

ベイズ統計から見た最小二乗法

岡山大学 異分野基礎科学研究所

大槻純也



ガウス分布 (Gaussian distribution)

ガウス分布 または 正規分布 (Normal distribution)

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

誤差を表す際に用いられる。数学的に扱いやすい。

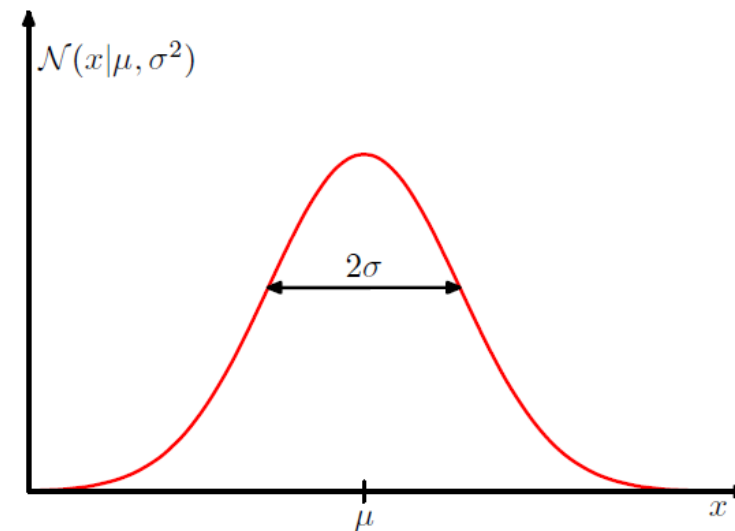
期待値 (Expectation value) 確率分布による平均値

$$\mathbb{E}[x] \equiv \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu \quad \text{mean (平均)}$$

$$\mathbb{E}[x^2] \equiv \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

$$\text{var}[x^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \quad \text{variance (分散)}$$

σ は standard deviation (標準偏差)



PRML Fig. 1.13

その他の性質

- Gaussian 2つの畳み込み積分はGaussian
- 多次元にも拡張可

尤度関数 (likelihood function)

線形回帰の前に、まずは、1変数 x を観測する問題を考える。

ガウス分布に従ってデータが生成されるとき、
値 x が観測される確率は

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$$

x を観測したとする。そのデータの生成過程、
すなわち (μ, σ^2) は知らないとする。

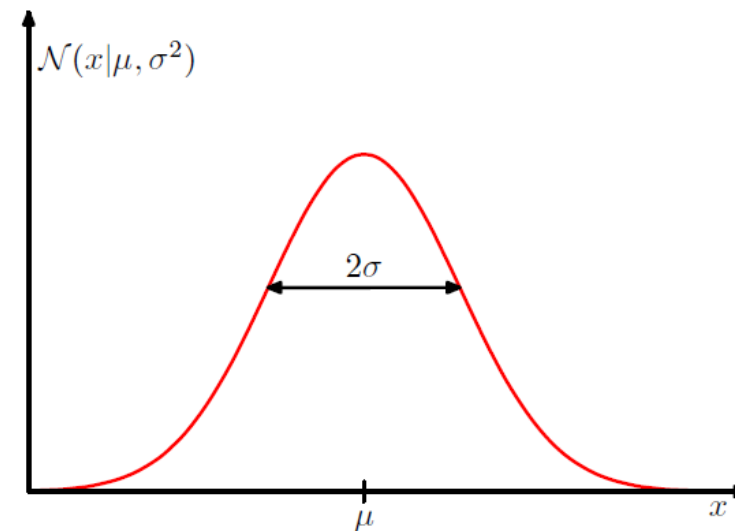
$p(x|\mu, \sigma^2)$ を (μ, σ^2) の関数と見る。このとき、

$p(x|\mu, \sigma^2)$ を **尤度関数 (likelihood function)** と呼ぶ。

測定を N 回行った場合

$$\mathbf{x} = \{x_1, x_2, \dots, x_N\}$$

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$



PRML Fig. 1.13

ベイズの定理

$$p(\mu, \sigma^2|\mathbf{x}) \propto p(\mathbf{x}|\mu, \sigma^2)p(\mu, \sigma^2)$$

知りたい量： (μ, σ^2)

観測データ： x

最尤推定 (Maximum likelihood)

尤度関数が最大になるように (μ, σ^2) を決定する

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x}|\mu, \sigma^2) = 0 \quad \text{より}$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad \text{sample mean}$$

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x}|\mu_{\text{ML}}, \sigma^2) = 0 \quad \text{より}$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 \quad \text{sample variance}$$

