

## ベイズ本 3.5 章

線形回帰モデルをガウス分布を使って構築し、係数パラメータの学習を行い、さらに未観測データの予測を行う。

### ■3.5.1 モデルの構築

線形回帰モデル

$$y_n = \mathbf{w}^T \mathbf{x}_n + \varepsilon_n \quad (1)$$

を考える。パラメータは、

- $y_n \in \mathbb{R}$ : 出力
- $\mathbf{x}_n \in \mathbb{R}^M$ : 入力値
- $\mathbf{w} \in \mathbb{R}^M$ : パラメータ
- $\varepsilon_n \in \mathbb{R}$ : ノイズ

$$\varepsilon_n \sim \mathcal{N}(\varepsilon_n | 0, \lambda^{-1}) \quad (2)$$

–  $\lambda \in \mathbb{R}_+$ : 精度パラメータ

ノイズが正規分布に従うという仮定の下では、出力  $y_n$  の分布は、

$$p(y_n | \mathbf{x}_n, \mathbf{w}) = \mathcal{N}(y_n | \mathbf{w}^T \mathbf{x}_n, \lambda^{-1}) \quad (3)$$

である。ここでは、パラメータ  $\mathbf{w}$  を観測データから学習したい。

以下の事前分布を考える。

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{\Lambda}^{-1}) \quad (4)$$

ただし、

- $\mathbf{m} \in \mathbb{R}^M$ : 平均パラメータ
- $\mathbf{\Lambda}^{-1}$ : 精度行列パラメータ。正定値。

はハイパーパラメータである。

### ■3.5.2 事後分布と予測分布の計算

上記で構築した線形回帰モデルを使って、データを観測した後の事後分布と予測分布を求める。

## 事後分布

ベイズの定理より,

$$\begin{aligned}
 p(\mathbf{w}|\mathbf{Y}, \mathbf{X}) &= \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{w})}{p(\mathbf{X}, \mathbf{Y})} \\
 &= \frac{p(\mathbf{w})p(\mathbf{X}, \mathbf{Y}|\mathbf{w})}{p(\mathbf{X}, \mathbf{Y})} = \frac{p(\mathbf{w})p(\mathbf{Y}|\mathbf{X}, \mathbf{w})p(\mathbf{X})}{p(\mathbf{X}, \mathbf{Y})} \\
 &= \frac{p(\mathbf{w})p(\mathbf{Y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{Y}|\mathbf{X})} = \frac{p(\mathbf{w}) \prod_{n=1}^N p(y_n|\mathbf{x}_n, \mathbf{w})}{p(\mathbf{Y}|\mathbf{X})} \\
 &\propto p(\mathbf{w}) \prod_{n=1}^N p(y_n|\mathbf{x}_n, \mathbf{w})
 \end{aligned} \tag{5}$$

が得られる. (5) を  $\mathbf{w}$  について整理し,  $\mathbf{w}$  の分布を明らかにする. (5) の対数をとる.

$$\begin{aligned}
 \ln p(\mathbf{w}|\mathbf{Y}, \mathbf{X}) &= \ln \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Lambda}^{-1}) + \sum_{n=1}^N \ln \mathcal{N}(y_n|\mathbf{w}^T \mathbf{x}_n, \lambda^{-1}) + \text{const.} \\
 &= -\frac{1}{2}(\mathbf{w} - \mathbf{m})^T \mathbf{\Lambda}(\mathbf{w} - \mathbf{m}) - \frac{1}{2} \sum_{n=1}^N \lambda(y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \text{const.}
 \end{aligned} \tag{6}$$

である. ここで,

$$(\mathbf{w} - \mathbf{m})^T \mathbf{\Lambda}(\mathbf{w} - \mathbf{m}) = \mathbf{w}^T \mathbf{\Lambda} \mathbf{w} - 2\mathbf{w}^T \mathbf{\Lambda} \mathbf{m} + \text{const.} \tag{7}$$

$$\begin{aligned}
 \lambda(y_n - \mathbf{w}^T \mathbf{x}_n)^2 &= -2\lambda y_n \mathbf{w}^T \mathbf{x}_n + \lambda(\mathbf{w}^T \mathbf{x}_n)(\mathbf{w}^T \mathbf{x}_n) + \text{const.} \\
 &= -2\lambda \mathbf{w}^T \mathbf{x}_n y_n + \lambda \mathbf{w}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{w} + \text{const.}
 \end{aligned} \tag{8}$$

