§3.5 線形回帰の例.

線形回帰モデルを，Gauss分布を使って構築.

3.5.1 モデルの構築.

$\mathbf{x}_n \in \mathbb{R}^M$: 入力値，$\mathbf{w} \in \mathbb{R}^M$: パラメータ，

$\varepsilon_n \sim \mathcal{N}(\varepsilon_n | 0, \lambda^{-1})$: ノイズ.　($\lambda > 0$: given)

出力値 $y_n \in \mathbb{R}$ を次のようにモデル化：

$$y_n = \mathbf{w}^T \mathbf{x}_n + \varepsilon_n$$ ← $\mathbf{w}^T\mathbf{x}_n$ から ε_n だけずれる.

⇒ y_n の確率分布は，

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

と表せる.

- やりたいこと：$\mathbf{w}$ の学習.

$\mathbf{w}$ について，次のような事前分布を仮定する：

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

$\mathbf{m} \in \mathbb{R}^M$, $\Lambda \in M_M(\mathbb{R})$, $\Lambda > O$: ハイパーパラメータ.

→ 分布の組みあわせ，構成する分布の種類，ハイパーパラメータの値の設定により，一つの確率モデル（データの生成過程についての一つの仮説）を明記できる.

モデルが明記できると…

サンプリングにより，仮説がカバーしている実現例を視覚化できる．

→モデルの特性・妥当性をチェックできる．

3.5.2 事後分布と予測分布の計算．

- データ $\mathcal{Y} := \{y_1, \cdots, y_N\}$, $\mathcal{X} := \{x_1, \cdots, x_N\}$ を観測後の事後分布を求める．

Bayes' Thm. より，

$$
\begin{aligned}
p(w|\mathcal{Y}, \mathcal{X}) &= \frac{p(\mathcal{Y}|\mathcal{X}, w)p(w)}{p(\mathcal{Y}|\mathcal{X})} \\
&\propto p(\mathcal{Y}|\mathcal{X}, w)p(w) \\
&= \left(\prod_{n=1}^{N} p(y_n|x_n, w)\right)p(w) \\
&= \left(\prod_{n=1}^{N} \mathcal{N}(y_n|w^T x_n, \lambda^{-1})\right)\mathcal{N}(w|m, \Lambda^{-1})
\end{aligned}
$$

$$
\begin{aligned}
&\log p(w|\mathcal{Y}, \mathcal{X}) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(y_n|w^T x_n, \lambda^{-1}) + \log \mathcal{N}(w|m, \Lambda^{-1}) + \text{const.} \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\left(\lambda(y_n - w^T x_n)^2 \underbrace{- \log\lambda}_{\text{const.}}\right)\right) \\
&\quad - \frac{1}{2}\left((w-m)^T\Lambda(w-m) \underbrace{- \log(\det\Lambda)}_{\text{const.}}\right) + \text{const.}
\end{aligned}
$$

$$= -\frac{\lambda}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 - \frac{1}{2}(\mathbf{w}-\mathbf{m})^T\Lambda(\mathbf{w}-\mathbf{m}) + \text{const.}$$

$$= -\frac{1}{2}\Big(\lambda\sum_{n=1}^{N}(\underbrace{y_n^2}_{\text{const.}} - 2\mathbf{w}^T\mathbf{x}_n y_n + \mathbf{w}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{w})$$

$$+ \mathbf{w}^T\Lambda\mathbf{w} - 2\mathbf{w}^T\Lambda\mathbf{m} + \underbrace{\mathbf{m}^T\Lambda\mathbf{m}}_{\text{const.}}\Big) + \text{const.}$$

$$= -\frac{1}{2}\left(\mathbf{w}^T\left(\lambda\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + \Lambda\right)\mathbf{w} - 2\mathbf{w}^T\left(\lambda\sum_{n=1}^{N}\mathbf{x}_n y_n + \Lambda\mathbf{m}\right)\right) + \text{const.}$$

↑ M次元Gauss分布のpdfのlogをとったもの

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

$$\hat{\Lambda} := \lambda\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T + \Lambda,\quad \hat{\mathbf{m}} := \hat{\Lambda}^{-1}\left(\lambda\sum_{n=1}^{N}\mathbf{x}_n y_n + \Lambda\mathbf{m}\right).$$

- 新規入力値 $\mathbf{x}_*$: given のときの出力値 y_* の予測分布を求める.

