ベイズ深層学習

3. ベイズ推論の基礎：後編

joeyoji

July 1, 2021

目次

ベイズ深層学習

joeyoji

本発表にあたっていくつか説明事項があります.

- 正定値行列は実対称行列に対して定義しています.
- 本発表に掲載されている命題等の証明は独自に証明しているものです. それ故誤りがございましたらご指摘ください.
- 本発表ではいくつか実験を行なっていますが, その実験には R 言語を用いました.
- 式番号について,(2. ∼), (3. ∼), (A. ∼) と表されるものはセミナーの本のものと揃えました. これらは必ずしも順番通りには掲載しておりません. それ以外の式番号は独自のものです.
- 途中よく分かるようでよく分からないアニメがありますが, 出来れば温かい目でご覧ください.

線形回帰モデルの定義

Definition 1 (線形回帰モデル)

入力 $\boldsymbol{x} \in \mathbb{R}^{H_0}$ から出力 $y \in \mathbb{R}$ を予測したい. パラメータ $\boldsymbol{w} \in \mathbb{R}^{H_1}$ と特徴量関数 $\boldsymbol{\phi} : \mathbb{R}^{H_0} \to \mathbb{R}^{H_1}$ を用いて次のように定義されるモデルを線形回帰モデルという.

$$y = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) + \epsilon \tag{2.1}$$

但し ϵ は予測 $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$ と出力 y の誤差を表す.

本発表では主にこの線形回帰モデルを扱い, ベイズ推論によるモデルの学習とテストデータの予測を行う.
線形回帰モデル (2.1) に於いて誤差 ϵ が平均 0, 固有の分散 $\sigma_y^2 > 0$ の正規分布に従うとすれば, その確率密度函数が存在し, パラメータ $\boldsymbol{w}$ と入力 $\boldsymbol{x}$ が与えられたときの出力 y の条件付き分布は

$$y \mid \boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}), \sigma_y^2) \tag{3.67}$$

というように表すことができる.

ベイズ線形回帰モデル

ベイズ法ではパラメータに対して事前分布を導入して予測を行う．今回はパラメータ $\boldsymbol{w}$ の事前分布を平均 $\mathbf{0}$ 分散 $\sigma_w^2 \boldsymbol{I}$ の正規分布で与える $(\sigma_w^2 > 0)$. この事前分布は確率密度函数を持つ.

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \boldsymbol{I}) \tag{3.68}$$

N 組の入出力 $\boldsymbol{X} \times \boldsymbol{Y} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ に対して, 各入出力が独立であるとするとベイズ線形回帰モデルの同時分布は次のように表される.

$$p(\boldsymbol{Y}, \boldsymbol{w} | \boldsymbol{X}) = p(\boldsymbol{w}) p(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{i=1}^{N} p(y_i | \boldsymbol{x}_i, \boldsymbol{w}) \tag{3.66}$$

事前分布からのサンプリング

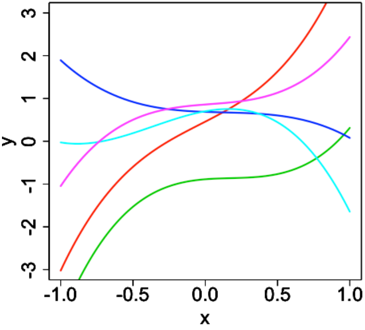

Figure 1: 事前分布からのサンプル.$H_1 = 4$ とし特徴量函数は $\boldsymbol{\phi}(x) = (x^3, x^2, x, 1)^\top$ で事前分布の分散$\sigma_w^2 = 1$ とした.

ベイズ回帰モデルでは, データを観測する以前に候補となる函数を事前分布からサンプリングし, モデル函数の具体例を表せる.

