1次元ガウス分布の学習と予測

復習: 1次元ガウス分布

1次元ガウス分布を以下に示す.

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2.64}$$

また, 計算の都合上, 対数を取ったものと比較することが多いのでそちらも示す.

$$\ln\mathcal{N}(x|\mu,\sigma^2)=-\frac{1}{2}\left\{\ln 2\pi+\ln\sigma^2+\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2.65}$$

1次元のガウス分布は平均値 μ と分散 σ^2 の2つのパラメータを持つ. したがって, 平均値のみを学習する場合, 分散のみを学習する場合, さらに両方を学習する場合の3パターンに分けることができる.

なお, 説明を簡単にするために, ここでは分散 σ^2 の代わりに逆数である**精度(precision)** $\lambda=\sigma^{-2}$ をパラメータとして導入する.

平均が未知のとき

条件

- ガウス分布に従うと仮定している N 個のデータを用いる.
- 今回はガウス分布の平均値 $\mu\in\mathbb{R}$ のみを学習する設定で推論を行う.
- そのため, ここでは精度パラメータ $\lambda\in\mathbb{R}^+$ は固定である.

この条件ってどんなとき?

精度は既知である状況を想像しよう

- 標準化された測定器具を例に考えてみる. 測定器具が一定の精度で測定を行うとき, その分散は既知とみなせるが, 測定対象の平均値は未知である.
- 例えば, 気温の長期間監視がこれに該当する. 温度計の精度(分散)は既知であるが, 測定する期間の平均気温は未知である.
- 他にもリスク管理はこれに当てはまる. 金融市場においてある投資商品のリスク(分散)はわかっているが, 期待リターン(平均)が不明確な場合, 投資家はリスクを基にして平均リターンの推定を行う.

ある観測値 $x\in\mathbb{R}$ について, 次のようなガウス分布を考える.

$$p(x|\mu)=\mathcal{N}(x|\mu,\lambda^{-1}) \tag{3.47}$$

μ に対しては, 次のようなガウス事前共役分布であることが知られている.

$$p(\mu)=\mathcal{N}(\mu|m,\lambda_\mu^{-1}) \tag{3.48}$$

$m \in \mathbb{R}$ および, $\lambda_\mu^{-1} \in \mathbb{R}^+$ は今回固定された超パラメータである. ここではこのように, 2種類の異なるガウス分布が登場するので注意. 平均値を未知としたガウス分布の学習モデルの共役事前分布がたまたま同じガウス分布になっているだけである.

平均未知の1次元ガウス分布の共役事前分布が1次元ガウス分布になることの導出

ガウス分布に従うと仮定している N 個の1次元連続データ $\mathbf{X} = \{x_1, \dots x_N\}$ を観測したとする. ベイズの定理を用いれば, 事後分布は以下のように書ける.

$$
\begin{aligned}
p(\mu|\mathbf{X}) &\propto p(\mathbf{X}|\mu)p(\mu) \\
&= \left\{ \prod_{n=1}^{N} p(x_n|\mu)p(\mu) \right\} \\
&= \left\{ \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \right\} \mathcal{N}(\mu|m, \lambda_\mu^{-1})
\end{aligned}
\tag{3.49}
$$

対数変換を行うことで, μ に関する関数形式を調べる.

$$
\begin{aligned}
\ln p(\mu|\mathbf{X}) &= \sum_{n=1}^{N} \ln \mathcal{N}(x_n|\mu, \lambda^{-1}) + \ln \mathcal{N}(\mu|m, \lambda_\mu^{-1}) + \text{const.} \\
&= -\frac{1}{2}\{(N\lambda + \lambda_\mu)\mu^2 - 2(\sum_{n=1}^{N} x_n\lambda + m\lambda_\mu)\mu\} + \text{const.}
\end{aligned}
\tag{3.50}
$$

一方で, 事後分布が次のようなガウス分布で書けるとする.

$$
p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\hat{m}, \hat{\lambda}_\mu^{-1}) \tag{3.51}
$$

これを μ について整理すると次のように書ける.

