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

$$y_n = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + \varepsilon_n \tag{1}$$

を考える．パラメータは,

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

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

  - $\lambda \in \mathbb{R}_+$: 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,

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

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

以下の事前分布を考える．

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

ただし,

- $\boldsymbol{m} \in \mathbb{R}^M$: 平均パラメータ
- $\boldsymbol{\Lambda}^{-1}$: 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，

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

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

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

である．ここで，

$$
(\boldsymbol{w}-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{w}-\boldsymbol{m}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{w} - 2\boldsymbol{w}^{\mathrm{T}}\Lambda\boldsymbol{m} + \mathrm{const.} \qquad (7)
$$

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

より，

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

を得る．したがって，$\boldsymbol{w}$ の事後分布は,

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

である．ただし,

$$
\hat{\boldsymbol{\Lambda}} = \lambda\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathrm{T}} + \boldsymbol{\Lambda} \qquad (11)
$$

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

である．

予測分布

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

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

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

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

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

を得る．対数をとると，

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

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

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

を得る．ただし，

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

と表される．(3), (14), (15) より，

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

となる．ここで，

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

であり，

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

である．また，

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

および

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

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

$$\ln p(y_*|\boldsymbol{x}_*) = -\frac{1}{2}\left[(\lambda - \lambda^2 \boldsymbol{x}_*^{\mathrm{T}}(\lambda \boldsymbol{x}_* \boldsymbol{x}_*^{\mathrm{T}} + \boldsymbol{\Lambda})^{-1}\boldsymbol{x}_*)y_*^2 - 2\boldsymbol{x}_*^{\mathrm{T}}\lambda(\lambda \boldsymbol{x}_* \boldsymbol{x}_*^{\mathrm{T}} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\boldsymbol{m}y_*\right] + \mathrm{const.} \tag{22}$$

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,

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

である．ただし，

$$\mu_* = \boldsymbol{m}^{\mathrm{T}}\boldsymbol{x}_* \tag{24}$$
$$\lambda_*^{-1} = \lambda^{-1} + \boldsymbol{x}_*^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{x}_* \tag{25}$$

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 $\boldsymbol{m}, \boldsymbol{\Lambda}$ の代わりに，事後分布のパラメータ $\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Lambda}}$ 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 $\boldsymbol{X}$ は常に与えられているので，$p(\boldsymbol{Y}|\boldsymbol{X})$ を比較すればよい．(5) より，

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

を得る．よって，

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

である．ここで，

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

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

$$-\ln \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Lambda}}^{-1}) = -\frac{1}{2}\left(\ln|\hat{\boldsymbol{\Lambda}}| - M\ln 2\pi\right) + \frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{m}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}(\boldsymbol{w} - \hat{\boldsymbol{m}}) \tag{30}$$

である．また，

$$\frac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{m}})^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}-\hat{\boldsymbol{m}}) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\boldsymbol{w} - \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{m}} + \frac{1}{2}\hat{\boldsymbol{m}}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{m}} \tag{31}$$

であり，第 1 項と第 2 項はそれぞれ，

$$\begin{aligned}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\boldsymbol{w} &= \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\left(\lambda\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathrm{T}} + \boldsymbol{\Lambda}\right)\boldsymbol{w} \\ &= \frac{\lambda}{2}\sum_{n=1}^{N}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{w} + \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{w} \qquad (32) \\ -\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{m}} &= -\boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^{-1}\left(\lambda\sum_{n=1}^{N}y_n\boldsymbol{x}_n + \boldsymbol{\Lambda}\boldsymbol{m}\right) \\ &= -\lambda\sum_{n=1}^{N}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n y_n - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{m} \qquad (33)\end{aligned}$$

である．(27) – (33) より，周辺尤度の対数は，

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

である．

最近傍法

新しい入力値 $\boldsymbol{x}_* \in \mathbb{R}^M$ 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 $\boldsymbol{x}_n \in \mathbb{R}^M$ を学習データから探す．

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

得られた $n_{\mathrm{opt.}}$ を使って予測値を $y_* = y_{n_{\mathrm{opt.}}}$ 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
  - ベイズモデルではない．