事後分布の計算

以上でモデル関数と事前分布を導入した．ここからは事後分布の具体形を求める. 事後分布は

$$p(\boldsymbol{w}|\boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})}{\int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}} \tag{3.69}$$

で与えられる. 但し各データは独立である仮定を置いているので $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{w})$ である.
事後分布は $\boldsymbol{w}$ についての確率密度函数となるので, $\boldsymbol{w}$ に関して式を整理する. 式 (3.67),(3.68) は共に正規分布であり, いずれも $\boldsymbol{w}$ は指数部に含まれる為ここに注目して考える．すなわち対数をとって考える.

$\boldsymbol{z} \in \mathbb{R}^n$, 対称行列 $\boldsymbol{A}, \boldsymbol{A}' \in \mathbb{R}^{n\times n}, \boldsymbol{b}, \boldsymbol{b}' \in \mathbb{R}^n, c, c' \in \mathbb{R}$ とする. これらに対して $\boldsymbol{A}'' = \boldsymbol{A} + \boldsymbol{A}', \boldsymbol{b}'' = \boldsymbol{b} + \boldsymbol{b}', c'' = c + c'$ とする. 但し以下では $\boldsymbol{A}, \boldsymbol{A}''$ は正則を仮定している. 平方完成の式は

$$\begin{aligned}\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}^\top \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{b} + c &= (\boldsymbol{z}^\top + \boldsymbol{b}^\top \boldsymbol{A}^{-1})\boldsymbol{A}(\boldsymbol{z} + \boldsymbol{A}^{-1}\boldsymbol{b}) - \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} + c \\ &= (\boldsymbol{z} + \boldsymbol{A}^{-1}\boldsymbol{b})^\top \boldsymbol{A}(\boldsymbol{z} + \boldsymbol{A}^{-1}\boldsymbol{b}) - \boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{b} + c\end{aligned} \tag{b}$$

である. これを用いると

$$\begin{aligned}(\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}^\top \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{b} + c) &+ (\boldsymbol{z}^\top \boldsymbol{A}'\boldsymbol{z} + \boldsymbol{b}'^\top \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{b}' + c') \\ &= \boldsymbol{z}^\top \boldsymbol{A}''\boldsymbol{z} + \boldsymbol{b}''^\top \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{b}'' + c'' \\ &= (\boldsymbol{z} + \boldsymbol{A}''^{-1}\boldsymbol{b}'')^\top \boldsymbol{A}''(\boldsymbol{z} + \boldsymbol{A}''^{-1}\boldsymbol{b}'') - \boldsymbol{b}''^\top \boldsymbol{A}''^{-1}\boldsymbol{b}'' + c''\end{aligned} \tag{c}$$

となる.

事後分布の計算 (2)

尤度関数 $p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})$ の指数部はデータの独立性と (3.67) により

$$-\frac{1}{2}\left\{\boldsymbol{w}^\top\left(\sigma_y^{-2}\sum_{i=1}^N \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^\top\right)\boldsymbol{w}-\left(\sigma_y^{-2}\sum_{i=1}^N y_i\boldsymbol{\phi}(\boldsymbol{x}_i)^\top\right)\boldsymbol{w}\right.$$
$$\left.-\boldsymbol{w}^\top\left(\sigma_y^{-2}\sum_{i=1}^N y_i\boldsymbol{\phi}(\boldsymbol{x}_i)\right)+\left(\sigma_y^{-2}\sum_{i=1}^N y_i^2\right)\right\} \qquad \text{(d)}$$

であり事前分布 $p(\boldsymbol{w})$ の指数部は (3.68) により

$$-\frac{1}{2}\boldsymbol{w}^\top(\sigma_w^{-2}\boldsymbol{I})\boldsymbol{w} \qquad \text{(e)}$$

である.