$$
\ln p(\mu|\mathbf{X}) = -\frac{1}{2}\{\hat{\lambda}_\mu\mu^2 - 2\hat{m}\hat{\lambda}_\mu\mu\} + \text{const.} \tag{3.52}
$$

係数の比較を行うことで, 事後分布のパラメータ $\hat{m}, \hat{\lambda_\mu}$ は次のように求まる.

$$
\hat{\lambda}_\mu = N\lambda + \lambda_\mu \tag{3.53}
$$

$$
\begin{aligned}
\hat{m} &= \frac{\lambda \sum_{n=1}^{N} x_n + \lambda_\mu m}{\hat{\lambda}_\mu} \\
&= \frac{\lambda \sum_{n=1}^{N} x_n + \lambda_\mu m}{N\lambda + \lambda_\mu}
\end{aligned}
\tag{3.54}
$$

式の気持ち

- (3.53) を見ると, 事後分布の精度は事前分布の精度に $N\lambda$ 足したものになっている. つまり, データ数 N が大きくなればなるほど, 平均 μ に対する事後分布の精度が単純に上昇していく.
- (3.54) は事前分布を尤度関数から得られる重み付き和となっている.
- データを観測すればするほど, 事前分布の平均 m による影響は次第に薄れ, データによって観測するほど単純な平均値 $\frac{1}{N}\sum_{n=1}^{N} x_n$ が支配的になってくることがわかる.

予測分布

事前分布 $p(\mu)$ を利用した未観測データ x_* に対する予測分布を見ていく. これは次のように周辺分布を計算することに対応する.

$$
\begin{aligned}
p(x_*) &= \int p(x_*|\mu)p(\mu)\mathrm{d}\mu \\
&= \int \mathcal{N}(x_*|\mu, \lambda^{-1})\mathcal{N}(\mu|m, \lambda_\mu^{-1})
\end{aligned}
\tag{3.55}
$$

ベイズの定理を使うと, いま求めたい予測分布 $p(x_*)$ と事前分布 $p(\mu)$ の間の関係は以下のようになる.

$$
p(\mu|x_*) = \frac{p(x_*|\mu)p(\mu)}{p(x_*)} \tag{3.56}
$$

対数をとって $p(x_*)$ に関して求めると, 以下のようになる

$$
\ln p(x_*) = \ln p(x_*|\mu) - \ln p(\mu|x_*) + \text{const.} \tag{3.57}
$$

ところで, $p(\mu|x_*)$ は x_* が与えられたときの μ の条件つき分布である. これは先程の事後分布の計算と同様に行えば良い. つまり,

$$
p(\mu|x_*) = \mathcal{N}(\mu|m(x_*), (\lambda + \lambda_\mu)^{-1}) \tag{3.58}
$$

ただし,

$$
m(x_*) = \frac{\lambda x_* + \lambda_\mu m}{\lambda + \lambda_\mu} \tag{3.59}
$$

これは式 (3.53) と式 (3.54) の $\mathbf{X}$ を x_* に書き換えただけ. これを (3.57) に代入すると, 以下のようになる.

$$
\begin{aligned}
\ln p(x_*) &= -\frac{1}{2}\{\lambda(x_* - \mu)^2 - (\lambda + \lambda_\mu)(\mu - m(x_*))^2\} + \text{const.} \\
&= -\frac{1}{2}(\frac{\lambda\lambda_\mu}{\lambda + \lambda_\mu}x_*^2 - \frac{2m\lambda\lambda_\mu}{\lambda + \lambda_\mu}x_*) + \text{const.}
\end{aligned}
\tag{3.60}
$$

これによって, x_* に関する2次関数として求めることができる. ここから平均と精度を計算する.

$$
p(x_*) = \mathcal{N}(x_*|\mu_*, \lambda_\mu^{-1}) \tag{3.61}
$$

ただし,

$$
\begin{aligned}
\lambda_* &= \frac{\lambda\lambda_\mu}{\lambda + \lambda_\mu} \\
\mu_* &= m
\end{aligned}
\tag{3.62}
$$

精度は逆数をとって分散として解釈してみると以下のようになる.

$$
\lambda_*^{-1} = \lambda^{-1} + \lambda_\mu^{-1} \tag{3.63}
$$

したがって, 予測分布の不確かさは, 観測分布と事前分布のそれぞれの不確かさを足し合わせたものである.

