

Data Science and Machine Learning

照屋 佑喜仁

July 3, 2025

- 1 2.9 Bayesian Learning
- 2 EX2.7 Normal Model
- 3 inverse gamma distribution
- 4 Example 2.8 Normal Model

- 1 2.9 Bayesian Learning
- 2 EX2.7 Normal Model
- 3 inverse gamma distribution
- 4 Example 2.8 Normal Model

Bayesian Learning

Review (Def 2.4)

事後確率密度 $g(\boldsymbol{\theta} \mid \tau)$ は事前確率密度 $g(\boldsymbol{\theta})$ と尤度 $g(\tau \mid \boldsymbol{\theta})$ の積に比例するのであった.

$$g(\boldsymbol{\theta} \mid \tau) \propto g(\tau \mid \boldsymbol{\theta})g(\boldsymbol{\theta})$$

- $g(\boldsymbol{\theta})$ は \mathcal{W}_p 内の KL を最小化する最適な (ベイズ学習機を学習する前の) 密度に関する事前情報を与える. この $g(\boldsymbol{\theta})$ を用いて, $f(\mathbf{x})$ のベイズ近似は事前予測密度 (prior predictive density) となる.

$$g(\mathbf{x}) = \int g(\mathbf{x} \mid \boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Bayesian Learning

- 一方、事後確率密度は \mathcal{W}_p 内の τ によって学習されたあとの最適な密度に関して改善された知識を与える。事後確率密度 $g(\boldsymbol{\theta} \mid \tau)$ を用いて、ベイズ学習器 $f(\boldsymbol{x})$ は事後予測密度 (posterior predictive density) となる。

$$g_{\tau}(\boldsymbol{x}) = g(\boldsymbol{x} \mid \tau) = \int g(\boldsymbol{x} \mid \boldsymbol{\theta})g(\boldsymbol{\theta} \mid \tau)d\boldsymbol{\theta}$$

ここで $g(\boldsymbol{x} \mid \boldsymbol{\theta}, \tau) = g(\boldsymbol{x} \mid \boldsymbol{\theta})$ を仮定した。つまり、尤度は $\boldsymbol{\theta}$ を介してのみ τ に依存する。 ($\tau \rightarrow \boldsymbol{\theta} \rightarrow$ 尤度)

Bayesian Learning

事前確率密度の選択は、通常次の2つを考慮して行われる。

- 事後確率密度の計算あるいはシミュレートが容易となるために十分にシンプルであるべき
- 関心のあるパラメータについて無知をモデル化するために十分に一般的であるべき
(気持ち：パラメータについてもちろん知らないことがあるから無難なものを使おう?)

パラメータの情報をそれほど多くは持たない事前確率密度は *uninformative*(無情報) と呼ばれる。一様または *flat*(平坦) な事前確率密度がよく用いられる。(Example 2.9)

Bayesian Learning

Remark

解析や数値計算のために、 θ を確率密度 $g(\theta)$ を持つ確率変数ベクトルとみることができ、学習後に事後確率密度 $g(\theta | \tau)$ に更新する.

これまでのスライドの議論により、次のように書くことができる.

$$g(\mathbf{x} | \tau) \propto \int g(\mathbf{x} | \theta) g(\tau | \theta) g(\theta) d\theta$$

したがって密度関数に影響しない定数は無視しても良い.

- 1 2.9 Bayesian Learning
- 2 EX2.7 Normal Model**
- 3 inverse gamma distribution
- 4 Example 2.8 Normal Model

Example 2.7 Normal Model

訓練データ $\mathcal{T} = \{X_1, \dots, X_n\}$ が尤度 $g(x \mid \boldsymbol{\theta})$ を用いてモデル化されると仮定し, $g(x \mid \boldsymbol{\theta})$ は

$$X \mid \boldsymbol{\theta} \sim \mathcal{N}(\mu, \sigma^2)$$

に従う確率密度関数とする. ただし, $\boldsymbol{\theta} := [\mu, \sigma^2]^\top$ である.

次に, モデルを完成させるために $\boldsymbol{\theta}$ の事前分布を決める必要がある. μ と σ^2 をそれぞれ別の事前分布を定めて, それらの積をとって $\boldsymbol{\theta}$ の事前分布を得る (独立性を仮定).