事後分布の計算 (3)

式 (c),(d),(e) により事後分布 (3.69) の分子の指数部は

$$-\frac{1}{2}\left\{(\boldsymbol{w}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{w}-\hat{\boldsymbol{\mu}})-\hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}+C\right\} \tag{f}$$

と表せる. 但し

$$\hat{\boldsymbol{\Sigma}}^{-1}=\sigma_y^{-2}\sum_{i=1}^N\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^\top+\sigma_w^{-2}\boldsymbol{I} \tag{3.72}$$

$$\hat{\boldsymbol{\mu}}=\hat{\boldsymbol{\Sigma}}\sigma_y^{-2}\sum_{i=1}^N y_i\boldsymbol{\phi}(\boldsymbol{x}_i) \tag{3.73}$$

$$C=\sigma_y^{-2}\sum_{i=1}^N y_i^2 \tag{g}$$

とした. 但し定義した順番により $\hat{\boldsymbol{\Sigma}}^{-1}$ の肩の -1 は形式的なものとし, 逆に $\hat{\boldsymbol{\Sigma}}$ を $\hat{\boldsymbol{\Sigma}}^{-1}$ の逆行列と定義する. いま事前分布の分散共分散行列は正定値行列であるので (3.72) は必ず逆行列 $\hat{\boldsymbol{\Sigma}}$ を持つ (命題 4).

事後分布の計算 (4)

よって事後分布 (3.69) の分子は改めて

$$p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})p(\boldsymbol{w}) = \frac{1}{\sqrt{2\pi}^{N+H_1}\sigma_y^N\sigma_w^{H_1}}\exp\left[-\frac{1}{2}\left\{(\boldsymbol{w}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{w}-\hat{\boldsymbol{\mu}})-\hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}+C\right\}\right] \tag{h}$$

となる. これを $\boldsymbol{w}$ に関して積分すると分母 (周辺尤度) になる (定理 2).

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\sqrt{\det\hat{\boldsymbol{\Sigma}}}}{\sqrt{2\pi}^N\sigma_y^N\sigma_w^{H_1}}\exp\left\{-\frac{1}{2}\left(-\hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}+C\right)\right\} \tag{i}$$

従って事後分布は次のように求まる.

$$p(\boldsymbol{w}|\boldsymbol{Y},\boldsymbol{X}) = \frac{1}{\sqrt{2\pi}^{H_1}\sqrt{\det\hat{\boldsymbol{\Sigma}}}}\exp\left\{-\frac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{w}-\hat{\boldsymbol{\mu}})\right\} \tag{3.71}$$

すなわち $\boldsymbol{w} \mid \boldsymbol{Y},\boldsymbol{X} \sim \mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ である. さらに尤度函数が正規分布に対して, 正規分布は共役事前分布であることがわかる.

予測分布の定義

次に予測分布を計算する. 予測分布の定義は以下である.

Definition 2 (予測分布)

予測分布はモデル関数を事後分布で平均したものである. すなわち入力 $\boldsymbol{x}_*$ が与えられたときの予測値を y_* とすると,

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X}) = \mathbb{E}_w[p(y_*|\boldsymbol{x}_*, \boldsymbol{w})] = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{Y}, \boldsymbol{X})d\boldsymbol{w} \quad \text{(k)}$$

で与えられる.

予測分布は $\boldsymbol{w}$ について積分するので, まず $\boldsymbol{w}$ について整理する. これは事後分布を求めたときと同様に, 被積分函数の指数部に注目すればよい.

予測分布の計算

事後分布の指数部は (3.71) により

$$-\frac{1}{2}\left\{\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{w}-(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}})^\top\boldsymbol{w}-\boldsymbol{w}^\top(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}})+\hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right\} \qquad \text{(l)}$$

また条件付き分布の指数部は (3.67) により

$$\begin{aligned}-\frac{1}{2}\Big\{\boldsymbol{w}^\top\left(\sigma_y^{-2}\boldsymbol{\phi}(\boldsymbol{x}_*)\boldsymbol{\phi}(\boldsymbol{x}_*)^\top\right)\boldsymbol{w}-\left(\sigma_y^{-2}y_*\boldsymbol{\phi}(\boldsymbol{x}_*)\right)^\top\boldsymbol{w}\\ -\boldsymbol{w}^\top\left(\sigma_y^{-2}y_*\boldsymbol{\phi}(\boldsymbol{x}_*)\right)+\sigma_y^{-2}y_*^2\Big\} \qquad \text{(m)}\end{aligned}$$

