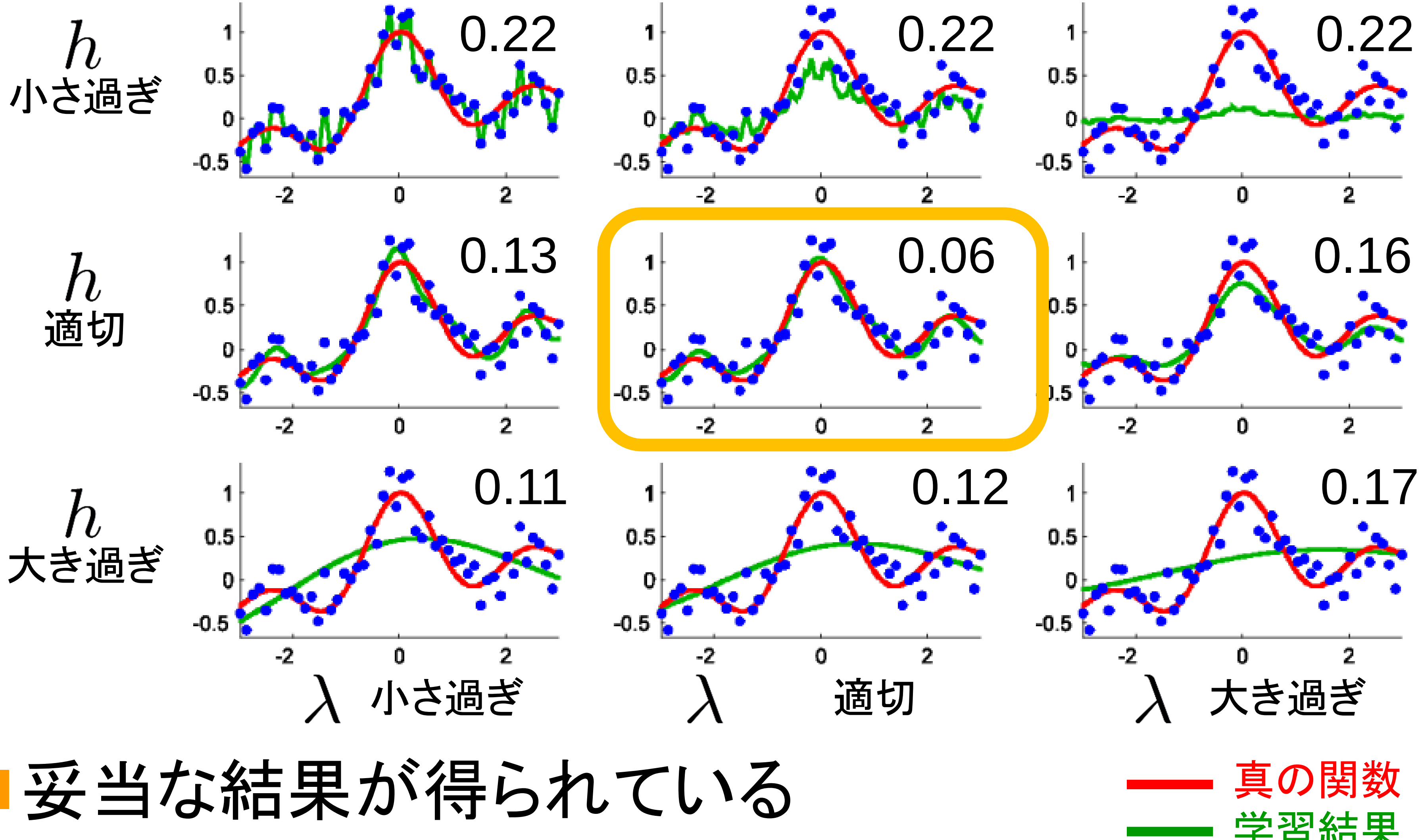


$\widehat{f}(\boldsymbol{x})$

$\left(\widehat{f}(\boldsymbol{x}')-y'\right)^2$

実行例

- ガウスカーネルモデル： $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_j\|^2}{2h^2} \right)$



- 妥当な結果が得られている

一つ抜き交差確認

- グループ数 k を標本数 n に設定する

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_i(\boldsymbol{x}_i)-y_i\right)^2$$