精度が未知の場合

条件

- ガウス分布に従うと仮定している N 個のデータを用いる.
- 今回はガウス分布の精度 λ のみを学習する設定で推論を行う.
- そのため, ここでは平均パラメータ μ は固定である.

$$
p(x|\lambda) = \mathcal{N}(x|\mu, \lambda^{-1}) \tag{3.64}
$$

この条件ってどんなとき?

- ある商品の平均的な月間販売数はわかっているけど, 月ごとの変動(分散) は不明な場合, 企業は在庫管理や生産計画を立てる際に平均販売数をもとに予測する.
- 一定の地域での平均的な電力消費量は把握できても, 季節や気候変動による消費の変動幅(分散)が不明な場合, 電力会社は電力供給の計画を立てる際に平均消費量を基にして予測を行う.

λ は正の実数値を持つので, 次のようなガンマ分布が考えられる.

$$p(\lambda) = \mathrm{Gam}(\lambda|a,b) \tag{3.65}$$

復習: ガンマ分布

ガンマ分布は正の実数 $\lambda \in \mathbb{R}^+$ を返す.

$$\mathrm{Gam}(\lambda|a,b) = C_G(a,b)\lambda^{a-1}e^{-b\lambda} \tag{2.56}$$

細かい計算の際には, 次のようにガンマ分布に対して対数を取った物を使うのが便利

$$\ln \mathrm{Gam}(\lambda|a,b) = (a-1)\ln\lambda - b\lambda + \ln C_G(a,b) \tag{2.58}$$

精度未知の1次元ガウス分布の共役事前分布がガンマ分布になることの導出

ベイズの定理を用いることで, λ の事前分布は次のように求められる.

$$\begin{aligned} p(\lambda|\mathbf{X}) &\propto p(\mathbf{X}|\lambda)p(\lambda) \\ &= \left\{\prod_{n=1}^{N} p(x_n|\lambda)\right\}p(\lambda) \\ &= \left\{\prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda^{-1})\right\}\mathrm{Gam}(\lambda|a,b) \end{aligned} \tag{3.66}$$

対数計算を行って, 具体的に λ に関する関数形式を調べる.

$$\begin{aligned} \ln p(\lambda|\mathbf{X}) &= \sum_{n=1}^{N} \ln \mathcal{N}(x_n|\mu,\lambda^{-1}) + \ln \mathrm{Gam}(\lambda|a,b) + \text{const.} \\ &= \left(\frac{N}{2} + a - 1\right)\ln\lambda - \left\{\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + b\right\}\lambda + \text{const.} \end{aligned} \tag{3.67}$$

λ と $\ln\lambda$ にかかる係数部分のみに注目すれば, これは次のようなガンマ分布になることがわかる.

$$p(\lambda|\mathbf{X}) = \mathrm{Gam}(\lambda|\hat{a},\hat{b}) \tag{3.68}$$

ただし,

$$\hat{a} = \frac{N}{2} + a \tag{3.68}$$

$$\hat{b} = \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + b \tag{3.69}$$

予測分布

予測分布 $p(x_*)$ は次のような積分計算によって導かれる.

$$p(x_*) = \int p(x_*|\lambda)p(\lambda)\,\mathrm{d}\lambda \tag{3.70}$$

ベイズの定理を用いれば, x_*, λ に関して次のような関係性が成立する.

$$p(\lambda|x_*) = \frac{p(x_*|\lambda)p(\lambda)}{p(x_*)} \tag{3.71}$$

対数を取って$p(\lambda)$ を無視すれば $p(\lambda)$ は次のように書ける.

$$\ln p(x_*) = \ln p(x_*|\lambda) - \ln p(\lambda|x_*) + \text{const.} \tag{3.72}$$

ここで, $p(\lambda|x_*)$ は1個の点x_*を観測したあとの事後分布と考えられるので, 式 (3.69) の結果を真似れば次のように書ける.