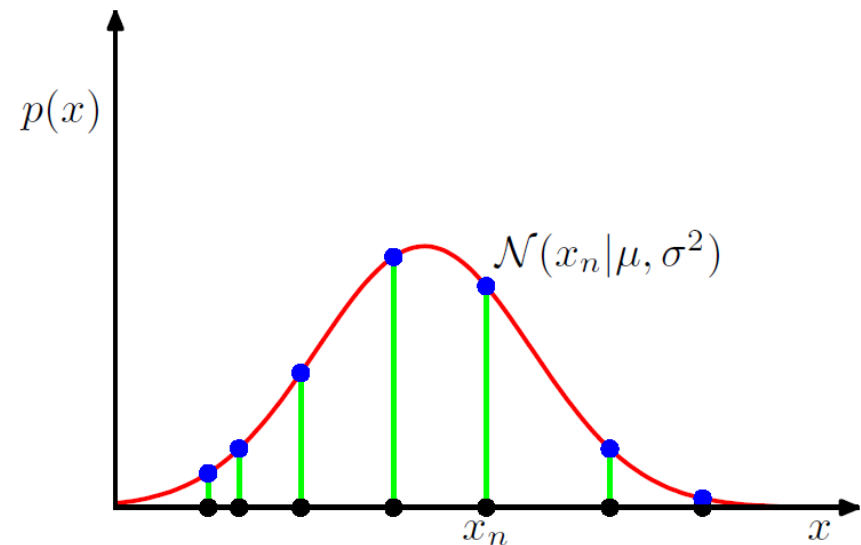
ひとつだけ注意：MLは分散を過小評価する

$$\mathbb{E}[\mu_{\text{ML}}] = \mu \quad \mathbb{E}[\sigma_{\text{ML}}^2] = \frac{N-1}{N} \sigma^2$$

導出はPRML Exercise 1.12

対数尤度 (log likelihood)

logは単調増加関数なので、対数尤度の微分でも解は同じ。logを取った方が計算が楽。



PRML Fig. 1.14

線形回帰

線形回帰に戻る

厳密な関数 $y(x, \mathbf{w})$ に誤差が乗ったデータが得られるとする

$$t = y(x, \mathbf{w}) + \epsilon \quad \text{誤差}$$

$$p(\epsilon) = \mathcal{N}(\epsilon|0, \beta^{-1}) \quad \text{誤差の確率分布}$$

誤差が確率分布するので t も確率分布する

$$\begin{aligned} p(t|x, \mathbf{w}, \beta) &= \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \\ &= \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t - y(x, \mathbf{w}))^2 \right] \end{aligned}$$

$(\mathbf{w}, \beta)$ の関数とみる → 尤度関数

$\beta = 1/\sigma^2$ は精度 (precision) パラメータと呼ばれる
 σ^2 よりも β の方が都合が良いことが多い

N 個のデータ点を得られたとする

$$\mathbf{x} = \{x_1, x_2, \dots, x_N\}$$

$$\mathbf{t} = \{t_1, t_2, \dots, t_N\}$$

線形回帰の尤度関数

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) &= \prod_{n=1}^N \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \\ &= \left(\frac{\beta}{2\pi} \right)^{N/2} \exp [-\beta E(\mathbf{w})] \end{aligned}$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2$$

統計力学のカノニカル分布と同じ形
統計力学の計算手法 (モンテカルロ法など)
が応用できそうなことが想像できる

線形回帰の最尤推定

対数尤度 (log likelihood) を最大化

$$\ln p(\mathbf{t}|\mathbf{w}, \beta) = -\beta E(\mathbf{w}) + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

↑ 条件 $\mathbf{x}$ は省略

$$\frac{\partial}{\partial \mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = 0 \quad \text{より}$$

$$\begin{aligned} \mathbf{w}_{\text{ML}} &= \arg \min_{\mathbf{w}} E(\mathbf{w}) \\ &= (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t} \end{aligned} \quad \text{最小二乗法と一致}$$

$$\frac{\partial}{\partial \beta} \ln p(\mathbf{t}|\mathbf{w}_{\text{ML}}, \beta) = 0 \quad \text{より}$$

$$\frac{1}{\beta_{\text{ML}}} = \frac{2}{N} E(\mathbf{w}_{\text{ML}}) \quad \text{平均二乗誤差}$$

t の予測分布 (Predictive distribution)

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$

↙ 新しい測定点

初めの仮定通りガウス分布

分散を過小評価する傾向がある点だけ注意

結局

最尤推定 = 最小二乗法

MAP推定

ベイズの定理

事後確率分布 尤度関数 事前確率分布

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

↑ 知りたいのは $\mathbf{w}$

Maximum posterior (MAP) 推定
(最大事後確率推定)

$$\mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{t}, \beta)$$

Maximum likelihood

$$\mathbf{w}_{\text{ML}} = \arg \max_{\mathbf{w}} p(\mathbf{t}, \beta|\mathbf{w})$$

事前確率として次の関数を考える

$$\begin{aligned} p(\mathbf{w}|\alpha) &= \prod_j \mathcal{N}(w_j|0, \alpha^{-1}) \\ &= \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \end{aligned}$$

すると、事後確率分布は

$$\ln p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \propto -\frac{\beta}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