より,

$$\ln p(\mathbf{w}|\mathbf{Y}, \mathbf{X}) = -\frac{1}{2} \left( \mathbf{w}^T \left( \lambda \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T + \mathbf{\Lambda} \right) \mathbf{w} - 2\mathbf{w}^T \left( \lambda \sum_{n=1}^N y_n \mathbf{x}_n + \mathbf{\Lambda} \mathbf{m} \right) \right) + \text{const.} \tag{9}$$

を得る. したがって,  $\mathbf{w}$  の事後分布は,

$$p(\mathbf{w}|\mathbf{Y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\hat{\mathbf{m}}, \hat{\mathbf{\Lambda}}^{-1}) \tag{10}$$

である. ただし,

$$\hat{\mathbf{\Lambda}} = \lambda \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T + \mathbf{\Lambda} \tag{11}$$

$$\hat{\mathbf{m}} = \hat{\mathbf{\Lambda}}^{-1} \left( \lambda \sum_{n=1}^N y_n \mathbf{x}_n + \mathbf{\Lambda} \mathbf{m} \right) \tag{12}$$

である.

## 予測分布

次に, 新規入力値  $\mathbf{x}_*$  が与えられたときの出力値  $y_*$  の予測分布  $p(y_*|\mathbf{x}_*, \mathbf{Y}, \mathbf{X})$  を求める.

- 事前分布を使った場合の予測分布  $p(y_*|\mathbf{x}_*)$  を求める.

- 事前分布を事後分布に置き換えて,  $p(y_*|\mathbf{x}_*, \mathbf{Y}, \mathbf{X})$  を求める.

新規の入力データのベクトル  $\mathbf{x}_*$  と未知の出力値  $y_*$  に対して, ベイズの定理より,

$$p(\mathbf{w}|y_*, \mathbf{x}_*) = \frac{p(\mathbf{w})p(y_*|\mathbf{x}_*, \mathbf{w})}{p(y_*|\mathbf{x}_*)} \quad (13)$$

を得る. 対数をとると,

$$\ln p(y_*|\mathbf{x}_*) = \ln p(y_*|\mathbf{x}_*, \mathbf{w}) - \ln p(\mathbf{w}|y_*, \mathbf{x}_*) + \text{const.} \quad (14)$$

となる. ここで,  $p(\mathbf{w}|y_*, \mathbf{x}_*)$  については, データ  $(\mathbf{x}_*, y_*)$  を得た後の事後分布とみなせるので,

$$p(\mathbf{w}|y_*, \mathbf{x}_*) = \mathcal{N}(\mathbf{w}|\mathbf{m}(y_*), (\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}) \quad (15)$$

を得る. ただし,

$$\mathbf{m}(y_*) = (\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m}) \quad (16)$$

と表される. (3), (14), (15) より,

$$\ln p(y_*|\mathbf{x}_*) = \ln \mathcal{N}(y_*|\mathbf{w}^T\mathbf{x}_*, \lambda^{-1}) - \ln \mathcal{N}(\mathbf{w}|\mathbf{m}(y_*), (\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}) + \text{const.} \quad (17)$$

となる. ここで,

$$\begin{aligned} \ln \mathcal{N}(y_*|\mathbf{w}^T\mathbf{x}_*, \lambda^{-1}) &= -\frac{1}{2}\lambda(y_* - \mathbf{w}^T\mathbf{x}_*)^2 + \text{const.} \\ &= -\frac{1}{2}\lambda y_*^2 + \lambda\mathbf{w}^T\mathbf{x}_*y_* + \text{const.} \end{aligned} \quad (18)$$

であり,

$$\begin{aligned} -\ln \mathcal{N}(\mathbf{w}|\mathbf{m}(y_*), (\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}) &= \frac{1}{2}(\mathbf{w} - \mathbf{m}(y_*))^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})(\mathbf{w} - \mathbf{m}(y_*)) + \text{const.} \\ &= -\mathbf{w}^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})\mathbf{m}(y_*) + \frac{1}{2}\mathbf{m}(y_*)^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})\mathbf{m}(y_*) + \text{const.} \end{aligned} \quad (19)$$

である. また,

$$\begin{aligned} -\mathbf{w}^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})\mathbf{m}(y_*) &= -\mathbf{w}^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m}) \\ &= -\lambda\mathbf{w}^T\mathbf{x}_*y_* + \text{const.} \end{aligned} \quad (20)$$