$\widehat{f}_i(\boldsymbol{x})$： $(\boldsymbol{x}_i, y_i)$ 以外の標本から学習した関数

  - 毎回一つだけ標本を抜く

- 一般に計算に非常に時間がかかるため，実用的でないが，線形モデルを用いた ℓ_2 -正則化回帰に対しては解析的に計算できる：

証明は宿題

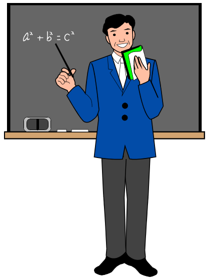
$$\frac{1}{n}\|\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{H}\boldsymbol{y}\|^2$$

$$\boldsymbol{H}=\boldsymbol{I}-\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^\top$$

$\widetilde{\boldsymbol{H}}$： $\boldsymbol{H}$ と同じ対角成分を持ち，非対角成分は零

正則化回帰：まとめ

- 最小二乗回帰は雑音に過適合しやすい
- パラメータの探索範囲をℓ_2-ノルムを用いて制約する
- 解は解析的に求められる
- モデル選択が重要

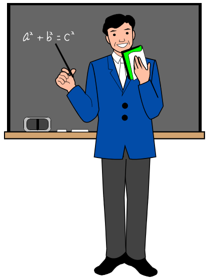


講義の流れ

1. 学習モデル（2章）
2. 最小二乗回帰（3章）
3. 正則化回帰（4章）

まとめ

- 関数を学習するためのモデル:
  - 線形モデル，カーネルモデル，非線形モデル
- 最小二乗回帰:
  - 訓練標本との二乗誤差を最小化
  - 解は解析的に計算できる
- ℓ_2 -正則化回帰:
  - 最小二乗回帰の過適合を軽減
  - 解は解析的に計算できる
  - モデル選択には交差確認法を用いる



次回の予告

- スパース回帰（5章）

宿題1

- ガウスカーネルモデル

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2h^2} \right)$$

に対する ℓ_2 -正則化を用いた最小二乗回帰の交差確認法を実装し，正則化パラメータとガウス幅を決定せよ

宿題2

■線形モデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$ を用いた ℓ_2 -正則化回帰に対する一つ抜き交差確認による二乗誤差は，次式で解析的に計算できることを示せ：

$$\frac{1}{n} \|\widetilde{\boldsymbol{H}}^{-1} \boldsymbol{H} \boldsymbol{y}\|^2$$

$$\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}^\top$$

$\widetilde{\boldsymbol{H}}$： $\boldsymbol{H}$ と同じ対角成分を持ち，非対角成分は零

宿題2(続き)

■ヒント1:

- $(\boldsymbol{x}_i, y_i)$ を抜いて学習したパラメータ $\widehat{\boldsymbol{\theta}}_i$ を

$$\boldsymbol{\phi}_i = (\phi_1(\boldsymbol{x}_i), \phi_2(\boldsymbol{x}_i), \ldots, \phi_b(\boldsymbol{x}_i))^\top$$

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n)^\top,$$

$$\boldsymbol{y},\ y_i,\ \boldsymbol{U} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I}$$

  を用いて表す

- 以下の逆行列の公式を用いる(Sherman-Morrison-Woodbury公式の特別な形):

$$(\boldsymbol{U} - \boldsymbol{u}\boldsymbol{u}^\top)^{-1} = \boldsymbol{U}^{-1} + \frac{\boldsymbol{U}^{-1}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{U}^{-1}}{1 - \boldsymbol{u}^\top\boldsymbol{U}^{-1}\boldsymbol{u}}$$

宿題2（続き）

- ヒント2：
  - $(\boldsymbol{x}_i, y_i)$ を抜いて学習したパラメータからの y_i の予測値 $\phi_i^\top \widehat{\boldsymbol{\theta}}_i$ を

$$\boldsymbol{\phi}_i = (\phi_1(\boldsymbol{x}_i), \phi_2(\boldsymbol{x}_i), \ldots, \phi_b(\boldsymbol{x}_i))^\top$$

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n)^\top,$$

$$\boldsymbol{y},\ y_i,\ \boldsymbol{U} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I}$$

を用いて表し，逆行列公式で現れる分母と通分を繰り返す