μ の事前分布として次のように決めることができる.

$$\mu \sim \mathcal{N}(\nu, \phi^2)$$

ベイズモデルでは, 事前確率密度のパラメータは通常 hyperparameters と呼ばれる.

Example 2.7 Normal Model

直接 σ^2 を与える代わりに、 $1/\sigma^2$ に対して次の事前分布を設定するのが便利である

$$\frac{1}{\sigma^2} \sim \mathbf{Gamma}(\alpha, \beta)$$

α と β が小さいほど、事前分布の情報量は少なくなる。この事前分布のもとで、 σ^2 は inverse gamma distribution(逆ガンマ分布)を持つと呼ばれる。

もし $1/Z \sim \mathbf{Gamma}(\alpha, \beta)$ なら、 Z の確率密度関数は $\exp(-\beta/x)/x^{\alpha+1}$ に比例する。(後に示す。Exercise 19)

Example 2.7 Normal Model

事後分布を求めるために (次スライドの計算のために) $g(\tau \mid \mu, \sigma^2)$ について考えると,

$$\begin{aligned}
 g(\tau \mid \mu, \sigma^2) &\propto \prod_i^n \frac{\exp \{ (x_i - \mu)^2 / (2\sigma^2) \}}{(\sigma^2)^{1/2}} \\
 &= \frac{\exp \{ \sum_i^n (x_i - \mu)^2 / (2\sigma^2) \}}{(\sigma^2)^{n/2}} \\
 &= (\sigma^2)^{-n/2} \exp \left\{ \frac{\frac{1}{n} \sum_i^n (x_i - \mu)^2}{(2\sigma^2)/n} \right\} \\
 &= (\sigma^2)^{-n/2} \exp \left\{ \frac{(\mu - \bar{x}_n)^2 + S_n^2}{(2\sigma^2)/n} \right\}
 \end{aligned}$$

ただし $S_n^2 = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ は (平均) 標本分散である.

Example 2.7 Normal Model

ベイズ事後確率密度は,

$$\begin{aligned}
 g(\mu, \sigma^2 \mid \tau) &\propto g(\mu) \times g(\sigma^2) \times g(\tau \mid \mu, \sigma^2) \\
 &\propto \exp \left\{ -\frac{(\mu - \nu)^2}{2\phi^2} \right\} \times \frac{\exp \{ -\beta/\sigma^2 \}}{(\sigma^2)^{\alpha+1}} \times \frac{\exp \{ \sum_i^n (x_i - \mu)^2 / (2\sigma^2) \}}{(\sigma^2)^{n/2}} \\
 &\propto (\sigma^2)^{-n/2-\alpha-1} \exp \left\{ -\frac{(\mu - \nu)^2}{2\phi^2} - \frac{\beta}{\sigma^2} - \frac{(\mu - \bar{x}_n)^2 + S_n^2}{2\sigma^2/n} \right\}
 \end{aligned}$$

(μ, σ^2) のすべての推論は事後確率密度関数によって表される.

Example 2.7 Normal Model

計算を簡単にするために、事後確率密度関数がよく知っている分布に属しているか調べることが有用である。

例えば、 σ^2 と τ が与えられたときの μ の条件付き確率密度は

$$g(\mu \mid \sigma^2, \tau) \propto \exp \left\{ -\frac{(\mu - \nu)^2}{2\phi^2} - \frac{(\mu - \bar{x}_n)^2}{2\sigma^2/n} \right\}$$

となる． (μ 以外を定数とみて、 μ についてできるだけ簡単にした) 更に簡単にすれば

$$(\mu \mid \sigma^2, \tau) \sim \mathcal{N}(\gamma_n \bar{x}_n + (1 - \gamma_n)\nu, \gamma_n \sigma^2/n)$$

ただし $\gamma_n := \frac{n}{\sigma^2} / \left(\frac{1}{\phi^2} + \frac{n}{\sigma^2} \right)$ と定義する．

Example 2.7 Normal Model

事後平均 $\mathbf{E}[\mu \mid \sigma^2, \tau] = \gamma_n \bar{x}_n + (1 - \gamma_n)\nu$ は事前平均 ν と標本平均 \bar{x}_n の重み付き線形和となっている. また, $n \rightarrow \infty$ とするとき, $\gamma_n \rightarrow 1$ で, 事後平均は最尤推定値 \bar{x}_n に近づく.