予測分布の計算 (2)

式 (c),(l),(m) により式 (k) の被積分函数の指数部は

$$-\frac{1}{2}\left\{(\boldsymbol{w}-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}})^\top\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{w}-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}})-\tilde{\boldsymbol{\mu}}^\top\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}}+D\right\} \qquad \text{(n)}$$

と表せる. 但し

$$\tilde{\boldsymbol{\Sigma}}^{-1}=\hat{\boldsymbol{\Sigma}}^{-1}+\sigma_y^{-2}\boldsymbol{\phi}(\boldsymbol{x}_*)\boldsymbol{\phi}(\boldsymbol{x}_*)^\top \qquad \text{(o)}$$

$$\tilde{\boldsymbol{\mu}}=\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}+\sigma_y^{-2}y_*\boldsymbol{\phi}(\boldsymbol{x}_*) \qquad \text{(p)}$$

$$D=\hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}+\sigma_y^{-2}y_*^2 \qquad \text{(q)}$$

とした.

予測分布の計算 (3)

式 (n) により予測分布 (k) は

$$
\begin{aligned}
&p(y_* | \boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X}) = \\
&\int \frac{1}{\sqrt{2\pi}^{H_1+1} \sigma_y \sqrt{\det \hat{\boldsymbol{\Sigma}}}} \exp\left[-\frac{1}{2} \left\{ (\boldsymbol{w} - \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}})^\top \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{w} - \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}}) - \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}} + D \right\} \right] d\boldsymbol{w} \\
&= \frac{\sqrt{2\pi}^{H_1} \sqrt{\det \tilde{\boldsymbol{\Sigma}}}}{\sqrt{2\pi}^{H_1+1} \sigma_y \sqrt{\det \hat{\boldsymbol{\Sigma}}}} \exp\left\{ -\frac{1}{2} \left(-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}} + D \right) \right\} \\
&= \frac{\sqrt{\det \tilde{\boldsymbol{\Sigma}}}}{\sqrt{2\pi} \sigma_y \sqrt{\det \hat{\boldsymbol{\Sigma}}}} \exp\left\{ -\frac{1}{2} \left(-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\mu}} + D \right) \right\} \qquad \text{(r)}
\end{aligned}
$$

となる. あとは行列式の整理をし被指数函数部分を y_* についいて整理をする.

準備 (行列式)

まずは行列式の整理から行う.

Proposition 1 (ブロック行列の行列式)

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ をそれぞれ $m \times m, m \times n, n \times m, n \times n$ 行列とし, $\boldsymbol{D}$ は正則であるとする. このとき

$$\det \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} = \det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}) \det \boldsymbol{D} \tag{s}$$

Proof.

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{O}_{m,n} \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{O}_{n,m} & \boldsymbol{D} \end{pmatrix}$$

である. 両辺の行列式をとれば式 (s) を得る. □

Proposition 2 (Woodbury の公式)

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{U}, \boldsymbol{V}$ をそれぞれ $m \times m, n \times n, m \times n, n \times m$ 行列とし, $\boldsymbol{A}, \boldsymbol{B}$ は正則であるとする. このとき $\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}$ が正則であるならば

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} \tag{A.1}$$

Proof.