この最大化は正則化付き最小二乗法と等価
正則化パラメータ $\lambda = \alpha/\beta$

まとめ

ベイズ統計の言葉を使うと

最小二乗法 → 最尤推定
正則化付き最小二乗法 → MAP推定
正則化 → 事前分布

パラメータ w の分布を無視する近似

物理の言葉を使うと

- ・パラメータ w のゆらぎを無視する近似
- ・平均場近似、鞍点近似

ベイズの定理

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

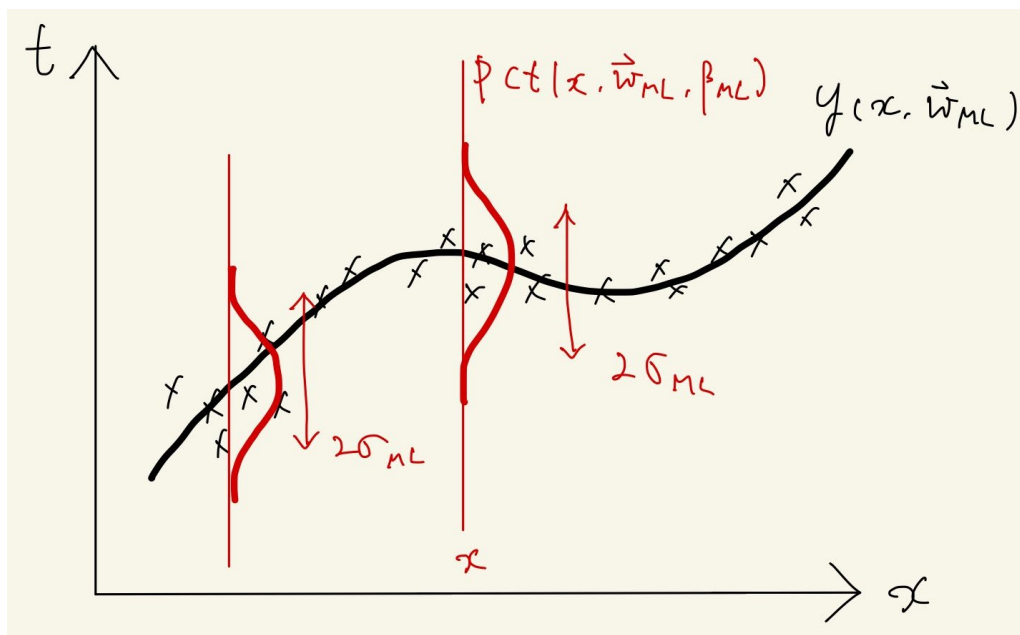
事後確率分布

実際は w は分布している

次回の予告

最尤推定の予測分布 (Predictive distribution)

$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$

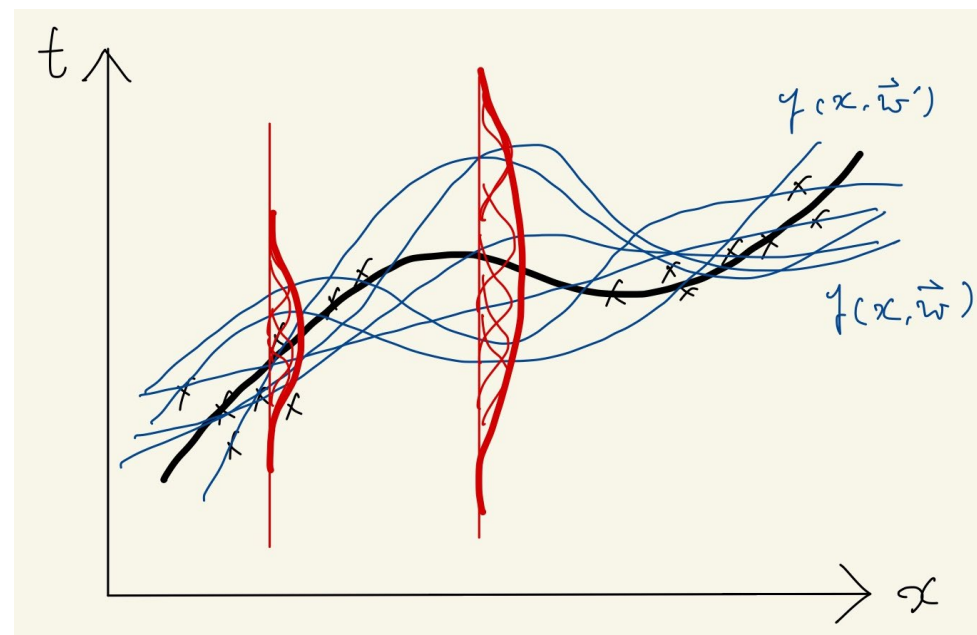


σ は x に依らない

$$\sigma = \sigma_{\text{ML}} = \beta_{\text{ML}}^{-1/2}$$

ベイズ推定の予測分布 (Predictive distribution)

パラメータ $\mathbf{w}$ の分布を考慮に入れる



σ が x に依存する

$$\sigma = \sigma(x)$$