- 1 2.9 Bayesian Learning
- 2 EX2.7 Normal Model
- 3 inverse gamma distribution**
- 4 Example 2.8 Normal Model

inverse gamma distribution

後回しにしていた Excercse19 を簡単に示す.

Theorem

$1/Z \sim \mathbf{Gamma}(\alpha, \beta)$ ならば, Z の確率密度関数は $\exp(-\beta/x)/x^{\alpha+1}$ に比例する.

Proof.

確率変数の変数変換についての定理 (Them C.4) を用いる.

確率変数 X を $X = 1/Z \sim \mathbf{Gamma}(\alpha, \beta)$ と定めると, $Z = g(X) = \frac{1}{X}$ となる.

ヤコビアンは $\det J = \det \frac{\partial X}{\partial Z} = -\frac{1}{Z^2}$

X について確率密度関数は $f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$ よって

$$f_Z(z) = f_X(g^{-1}(z))|J| = \frac{\lambda^\alpha (1/z)^{\alpha-1} e^{-\lambda(1/z)}}{\Gamma(\alpha)} (1/z^2) \propto \frac{\exp(-\beta/z)}{z^{\alpha+1}}$$

□

- 1 2.9 Bayesian Learning
- 2 EX2.7 Normal Model
- 3 inverse gamma distribution
- 4 Example 2.8 Normal Model**

Example 2.8 Normal Model

$\int g(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$ という意味で真に確率密度ではない事前分布 $g(\boldsymbol{\theta})$ を用いることがある。ただし、結果として得られる事後分布 $g(\boldsymbol{\theta} | \tau) \propto g(\tau | \boldsymbol{\theta})g(\boldsymbol{\theta})$ が適切な確率密度関数になる場合に限る。このような事前分布は improper prior(広義の事前分布, 変則事前分布) と呼ばれる。

例として、 $\mu \sim \mathcal{N}(\nu, \phi^2)$ (2.43) において $\phi \rightarrow \infty$ を考える。(ϕ が大きくなるほど事前分布が無情報となる。)

$$g(\mu) \propto \exp\left(-\frac{(\mu - \nu)^2}{2\phi^2}\right) \rightarrow 1$$

これより平坦な事前分布となるが $\int g(\mu)d\mu = \infty$ となってしまう。事後分布は適切な密度となる。

Example 2.8 Normal Model

特に, $(\mu \mid \sigma^2, \tau)$ の条件付き事後分布は

$$(\mu \mid \sigma^2, \tau) \sim \mathcal{N}(\bar{x}_n, \sigma^2/n)$$

と簡単になる. これは重みパラメータ γ が 1 に収束するからである. また, σ^2 の周辺事後密度は次のように簡単になる.

$$g(\sigma^2 \mid \tau) = \int g(\mu, \sigma^2 \mid \tau) d\mu \propto (\sigma^2)^{-(n-1)/2 - \alpha - 1} \exp\left(-\frac{\beta + nS_n^2/2}{\sigma^2}\right)$$

これは次の関係を持つことが (以前の議論より) わかる.

$$\frac{1}{\sigma^2} \mid \tau \sim \mathbf{Gamma}\left(\alpha + \frac{n-1}{2}, \beta + \frac{1}{n}S_n^2\right)$$

Example 2.8 Normal Model

$g(\mu)$ だけでなく, σ^2 に対しても広義の事前分布を考えることができる.

$\frac{1}{\sigma^2} \sim \mathbf{Gamma}(\alpha, \beta)$ (2.44) において $\alpha \rightarrow 0, \beta \rightarrow 0$ とすれば広義の事前分布 $g(\sigma^2) \propto 1/\sigma^2$ または $g(1/\sigma^2) \propto 1/\sigma^2$ を得る.

$$\begin{aligned} g(\sigma^2) &\propto \frac{\exp(-\beta/\sigma^2)}{(\sigma^2)^{\alpha+1}} \\ &\longrightarrow \frac{1}{\sigma^2} \quad \left(\begin{matrix} \alpha \rightarrow 0 \\ \beta \rightarrow 0 \end{matrix} \right) \end{aligned}$$

変数変換など用いれば $g(1/\sigma^2) \propto 1/\sigma^2$ を得る. この場合, σ^2 の周辺事後密度は