計算して確かめれば良い.

$$\begin{aligned}
(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})\{\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}\} \\
&= (\boldsymbol{I} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1})\{\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}\} \\
&= (\boldsymbol{I} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1})\{\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\} \\
&= \boldsymbol{I} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} - \boldsymbol{U}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} \\
&\quad - \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} \\
&= \boldsymbol{I} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} - \boldsymbol{U}(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} \\
&= \boldsymbol{I} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} - \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}^{-1} = \boldsymbol{I}
\end{aligned}$$

□

予測分布の計算 (4)

$\det \hat{\boldsymbol{\Sigma}}$ と $\det \tilde{\boldsymbol{\Sigma}}$ の関係を導く為に以下のようなブロック行列を考える.

$$\boldsymbol{H} \triangleq \begin{pmatrix} \sigma_y^2 & -\boldsymbol{\phi}^\top \\ \boldsymbol{\phi} & \hat{\boldsymbol{\Sigma}}^{-1} \end{pmatrix} \tag{t}$$

この行列の行列式は命題 1 により

$$\det \boldsymbol{H} = \det \hat{\boldsymbol{\Sigma}}^{-1} \det(\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}) \tag{u}$$

次に $\boldsymbol{H}$ に対して右から以下のような行列をかける.

$$\boldsymbol{H} \begin{pmatrix} 1 & \sigma_y^{-2}\boldsymbol{\phi}^\top \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\phi}^\top - \boldsymbol{\phi}^\top \\ \boldsymbol{\phi} & \boldsymbol{\phi}\sigma_y^{-2}\boldsymbol{\phi}^\top + \hat{\boldsymbol{\Sigma}}^{-1} \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \mathbf{0}^\top \\ \boldsymbol{\phi} & \tilde{\boldsymbol{\Sigma}}^{-1} \end{pmatrix} \tag{v}$$

予測分布の計算 (5)

式 (v) の両辺の行列式をとると,

$$\det \boldsymbol{H} = \det \sigma_y^2 \det \tilde{\boldsymbol{\Sigma}}^{-1} \tag{w}$$

従って式 (u) と (w) から

$$\frac{\det \hat{\boldsymbol{\Sigma}}^{-1}}{\sigma_y^2} = \frac{\det \tilde{\boldsymbol{\Sigma}}^{-1}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \tag{x}$$

逆行列の行列式は元の行列の行列式の逆数で与えられるので,

$$\frac{\det \tilde{\boldsymbol{\Sigma}}}{\det \hat{\boldsymbol{\Sigma}}} = \frac{\sigma_y^2}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \tag{y}$$

が得られる. なお $\hat{\boldsymbol{\Sigma}}$ は正定値行列であるので (x) と (y) の分母は 0 ではない (命題 5).

予測分布の計算 (6)

次に被指数函数部分を y_* について整理する．まず式 (o) と Woodbury の公式 (命題 2) から

$$\begin{aligned}\tilde{\mathbf{\Sigma}} &= \left(\hat{\mathbf{\Sigma}}^{-1} + \boldsymbol{\phi}\sigma_y^{-2}\boldsymbol{\phi}^\top\right)^{-1} \\ &= \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\boldsymbol{\phi}(\sigma_y^2 + \boldsymbol{\phi}^\top\hat{\mathbf{\Sigma}}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^\top\hat{\mathbf{\Sigma}} \\ &= \hat{\mathbf{\Sigma}} - \frac{1}{\sigma_y^2 + \boldsymbol{\phi}^\top\hat{\mathbf{\Sigma}}\boldsymbol{\phi}}\hat{\mathbf{\Sigma}}\boldsymbol{\phi}\boldsymbol{\phi}^\top\hat{\mathbf{\Sigma}} \qquad \text{(z)}\end{aligned}$$

である．

予測分布の計算 (7)

それでは $-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} + D$ を計算する．まず (o),(p) をそのまま代入すると

$$-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} + D = -(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi})^\top \tilde{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi}) + \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_*^2 \tag{aa}$$