および

$$\begin{aligned} &\frac{1}{2}\mathbf{m}(y_*)^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})\mathbf{m}(y_*) \\ &= \frac{1}{2}[(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m})]^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m}) \\ &= \frac{1}{2}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m})^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}(\lambda y_*\mathbf{x}_* + \mathbf{\Lambda}\mathbf{m}) \\ &= \frac{1}{2}\lambda^2\mathbf{x}_*^T(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}\mathbf{x}_*y_* + \mathbf{x}_*^T\lambda(\lambda\mathbf{x}_*\mathbf{x}_*^T + \mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{m}y_* + \text{const.} \end{aligned} \quad (21)$$

を得る。ここで、(21)における2つ目の等号は、 $\lambda \mathbf{x}_* \mathbf{x}_*^T + \mathbf{\Lambda}$ が正定値 (対称) であることを用いた。よって、(17) – (21) より、

$$\ln p(y_* | \mathbf{x}_*) = -\frac{1}{2} [(\lambda - \lambda^2 \mathbf{x}_*^T (\lambda \mathbf{x}_* \mathbf{x}_*^T + \mathbf{\Lambda})^{-1} \mathbf{x}_*) y_*^2 - 2 \mathbf{x}_*^T (\lambda \mathbf{x}_* \mathbf{x}_*^T + \mathbf{\Lambda})^{-1} \mathbf{\Lambda} \mathbf{m} y_*] + \text{const.} \quad (22)$$

のように、密度関数の対数は  $y_*$  の2次関数として表される。これは、1次元のガウス分布の密度関数の対数である。よって、予測分布は、

$$p(y_* | \mathbf{x}_*) = \mathcal{N}(y_* | \mu_*, \lambda_*^{-1}) \quad (23)$$

である。ただし、

$$\mu_* = \mathbf{m}^T \mathbf{x}_* \quad (24)$$

$$\lambda_*^{-1} = \lambda^{-1} + \mathbf{x}_*^T \mathbf{\Lambda}^{-1} \mathbf{x}_* \quad (25)$$

となる。データを観測した後の予測分布は、事前分布のパラメータ  $\mathbf{m}, \mathbf{\Lambda}$  の代わりに、事後分布のパラメータ  $\hat{\mathbf{m}}, \hat{\mathbf{\Lambda}}$  を当てはめればよい。

### ■3.5.3 モデルの比較

データ解析の分野では、あるデータセット  $\mathcal{D}$  に対して複数のモデルの良さを比較したい場合がある (モデル選択)。ベイズ学習においては、周辺尤度 (marginal likelihood) あるいはモデルエビデンス (model evidence)  $p(\mathcal{D})$  を複数のモデル同士で直接比較してモデル選択する方法が一般に行われている。これは、あるモデルに対する  $p(\mathcal{D})$  の値が、データ  $\mathcal{D}$  を生成する尤もらしさを表しているとされているためである。

線形回帰モデルでは、入力値  $\mathbf{X}$  は常に与えられているので、 $p(\mathbf{Y} | \mathbf{X})$  を比較すればよい。(5) より、

$$p(\mathbf{Y} | \mathbf{X}) = \frac{p(\mathbf{w}) \prod_{n=1}^N p(y_n | \mathbf{x}_n, \mathbf{w})}{p(\mathbf{w} | \mathbf{Y}, \mathbf{X})} \quad (26)$$

を得る。よって、

$$\begin{aligned} \ln p(\mathbf{Y} | \mathbf{X}) &= \ln p(\mathbf{w}) + \sum_{n=1}^N \ln p(y_n | \mathbf{x}_n, \mathbf{w}) - \ln p(\mathbf{w} | \mathbf{Y}, \mathbf{X}) \\ &= \ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{\Lambda}^{-1}) + \sum_{n=1}^N \mathcal{N}(y_n | \mathbf{w}^T \mathbf{x}_n, \lambda^{-1}) - \mathcal{N}(\mathbf{w} | \hat{\mathbf{m}}, \hat{\mathbf{\Lambda}}^{-1}) \quad (\because (3), (4), (10)) \end{aligned} \quad (27)$$

である。ここで、

$$\begin{aligned} \ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{\Lambda}^{-1}) &= \frac{1}{2} (\ln |\mathbf{\Lambda}| - M \ln 2\pi) - \frac{1}{2} (\mathbf{w} - \mathbf{m})^T \mathbf{\Lambda} (\mathbf{w} - \mathbf{m}) \\ &= \frac{1}{2} (\ln |\mathbf{\Lambda}| - M \ln 2\pi) - \frac{1}{2} \mathbf{w}^T \mathbf{\Lambda} \mathbf{w} + \mathbf{w}^T \mathbf{\Lambda} \mathbf{m} - \frac{1}{2} \mathbf{m}^T \mathbf{\Lambda} \mathbf{m} \end{aligned} \quad (28)$$