Bayes' Thm. より,

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

$$\therefore\ \log p(y_*|\mathbf{x}_*)$$

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

上の事後分布の結果を用いると,

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

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

$$\therefore \log p(\boldsymbol{w} \mid y_*, \boldsymbol{x}_*)$$

$$= \log \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}(y_*), (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1})$$

$$= -\frac{1}{2}\Big((\boldsymbol{w} - \boldsymbol{m}(y_*))^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)(\boldsymbol{w} - \boldsymbol{m}(y_*))$$

$$\underbrace{- \log(\det(\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda))}_{\text{const.}}\Big) + \text{const.}$$

$$= -\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}(y_*))^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)(\boldsymbol{w} - \boldsymbol{m}(y_*)) + \text{const.}$$

$$= -\frac{1}{2}\Big(\underbrace{\boldsymbol{w}^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{w}}_{\text{const.}} - \boldsymbol{w}^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{m}(y_*)$$

$$- \boldsymbol{m}(y_*)^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{w} + \boldsymbol{m}(y_*)^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{m}(y_*)\Big) + \text{const}$$

$$(\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^T = \lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda.$$

$$= -\frac{1}{2}\Big(\boldsymbol{m}(y_*)^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{m}(y_*)$$

$$- 2\boldsymbol{w}^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)\boldsymbol{m}(y_*)\Big) + \text{const.}$$

$$= -\frac{1}{2}\Big((\lambda \boldsymbol{x}_* y_* + \Lambda \boldsymbol{m})^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1}(\lambda \boldsymbol{x}_* y_* + \Lambda \boldsymbol{m})$$

$$- 2\boldsymbol{w}^T (\lambda \boldsymbol{x}_* y_* + \underbrace{\Lambda \boldsymbol{m}}_{\text{const.}})\Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\lambda^2 \boldsymbol{x}_*^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1} \boldsymbol{x}_* y_*^2 + 2\lambda \boldsymbol{x}_*^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1} \Lambda \boldsymbol{m} y_*$$

$$+ \underbrace{\boldsymbol{m}^T \Lambda (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1} \Lambda \boldsymbol{m}}_{\text{const.}} - 2\lambda \boldsymbol{w}^T \boldsymbol{x}_* y_*\Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\lambda^2 \boldsymbol{x}_*^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1} \boldsymbol{x}_* y_*^2$$

$$+ 2\lambda\big(\boldsymbol{x}_*^T (\lambda \boldsymbol{x}_* \boldsymbol{x}_*^T + \Lambda)^{-1} \Lambda \boldsymbol{m} - \boldsymbol{w}^T \boldsymbol{x}_*\big) y_*\Big) + \text{const.}$$

また，

$$\log p(y_* | x_*, w)$$

$$= \log \mathcal{N}(y_* | w^T x_*, \lambda^{-1})$$

$$= -\frac{1}{2}\left(\lambda (y_* - w^T x_*)^2 \underbrace{- \log \lambda}_{\text{const.}}\right) + \text{const.}$$

$$= -\frac{1}{2}\lambda\left(y_*^2 - 2w^T x_* y_* + \underbrace{(w^T x_*)^2}_{\text{const.}}\right) + \text{const.}$$

$$= -\frac{1}{2}\lambda\left(y_*^2 - 2w^T x_* y_*\right) + \text{const.}$$

$$\therefore \log p(y_* | x_*)$$

$$= -\frac{1}{2}\lambda\left(y_*^2 - 2w^T x_* y_*\right)$$

$$+ \frac{1}{2}\Big(\lambda^2 x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} x_* y_*^2$$

$$+ 2\lambda\left(x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} \Lambda m - w^T x_*\right) y_*\Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\left(\lambda - \lambda^2 x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} x_*\right) y_*^2$$

$$- 2\lambda x_*^T \left(\cancel{w} + (\lambda x_* x_*^T + \Lambda)^{-1} \Lambda m - \cancel{w}\right) y_*\Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\left(\lambda - \lambda^2 x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} x_*\right) y_*^2$$

$$- 2\lambda x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} \Lambda m y_*\Big) + \text{const.}$$

↖ Gauss分布のpdfのlogをとったもの

$$\therefore p(y_* | x_*) = \mathcal{N}(y_* | \mu_*, \lambda_*^{-1}),$$

$$\lambda_* := \lambda - \lambda^2 x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} x_*,$$

$$\mu_* := \lambda_*^{-1} \lambda x_*^T (\lambda x_* x_*^T + \Lambda)^{-1} \Lambda m.$$

もう少し計算をすすめる.