である．次に (z) を代入すると

$$\begin{aligned}
&- \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} + D \\
&= -(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi})^\top \left(\hat{\boldsymbol{\Sigma}} - \frac{\hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \right) (\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi}) + \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_*^2 \\
&= -(\hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} + \sigma_y^{-2} y_* \boldsymbol{\phi}^\top) \left(\hat{\boldsymbol{\Sigma}} - \frac{\hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \right) (\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi}) + \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_*^2 \\
&= -\left(\hat{\boldsymbol{\mu}}^\top - \frac{\hat{\boldsymbol{\mu}}^\top \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} + \sigma_y^{-2} y_* \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} - \sigma_y^{-2} y_* \frac{\boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \right) (\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_* \boldsymbol{\phi}) \\
&+ \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + \sigma_y^{-2} y_*^2 \\
&= -\left(\sigma_y^{-2} y_* \hat{\boldsymbol{\mu}}^\top \boldsymbol{\phi} - \frac{\hat{\boldsymbol{\mu}}^\top \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\mu}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} - \sigma_y^{-2} y_* \frac{\hat{\boldsymbol{\mu}}^\top \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} + \sigma_y^{-2} y_* \boldsymbol{\phi}^\top \hat{\boldsymbol{\mu}} \right. \\
&\left. + \sigma_y^{-4} y_*^2 \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} - \sigma_y^{-2} y_* \frac{\boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\mu}}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} - \sigma_y^{-4} y_*^2 \frac{\boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}}{\sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi}} \right) + \sigma_y^{-2} y_*^2
\end{aligned} \tag{ab}$$

予測分布の計算 (8)

式 (ab) について

$$\mu_* = \hat{\boldsymbol{\mu}}^\top \boldsymbol{\phi} \tag{3.76}$$

$$\sigma_*^2 = \sigma_y^2 + \boldsymbol{\phi}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \tag{3.77}$$

とおくと，

$$\begin{aligned}
-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} + D = &-\left\{ \sigma_y^{-2} y_* \mu_* - \frac{\mu_*^2}{\sigma_*^2} - \sigma_y^{-2} y_* \frac{\mu_*(\sigma_*^2 - \sigma_y^2)}{\sigma_*^2} + \sigma_y^{-2} y_* \mu_* + \sigma_y^{-4} y_*^2 (\sigma_*^2 - \sigma_y^2) \right. \\
&\left. -\sigma_y^{-2} y_* \frac{(\sigma_*^2 - \sigma_y^2)\mu_*}{\sigma_*^2} - \sigma_y^{-4} y_*^2 \frac{(\sigma_*^2 - \sigma_y^2)^2}{\sigma_*^2} \right\} + \sigma_y^{-2} y_*^2
\end{aligned} \tag{ae}$$

となる．あとは y_* の次数ごとでまとめると，

$$\begin{aligned}
-\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} + D &= \frac{1}{\sigma_*^2} \left[\mu_*^2 - 2\left\{ -\sigma_y^{-2} \mu_* (\sigma_*^2 - \sigma_y^2) + \sigma_y^{-2} \mu_* \sigma_*^2 \right\} y_* \right. \\
&\quad \left. + \left\{ -\sigma_y^{-4} \sigma_*^2 (\sigma_*^2 - \sigma_y^2) + \sigma_y^{-4} (\sigma_*^2 - \sigma_y^2)^2 + \sigma_*^2 \sigma_y^{-2} \right\} y_*^2 \right] \\
&= \frac{1}{\sigma_*^2} (\mu_*^2 - 2\mu_* y_* + y_*^2) = \frac{1}{\sigma_*^2} (y_* - \mu_*)^2
\end{aligned} \tag{af}$$

よって式 (y) と (af) のよって予測分布 (r) は改めて

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{\sqrt{2\pi\sigma_*^2}} \exp\left\{-\frac{(y_* - \mu_*)^2}{2\sigma_*^2}\right\} \tag{3.75}$$

と表せる. すなわち $y_*|\boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X} \sim \mathcal{N}(\mu_*, \sigma_*^2)$ である.

予測実験の設定

予測分布がどれくらい的を射ているかを視覚的に捉えやすいよういくつか実験を行う. 実験の共通設定は以下である.

