PRML輪講

３章：線形回帰モデル

担当：大木

章の構成

3.1 線形基底関数モデル

3.2 バイアス・バリアンス分解

3.3 ベイズ線形回帰

3.4 ベイズモデル比較

3.5 エビデンス近似

この章でやること

- 線形回帰を理解する(前半)
- ベイズ+線形回帰を理解する(後半)

線形回帰(前半)

3.1 線形基底関数モデル

3.2 バイアス・バリアンス分解

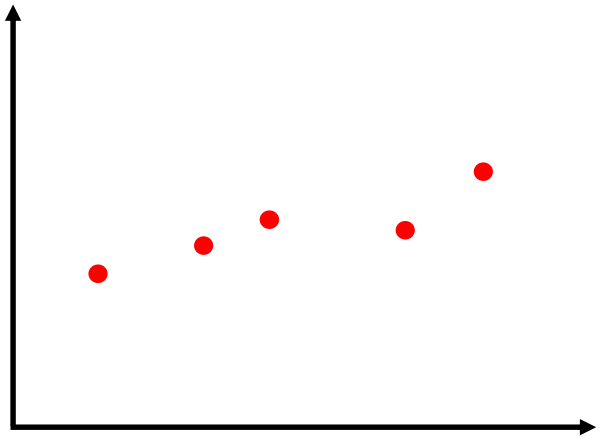
3.3 ベイズ線形回帰

3.4 ベイズモデル比較

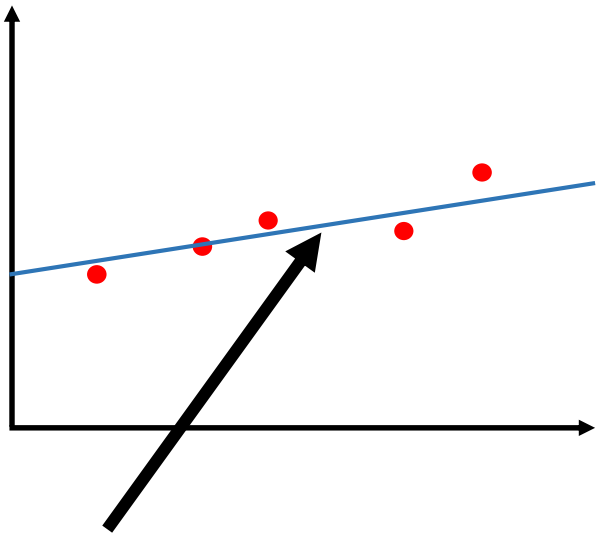
3.5 エビデンス近似

線形回帰とは

データを取得



データの従う関数をプロット



青線の傾き・切片を求めたい

回帰問題→入力(説明)変数xから
目的変数tを予測する

線形回帰→基底関数とパラメータの
線形結合でもって目的変数を表す

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$$

基底関数は
非線形OK

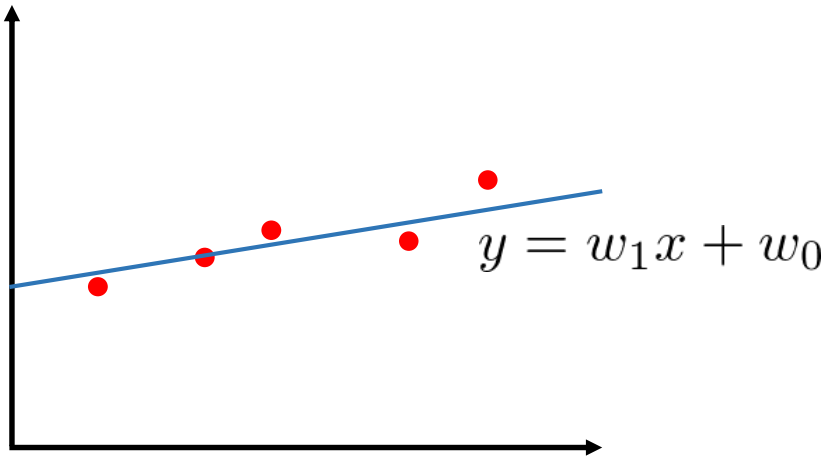
Point

- 回帰関数がパラメータに対する線形性を持つ

線形回帰の例

基底関数：多項式関数

$$\phi(\mathbf{x}) = (1\ x\ x^2 \cdots x^{M-1})^{\mathrm{T}}$$

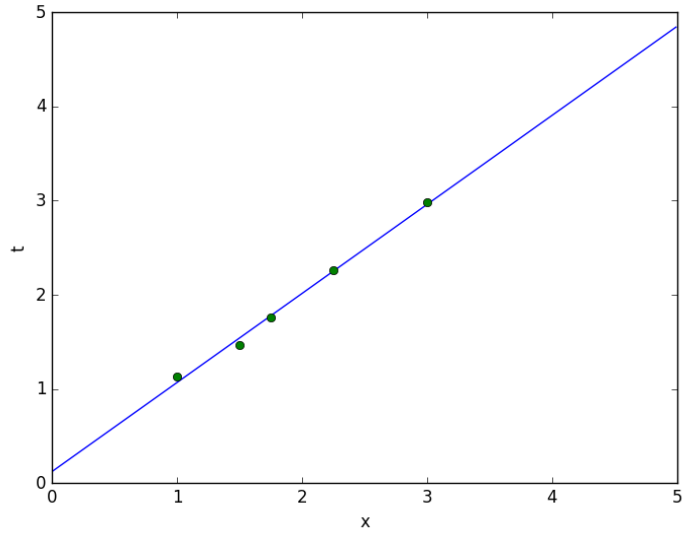


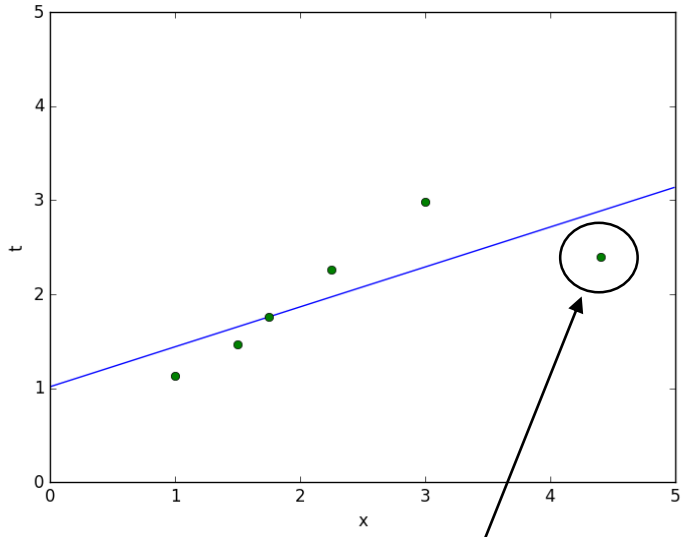
データから最小二乗法などでパラメータを決定

二乗和誤差を最小化

その他シグモイド関数(ロジット関数の逆関数)(3.6)やフーリエ関数を用いた基底がある

最小二乗法の注意点





外れ値検出

外れ値に敏感

・外れ値を除いて統計処理を行う
・最小二乗法よりも外れ値に対して
ロバストな手法を用いる

RANSAC
絶対誤差

目的変数が１次元の場合(多次元の場合も同様)