Sherman-Morrison公式より,

$$(\Lambda + \lambda x_* x_*^T)^{-1} = \Lambda^{-1} - (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} \lambda \Lambda^{-1} x_* x_*^T \Lambda^{-1}.$$

$\therefore \lambda_*$

$$= \lambda - \lambda^2 x_*^T \left(\Lambda^{-1} - (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} \lambda \Lambda^{-1} x_* x_*^T \Lambda^{-1} \right) x_*$$

$$= \lambda - \lambda^2 x_*^T \Lambda^{-1} x_* + \lambda^3 (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} (x_*^T \Lambda^{-1} x_*)^2$$

$$= \frac{\lambda + \cancel{\lambda^2 x_*^T \Lambda^{-1} x_*} - \cancel{\lambda^2 x_*^T \Lambda^{-1} x_*} - \cancel{\lambda^3 (x_*^T \Lambda^{-1} x_*)^2} + \cancel{\lambda^3 (x_*^T \Lambda^{-1} x_*)^2}}{1 + \lambda x_*^T \Lambda^{-1} x_*}$$

$$= \lambda (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1}.$$

$$\lambda_*^{-1} = \lambda^{-1} (1 + \lambda x_*^T \Lambda^{-1} x_*) = \lambda^{-1} + x_*^T \Lambda^{-1} x_*.$$

μ_*

$$= \lambda_*^{-1} \lambda x_*^T \left(\Lambda^{-1} - (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} \lambda \Lambda^{-1} x_* x_*^T \Lambda^{-1} \right) \Lambda m$$

$$= \lambda_*^{-1} \lambda \left(1 - \lambda (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} x_*^T \Lambda^{-1} x_* \right) x_*^T m$$

$$= \lambda_*^{-1} \lambda (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} (1 + \cancel{\lambda x_*^T \Lambda^{-1} x_*} - \cancel{\lambda x_*^T \Lambda^{-1} x_*}) m^T x_*$$

$$= \lambda_*^{-1} \lambda (1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1} m^T x_*$$

$$= \cancel{\lambda^{-1} (1 + \lambda x_*^T \Lambda^{-1} x_*)} \lambda \cancel{(1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1}} m^T x_*$$

$$= m^T x_*.$$

結局,

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

$$\mu_* = m^T x_*, \quad \lambda_*^{-1} = \lambda(1 + \lambda x_*^T \Lambda^{-1} x_*)^{-1}.$$

- N個のデータを観測したあとの予測分布は,上の式で m を $\hat{m}$ に, Λ を $\hat{\Lambda}$ にすればよい.

3.5.3 モデルの比較.

- モデル選択:データセット $\mathcal{D}$ に対して,複数のモデルの良さを比較する.
  - 予測分布の可視化をするのも一案.可視化できない場合は…?

→ Bayes推論では,周辺尤度(モデルエビデンス) $p(\mathcal{D})$ を比較するのが一般的.

モデルが $\mathcal{D}$ を生成するもっともらしさ.

- 線形回帰モデルの例では,

$$p(\mathcal{D}) = p(Y|X)$$

X は常に与えられている.

$$= \frac{p(w) \prod_{n=1}^{N} p(y_n|x_n, w)}{p(w|Y, X)}.$$

対数をとって,

$$\log p(Y|X)$$
$$= \log p(w) + \sum_{n=1}^{N} \log p(y_n|x_n, w) - \log p(w|Y, X)$$

$$= \log \mathcal{N}(w \mid m, \Lambda^{-1}) + \sum_{n=1}^{N} \log \mathcal{N}(y_n \mid w^T x_n, \lambda^{-1})$$

$$- \log \mathcal{N}(w \mid \hat{m}, \hat{\Lambda}^{-1})$$

$$= -\frac{1}{2}\left((w-m)^T \Lambda (w-m) - \log(\det \Lambda) + \cancel{M \log 2\pi}\right)$$

$$- \frac{1}{2}\sum_{n=1}^{N}\left(\lambda (y_n - w^T x_n)^2 - \log \lambda + \log 2\pi\right)$$

$$+ \frac{1}{2}\left((w-\hat{m})^T \hat{\Lambda} (w-\hat{m}) - \log(\det \hat{\Lambda}) + \cancel{M \log 2\pi}\right)$$

$$= -\frac{1}{2}\Big(\cancel{w^T \Lambda w} \cancel{- 2 w^T \Lambda m} + m^T \Lambda m - \log(\det \Lambda)$$

$$+ \lambda \sum_{n=1}^{N}\left(y_n^2 \cancel{- 2 w^T x_n y_n} + \cancel{(w^T x_n)^2}\right) - N \log \lambda + N \log 2\pi$$

$$\cancel{- w^T \hat{\Lambda} w} + \cancel{2 w^T \hat{\Lambda} \hat{m}} - \hat{m}^T \hat{\Lambda} \hat{m} + \log(\det \hat{\Lambda})\Big)$$

$$= -\frac{1}{2}\Big(\lambda \sum_{n=1}^{N} y_n^2 - N \log \lambda + N \log 2\pi$$

$$+ m^T \Lambda m - \log(\det \Lambda) - \hat{m}^T \hat{\Lambda} \hat{m} + \log(\det \hat{\Lambda})\Big).$$

$$\rightarrow M^* := \underset{M}{\operatorname{argmax}} \log p(Y \mid X)$$

とすればよい.

・モデルエビデンスはデータ数 N に依存することに注意.