- 真の分布 $y = f(x) + \epsilon$ に於いて $\epsilon \sim \mathcal{N}(0, 0.01)$
- パラメータの次元 $H_1 = 4$
- 特徴量函数 $\boldsymbol{\phi}(x) = (x^3, x^2, x, 1)^\top$
- 事前分布 $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
- 標本の数は 10 個,x は指定区間 $[a, b]$ からの一様分布 $U([a, b])$ に従う.

グラフの見方は以下のようである.

- 青い実線は真の分布の平均すなわち $y = f(x)$.
- 赤い破線はデータからの予測分布の平均すなわち $y = \mu_*(x)$.
- 緑の実線 (5 本) は事後分布からの母数標本を用いた回帰曲線すなわち $y = \boldsymbol{w}^\top \boldsymbol{\phi}(x)$.
- 薄水色の領域は予測平均からの差が予測標準偏差以下になる領域すなわち $\{\mu_*(x) - \sigma_*(x) \leq y \leq \mu_*(x) + \sigma_*(x)\}$.

予測の精度の指標:KL 情報量の計算

この実験では真の分布も予測分布も正規分布に従っており真のパラメータも既知なので, 直接それらの KL 情報量を計算することができる.
$p \sim \mathcal{N}(\mu, \sigma^2), q \sim \mathcal{N}(\nu, \tau^2)$ とする.

$$\log q = -\frac{1}{2}\log 2\pi - \log\tau - \frac{(x-\nu)^2}{2\tau^2} \tag{KL1}$$

であるから

$$\begin{aligned}
\int p \log q &= \int \left(-\frac{1}{2}\log 2\pi - \log\tau - \frac{(x-\nu)^2}{2\tau^2}\right) p \\
&= -\frac{1}{2}\log 2\pi - \log\tau - \frac{1}{2\tau^2}\int (x^2 - 2x\nu + \nu^2)p \\
&= \frac{1}{2}\log 2\pi - \log\tau - \frac{(\sigma^2+\mu^2) - 2\nu\mu + \nu^2}{2\tau^2} \\
&= \frac{1}{2}\log 2\pi - \log\tau - \frac{\sigma^2 + (\mu-\nu)^2}{2\tau^2} \qquad \text{(KL2)}
\end{aligned}$$

予測の精度の指標:KL 情報量の計算 (2)

(KL2) に於いて q を p に置き換えれば $\int p \log p = -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}$ を得る. 従って p からみた q の KL 情報量 $D_{KL}(p\|q) \triangleq \mathbb{E}_p[\log(p/q)]$ は,

$$
\begin{aligned}
D_{KL}(p\|q) &= \int p\log p - \int p \log q = -\log\sigma + \log\tau - \frac{1}{2} + \frac{\sigma^2 + (\mu-\nu)^2}{2\tau^2} \\
&= \log\frac{\tau}{\sigma} + \frac{(\sigma^2-\tau^2) + (\mu-\nu)^2}{2\tau^2} \qquad \text{(KL3)} \\
&\geq (1 - \frac{\sigma}{\tau}) + \frac{\sigma^2-\tau^2}{2\tau^2} + \frac{(\mu-\nu)^2}{2\tau^2} = \frac{(\sigma-\tau)^2 + (\mu-\nu)^2}{2\tau^2} \geq 0
\end{aligned}
$$

で求まる. 但し最後段左の不等式は任意の $x > 0$ に対して $\log x \leq x - 1$ であることから同値変形の $\log\frac{1}{x} \geq 1 - x$ を利用した. そのため等号成立は $\sigma = \tau$ である. また $D_{KL}(p\|q) = 0$ となる条件は更に $\mu = \nu$ が成立することである. これは二つの分布 p, q が一致することに他ならない.