データセット $(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), ..., (\mathbf{x}_N, t_N)$

測定に際し、ガウスノイズϵが混入

$$t_i = y(\mathbf{x}_i, \mathbf{w}) + \epsilon_i \ \ (i = 1, 2, ..., N)$$

→*i*番目のデータに対する尤度関数

βは精度(分散の逆数)
尤度関数はwの関数

$$p(t_i|\mathbf{x}_i, \mathbf{w}, \beta) = \mathcal{N}(t_i|y(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$$

以上から、データ集合に対する尤度関数が得られる

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1})$$

(復習)MLEは、尤度関数が最大となるパラメータを推定量とする

計算の簡略化のため、
対数尤度関数を最大化することを考える

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\beta E_D(\mathbf{w}) + const$$

ただし $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2$

→(復習)ガウスノイズの下で、MLE=LSM

実際に最尤推定量を計算してみる(計算ノート)

最小二乗法の幾何学的な解釈を考える(N>M)

ベクトルデータt:座標 $\xi = (\xi_1, \ldots, \xi_N)^{\mathrm{T}}$ の張る
N次元空間の点

※ $\Phi^{\mathrm{T}}\Phi$ が非正則に近いとき
→SVDによって擬似逆行列を表現し、特異性を解消

逐次学習(オンライン学習)

バッチ学習：MLEのように訓練データを すべて使って学習を行う

最急降下法(反復法によるアルゴリズム)