$$\begin{aligned} \sum_{n=1}^N \mathcal{N}(y_n | \mathbf{w}^T \mathbf{x}_n, \lambda^{-1}) &= \sum_{n=1}^N \left[ \frac{1}{2} (\ln \lambda - \ln 2\pi) - \frac{1}{2} \lambda (y_n - \mathbf{w}^T \mathbf{x}_n)^2 \right] \\ &= -\frac{1}{2} \sum_{n=1}^N (\lambda y_n^2 - \ln \lambda + \ln 2\pi) + \lambda \sum_{n=1}^N \mathbf{w}^T \mathbf{x}_n y_n - \frac{\lambda}{2} \sum_{n=1}^N \mathbf{w}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{w} \end{aligned} \quad (29)$$

$$- \ln \mathcal{N}(\mathbf{w} | \hat{\mathbf{m}}, \hat{\mathbf{\Lambda}}^{-1}) = -\frac{1}{2} (\ln |\hat{\mathbf{\Lambda}}| - M \ln 2\pi) + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{m}})^T \hat{\mathbf{\Lambda}} (\mathbf{w} - \hat{\mathbf{m}}) \quad (30)$$

である。また,

$$\frac{1}{2}(\mathbf{w} - \hat{\mathbf{m}})^T \hat{\mathbf{\Lambda}} (\mathbf{w} - \hat{\mathbf{m}}) = \frac{1}{2} \mathbf{w}^T \hat{\mathbf{\Lambda}} \mathbf{w} - \mathbf{w}^T \hat{\mathbf{\Lambda}} \hat{\mathbf{m}} + \frac{1}{2} \hat{\mathbf{m}}^T \hat{\mathbf{\Lambda}} \hat{\mathbf{m}} \quad (31)$$

であり, 第1項と第2項はそれぞれ,

$$\begin{aligned} \frac{1}{2} \mathbf{w}^T \hat{\mathbf{\Lambda}} \mathbf{w} &= \frac{1}{2} \mathbf{w}^T \left( \lambda \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T + \mathbf{\Lambda} \right) \mathbf{w} \\ &= \frac{\lambda}{2} \sum_{n=1}^N \mathbf{w}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{w} + \frac{1}{2} \mathbf{w}^T \mathbf{\Lambda} \mathbf{w} \end{aligned} \quad (32)$$

$$\begin{aligned} -\mathbf{w}^T \hat{\mathbf{\Lambda}} \hat{\mathbf{m}} &= -\mathbf{w}^T \hat{\mathbf{\Lambda}} \hat{\mathbf{\Lambda}}^{-1} \left( \lambda \sum_{n=1}^N y_n \mathbf{x}_n + \mathbf{\Lambda} \mathbf{m} \right) \\ &= -\lambda \sum_{n=1}^N \mathbf{w}^T \mathbf{x}_n y_n - \mathbf{w}^T \mathbf{\Lambda} \mathbf{m} \end{aligned} \quad (33)$$

である。(27) – (33) より, 周辺尤度の対数は,

$$\ln p(\mathbf{Y}|\mathbf{X}) = -\frac{1}{2} \left[ \sum_{n=1}^N (\lambda y_n^2 - \ln \lambda + \ln 2\pi) + \mathbf{m}^T \mathbf{\Lambda} \mathbf{m} - \ln |\mathbf{\Lambda}| - \hat{\mathbf{m}}^T \hat{\mathbf{\Lambda}} \hat{\mathbf{m}} + \ln |\hat{\mathbf{\Lambda}}| \right] \quad (34)$$

である。

### 最近傍法

新しい入力値  $\mathbf{x}_* \in \mathbb{R}^M$  に対して, ある誤差関数 (e.g. ユークリッド距離) に関して最も近い点  $\mathbf{x}_n \in \mathbb{R}^M$  を学習データから探す。

$$n_{\text{opt.}} = \underset{n \in \{1, \dots, N\}}{\operatorname{argmin}} \sum_{m=1}^M (x_{n,m} - x_{*,m})^2 \quad (35)$$

得られた  $n_{\text{opt.}}$  を使って予測値を  $y_* = y_{n_{\text{opt.}}}$  として採用するというアルゴリズム。

### パラメトリックモデルとノンパラメトリックモデル

- パラメトリックモデル
  - パラメータの数が固定である。
  - e.g. 線形回帰モデル
- ノンパラメトリックモデル
  - データ数に応じてモデルが変化する。
  - e.g. 最近傍法
  - バイズモデルではない。