$$p(\lambda|x_*) = \mathrm{Gam}\left(\lambda|\frac{1}{2} + a, b(x_*)\right) \tag{3.73}$$

ただし,

$$b(x_*) = \frac{1}{2}(x_* - \mu)^2 + b \tag{3.74}$$

$p(\lambda|x_*)$ および, $p(x_*|\lambda)$ を式 (3.72) に代入して計算を進めると, 途中計算でλ に関する項は消えてしまい, 結果的に以下のような式で書ける.

$$\ln p(x_*) = -\frac{2a-1}{2}\ln\left\{1 + \frac{1}{2b}(x_* - \mu)^2\right\} + \text{const.} \tag{3.75}$$

この結果は, **スチューデントのt分布(Student's t distribution)** と呼ばれる分布に対数を取ったものになっている.

スチューデントのt分布

$$\mathrm{St}(x|\mu_s, \lambda_s, \nu_s) = \Gamma\left(\frac{\frac{\nu_s+1}{2}}{\Gamma\left(\frac{\nu_s}{2}\right)}\right)\left(\frac{\lambda_s}{\pi\nu_s}\right)^{\frac{1}{2}}\left\{1 + \frac{\lambda_s}{\nu_s}(x-\mu_s)^2\right\}^{-\frac{\nu_s+1}{2}} \tag{3.76}$$

対数を取ると以下のようになる.

$$\ln \mathrm{St}(x|\mu_s, \lambda_s, \nu_s) = -\frac{\nu_s+1}{2}\ln\left\{1 + \frac{\lambda_s}{\nu_s}(x-\mu_s)^2\right\} + \text{const.} \tag{3.77}$$

これと(3.75)と比較すると予測分布は次のように書ける.

$$p(x_*) = \mathrm{St}(x_*|\mu_s, \lambda_s, \nu_s) \tag{3.78}$$

ただし,

$$\begin{aligned}\mu_s &= \mu \\ \lambda_s &= \frac{a}{b} \\ \nu_s &= 2a\end{aligned} \tag{3.79}$$

平均と精度が未知の場合

平均と精度がともに未知であるケースを考える.

$$p(x|\mu,\lambda)=\mathcal{N}(x|\mu,\lambda^{-1}) \tag{3.80}$$

このモデルに対してこれまでのアプローチを合体して2つの事前分布を導入してベイズ推論を行うこともできるが, 実は1次元ガウス分布では, 次のようなm,β,a,b をパラメータとした**ガウス・ガンマ分布(Gauss-gamma distribution)** を事前分布として仮定すると, まったく同じ形式の事後分布が得られることが知られている.

> **ガウス・ガンマ分布**
>
> $$\begin{aligned}p(\mu,\lambda)&=\mathrm{NG}(\mu,\lambda|m,\beta,a,b)\\&=\mathcal{N}(\mu|m,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)\end{aligned} \tag{3.81}$$
>
> ここでは, 平均パラメータ μ の精度が固定ではなく, $\beta\lambda$ に置き換わっている.

基本的にはベイズの定理を用いて事後分布を計算する.

はじめに平均値 μ にのみ注目してみる. これは式 (3.53) および式 (3.54) で示される計算結果に対して精度の部分を $\beta\lambda$ とおけば, 計算結果を流用できる. したがって, 事後分布 $p(\mu|\lambda,\mathbf{X})$ の部分は次のようになる.

$$p(\mu|\lambda,\mathbf{X})=\mathcal{N}(\mu|\hat{m},(\hat{\beta}\lambda)^{-1}) \tag{3.82}$$

ただし,

$$\begin{aligned}\hat{\beta}&=N+\beta\\\hat{m}&=\frac{1}{\hat{\beta}}\left(\sum_{n=1}^{N}x_n+\beta m\right)\end{aligned} \tag{3.83}$$

次に, 残りの $p(\lambda|\mathbf{X})$ を求める. まず, 同時分布を条件付き分布の積によって次のように単純に書き下ろす.

$$p(\mathbf{X},\mu,\lambda)=p(\mu|\lambda,\mathbf{X})p(\lambda|\mathbf{X})p(\mathbf{X}) \tag{3.84}$$