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \sum_{n=1}^{N} \nabla E_n(\mathbf{w}^{(\tau)})$$

学習率(正の小さな数)

バッチ学習は大規模なデータに対しては
計算時間が膨大になる
→パラメータ更新の際にデータの一つだけを
用いて学習を行うオンライン学習が効率的

確率的勾配降下法(SGD)の更新則

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

・最小二乗法では過学習がしばしば起こる(図1.9参照)
・多くの場合データが従うモデルが全く分からない

→目的関数に正則化項を付加

L_qノルム正則化 $\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$

→不必要なパラメータは減衰して0になる

$q=2$の正則化項を機械学習の分野では
荷重減衰と呼び、最もよく使用する

※$q=1$のときをlassoといい、適切なλの範囲で
スパースなパラメータ推定が可能(→計算ノート)

3.1 線形基底関数モデル

3.2 バイアス・バリアンス分解

3.3 ベイズ線形回帰

3.4 ベイズモデル比較

3.5 エビデンス近似

バイアスーバリアンス分解

MLEの過学習の問題を解消するために…

- 基底の数を制限→モデルの表現能力が低下
- 正則化項の追加→係数λをどう決めるか

誤差の最小化と適切なモデル選択を両立させたい

1.5.5節の議論から、期待(二乗)損失

$$\mathrm{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int \int \{h(\mathbf{x}) - t)\}^2 p(\mathbf{x}, t) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}$$

学習ではこの項を動かして期待損失を最小化する

※データ集合が有限であることから、
第一項は厳密に0にすることはできない

(3.37)式のデータ集合の取り方に
対する期待値をとれば

期待損失＝(バイアス)^2+バリアンス+ノイズ

バイアス：モデルの自由度が不十分であるために
真の回帰関数から生じてしまうずれ

バリアンス：ランダムなサンプルに基づく
推定量の確率的なゆらぎ

期待損失の最小化
＝(二乗バイアス＋バリアンス)の最小化

モデルの複雑度高：バイアス小,バリアンス大

モデルの複雑度低：バイアス大,バリアンス小

→ちょうどよいモデルを選ぶ必要

正則化項の利用

正則化係数大＝モデルの複雑度低
正則化係数小＝モデルの複雑度高

図3.5,図3.6参照

ベイズ線形回帰(後半)

3.1 線形基底関数モデル

3.2 バイアス・バリアンス分解

3.3 ベイズ線形回帰

3.4 ベイズモデル比較

3.5 エビデンス近似

過学習を避けるために、
データを訓練データ、テストデータに分ける

→データがもったいない

ベイズの枠組みで線形回帰を考える

1.2.5,1.2.6節の議論を再考(復習)する

MAP推定を行うことを考える
→パラメータの事前分布を(3.48)と与える

事後分布は(3.49)式のように書ける
(計算ノート)

Point
- 事前分布がガウス分布であるため、
  事後分布もガウス分布となる

→計算が楽(事後分布の更新が楽)

MAP推定について

対数事後確率の最大化(MAP推定)
＝二乗和誤差＋二次の正則化項の最小化(復習)

パラメータ空間全域を一様に分布しているような分布を事前分布とした場合
→MAP推定量＝最尤推定量

事後確率の逐次学習におけるふるまいを確認
(図2.3,図3.7参照)

確率の乗法・加法定理より、予測分布

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w}$$

$$\left.\begin{array}{l} p(t|\mathbf{w},\beta) \\ p(\mathbf{w}|\mathbf{t},\alpha,\beta) \end{array}\right\}$$ (3.8),(3.49)参照

(2.115)から、予測分布は
以下のように書ける(計算ノート)

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

(3.59)で定義

等価カーネル

線形基底モデルに対する事後分布の平均解(3.53)は、ガウス過程(cf. 6章)を含むカーネル法を導入する下で異なった解釈を与える

(3.53)に(3.3)代入

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} \beta \phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x}_n) t_n \quad \textbf{(3.60)}$$

等価カーネルを次のように定義

$$k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}') \quad (3.62)$$

この式を用いれば、(3.60)は

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}')t_n \quad (3.61)$$

と書ける

※このように、訓練データの目標変数の線形結合で与えられる回帰関数を線形平滑器という

図(3.10)からカーネルはxに近いデータ点ほど大きい

→カーネルによって、遠くの情報より近傍の情報を強く重みづけ

(3.63)より、近傍での予測平均は強い相関を持ち、離れた点では相関は小さくなる

※この性質はガウス基底に限らず、局所性を持たない基底についても成り立つ

以上より、カーネルを用いることで
これまでの線形回帰問題を異なった
形式で定式化できる

・基底関数の集合をあらかじめ定義しない
・カーネルを定義
→データが与えられた時に
カーネルを用いて予測値を計算

3.1 線形基底関数モデル

3.2 バイアス・バリアンス分解

3.3 ベイズ線形回帰

3.4 ベイズモデル比較

3.5 エビデンス近似

MLEの過学習の問題→パラメータについて
周辺化で回避可能

テストデータいらず

ベイズ線形回帰の強み

ベイズにおけるモデル選択とは？

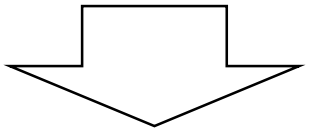
→ベイズモデル比較

問題設定

L個のモデル候補の中からモデルの事後分布

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \quad \textbf{(3.66)}$$

を評価し、モデルを１つ決める(モデル選択)



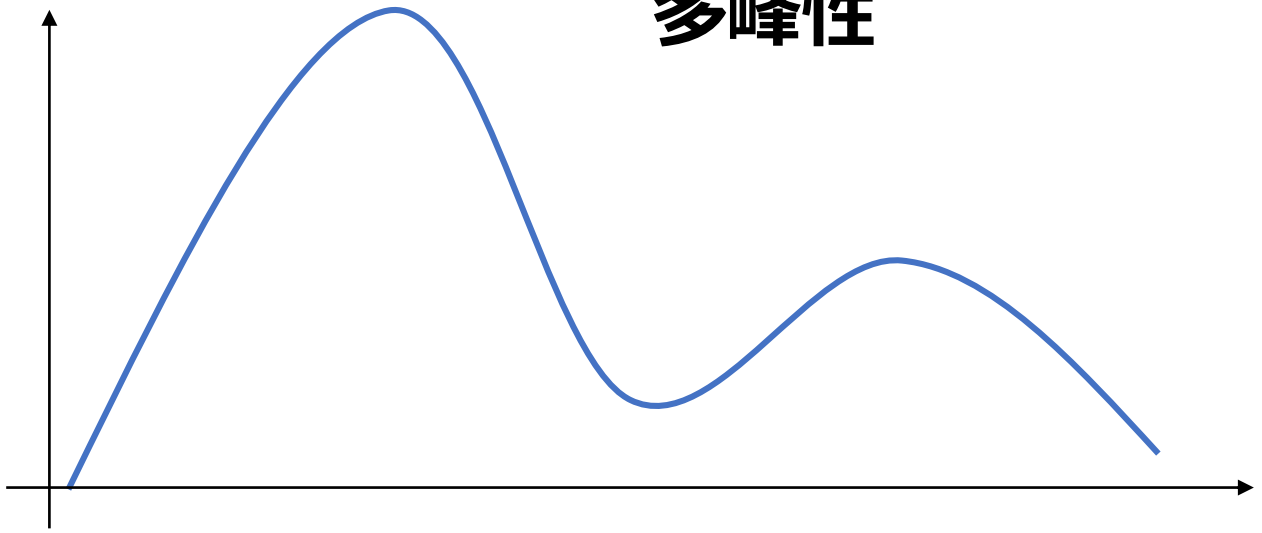
モデルエビデンス項 $p(\mathcal{D}|\mathcal{M}_i)$ を比較

(事前分布はすべてのモデルで等しいとする)

モデルエビデンス ＝ データから見た
モデルの好み

モデルの事後分布がわかれば、
予測分布が得られる(3.67)
このときの予測分布は混合分布となっている

多峰性



モデルエビデンス

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

※表記を簡便化するために
モデル依存性を省略

を具体的に計算することを考える

仮定

- **事後分布がモード近傍で鋭くとがっている**
- **事前分布が平坦である**

→積分を長方形型の分布で近似(図3.12参照)

パラメータが1つしかないとして、
エビデンスは以下のように近似できる

$$p(\mathcal{D}) \simeq p(\mathcal{D}|w_{\mathrm{MAP}}) \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$$

対数をとって

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right)$$

MAP推定値における
データへの当てはまり度

モデルが複雑になることに
対するペナルティ

同様に、パラメータがM個あるとき
すべてのパラメータに対して第二項が
一定であるとして以下が得られる

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + M \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right)$$