ここから以下の式に変形できる.

$$\begin{aligned}p(\lambda|\mathbf{X})&=\frac{p(\mathbf{X},\mu,\lambda)}{p(\mu|\lambda,\mathbf{X})p(\mathbf{X})}\\&\propto\frac{p(\mathbf{X},\mu,\lambda)}{p(\mu|\lambda.\mathbf{X})}\end{aligned} \tag{3.85}$$

こうすることで, モデルとしてはじめから与えられている同時分布 $p(\mathbf{X},\mu,\lambda)=p(\mathbf{X}|\mu,\lambda)p(\mu,\lambda)$ と, 式(3.82)で既に求めてある $p(\mu|\lambda,\mathbf{X})$ を使うことでλ の事後分布が明らかになりそうである.
式(3.85) の対数を取って, 実際にλ に関する関数形式を求めてみる.

$$\begin{aligned}\ln p(\lambda|\mathbf{X})=&\left(\frac{N}{2}+a-1\right)\ln\lambda\\&-\left\{\frac{1}{2}\left(\sum_{n=1}^{N}x_n^2+\beta m^2-\hat{\beta}\hat{m}^2\right)+b\right\}\lambda+\mathrm{const.}\end{aligned} \tag{3.86}$$

これとガンマ分布の定義式と照らし合わせると, 以下のようにまとめられる.

$$p(\lambda|\mathbf{X})=\mathrm{Gam}(\lambda|\hat{a},\hat{b}) \tag{3.87}$$

ただし,

$$
\begin{aligned}
\hat{a} &= \frac{N}{2} + a \\
\hat{b} &= \frac{1}{2}\left(\sum_{n=1}^{N} x_n^2 + \beta m^2 - \hat{\beta} m^2\right) + b
\end{aligned}
\tag{3.88}
$$

予測分布

予測分布は次のような積分計算を行って, 平均と精度の2つの変数を積分除去してやる必要がある.

$$
p(x_*) = \int\int p(x_*|\mu, \lambda) p(\mu, \lambda)\, \mathrm{d}\mu\, \mathrm{d}\lambda \tag{3.89}
$$

ベイズの定理を使って, x_*に無関係な項を無視すれば, 予測分布$p(x_*)$ に対して次のような式が成立する.

$$
\ln p(x_*) = \ln p(x_*|\mu, \lambda) - \ln p(\mu, \lambda|x_*) + \mathrm{const.} \tag{3.90}
$$

式 (3.80) と式 (3.83) の計算結果を流用すれば, 2つ目の項は以下のように書ける.

$$
p(\mu, \lambda|x_*) = \mathcal{N}(\mu|m(x_*), \{(1+\beta)\lambda^{-1}\})\mathrm{Gam}\left(\lambda|\frac{1}{2} + a, b(x_*)\right) \tag{3.91}
$$

ただし,

$$
\begin{aligned}
m(x_*) &= \frac{x_* + \beta m}{1 + \beta} \\
b(x_*) &= \frac{b}{2(1+\beta)}(x_* - m)^2 + b
\end{aligned}
\tag{3.92}
$$

これを式(3.90)に代入して, x_*に関わる項のみで整理すると, 以下のような式になる.

$$
\ln p(x_*) = -\frac{1+2a}{2}\ln\left\{1 + \frac{\beta}{2(1+\beta)b}\right\}(x_* - m)^2 + \mathrm{const.} \tag{3.93}
$$

計算の途中で, 2つの変数 μ, λが消えるのがポイントである. これは1次元のスチューデントのt分布に対数をとったものと同じで, 式(3.76)で表される定義に基づけば, 以下のように予測分布を解析的に求めることができる.

$$
p(x_*) = \mathrm{St}(x_*|\mu_s, \lambda_s, \nu_s) \tag{3.94}
$$

ただし,

$$
\begin{aligned}
\mu_s &= m \\
\lambda_s &= \frac{\beta a}{(1+\beta)b} \\
\nu_s &= 2a
\end{aligned}
\tag{3.95}
$$