Mが大きくなる(モデルが複雑になる)
→第一項↑、第二項↓

モデルエビデンスの最大化によって、
ちょうどよい表現力のモデルを選べる

ベイズモデル比較における注意

・考えているモデルの中に真の分布があることを仮定

→そうでない場合、誤った結果が得られることも

・用いる事前分布によってはうまくいかない

応用上テストデータは用意すべき

3.1 線形基底関数モデル

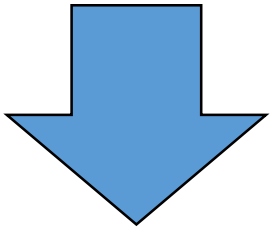
3.2 バイアス・バリアンス分解

3.3 ベイズ線形回帰

3.4 ベイズモデル比較

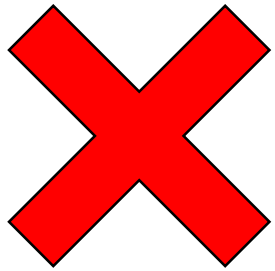
3.5 エビデンス近似

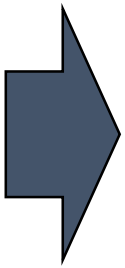
これまで、パラメータwに対する
周辺化を行ってきた



超パラメータα, βに対しても
周辺化予測を行う

すべてのパラメータに対して
解析的に周辺化





かわりに、パラメータwのみに
関して周辺化した周辺尤度を
超パラメータを動かして最大化

エビデンス近似

超パラメータに対して事前分布を導入

予測分布

$$p(t|\mathbf{t}) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})d\mathbf{w}d\alpha d\beta$$

事後分布が $\hat{\alpha},\hat{\beta}$ の周りでデルタ
関数的に尖っているとすると

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int p(t|\mathbf{w},\hat{\beta})p(\mathbf{w}|\mathbf{t},\hat{\alpha},\hat{\beta})d\mathbf{w}$$

と書ける

ここで、ベイズの定理から

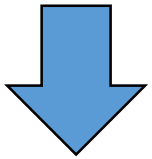
$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

事前分布がフラットなら、
MAP=MLEであったことから
事後分布を最大化する$\hat{\alpha}, \hat{\beta}$は
周辺尤度を最大化して得られる
→周辺尤度を最大化

モデルエビデンス

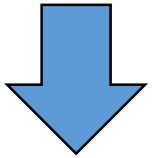
ここまでの流れ

(3.57)の予測分布でα, βに対しても
周辺化したい



解析計算きついので近似

事後分布 $p(\alpha, \beta|\mathbf{t})$がデルタ関数的であるなら、
そのピークでのα, βで予測分布は近似できる



そのα, βの値を求めよう

ここでも、対数をとって
対数エビデンスを最大化する

周辺尤度関数

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

これの停留点での超パラメータの値を求める

→計算ノート

注：求まったパラメータは
(3.92),(3.95)のように与えられる
これは、セルフコンシステントな解

エビデンス近似によって求めた解
→モデルの複雑さを最適化するために
テストデータを用いて検証する必要がない

これらの解についての解釈を考える

まずαについて

$\lambda_i \propto$ 尤度関数の歪み具合(曲率)
→図3.15より$\lambda_1 < \lambda_2$

$\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ が正定値行列→ λ_i は正

よって　　$0 < \frac{\lambda_i}{\lambda_i+\alpha} < 1$

$\lambda_i \ll \alpha$　パラメータは最尤推定値に近づく

$\lambda_i \gg \alpha$　パラメータは0に近づく

続いて、βのベイズ推定量(3.95)についても考察

分散の最尤推定量

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \qquad \textbf{(3.96)}$$

分散の不偏推定量

$$\sigma^2_{\mathrm{MAP}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \qquad \textbf{(3.97)}$$

最尤推定では、分散を過小評価している
→バイアスを取り除くために自由度の１つを用いていると解釈

一方、分散のベイズ推定量は

$$\frac{1}{\beta} = \frac{1}{N-\gamma} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2$$

→有効パラメータ数γがバイアスを補正

固定された基底関数の限界

固定された非線形関数を線形結合したモデル
(3章で扱った)

利点
- 閉じた解が求まる
- 任意の非線形返変換を表現可能

欠点
- 入力データの次元数に対して、
  指数的に基底関数の数を増やす
  必要あり(次元の呪い)

※現実のデータの本質的な次元数は
概してあまり大きくない