予測の精度の指標:KL 情報量の計算 (3)

いま真の分布は $\boldsymbol{y} \mid x \sim \mathcal{N}(f(x), 0.01)$ で予測分布は $\boldsymbol{y} \mid x \sim \mathcal{N}(\mu_*, \sigma_*^2)$ である. また $x \sim U([a, b])$ である. 真の分布を $p(y|x)$, 予測分布を $q(y|x)$ と書くと,(KL3) を用いて

$$
\begin{aligned}
D_{KL}(p\|q) &= \iint p(x)p(y|x)\log\frac{p(y|x)}{q(y|x)}dxdy \\
&= \int_a^b p(x)\left(\int p(y|x)\log\frac{p(y|x)}{q(y|x)}dy\right)dx \\
&= \frac{1}{b-a}\int_a^b \log\frac{\sigma_*}{0.1} + \frac{0.01-\sigma_*^2+(f(x)-\mu_*)^2}{2\sigma_*^2}dx \\
&= \frac{1}{b-a}\int_a^b \log\frac{\sqrt{0.01+\boldsymbol{\phi}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\phi}}}{0.1} + \frac{-\boldsymbol{\phi}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\phi}+(f(x)-\hat{\boldsymbol{\mu}}^\top\boldsymbol{\phi})^2}{2(0.01+\boldsymbol{\phi}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\phi})}dx
\end{aligned} \tag{KL4}
$$

この実験ではこの値 (KL4) を Simpson 公式 [3] により数値的に求めた.

予測実験 1

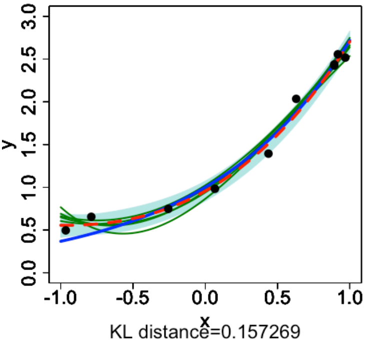

Figure 2: 真の分布の平均 (青い実線) は $y = e^x$ で $x \sim U([-1, 1])$. KL 情報量の標本平均は 0.274623

予測実験 2

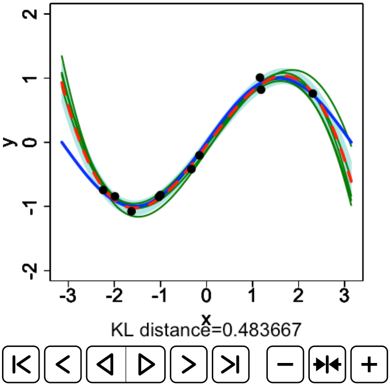

Figure 3: 真の分布の平均 (青い実線) は $y = \sin x$ で $x \sim U([-\pi, \pi])$. KL 情報量の標本平均は 0.5016577

予測実験 3

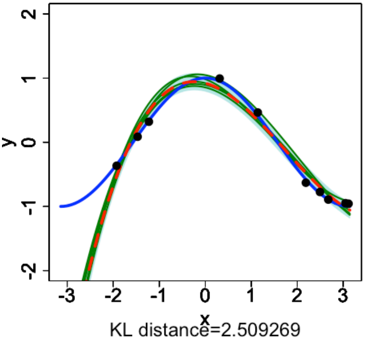

Figure 4: 真の分布の平均 (青い実線) は $y = \cos x$ で $x \sim U([-\pi, \pi])$
KL 情報量の標本平均は 2.030117

実験の考察

3 つの実験の結果から以下のことが分かった.

- KL 情報量の観点から精度の高い順に $\exp > \sin \gg \cos$ である.
- 薄水色の領域が広い $\simeq$ 分散が大きいことよりも, 青色線と赤色破線の位置の違い $\simeq$ 平均の差が大きいことの方が KL 情報量を大きくする.
- 最後の $\cos$ の推定が特に悪いのは, 指定区間 $[-\pi, \pi]$ では $\cos$ が 4 次函数で近似できるのに対し, 特徴量函数が 3 次までしか表現力できないことが影響していると考えられる.

最尤推定量の計算

最尤法との比較をする為まずは最尤推定量 $\hat{\boldsymbol{w}}^{ML}$ の導出を行う.
尤度函数は $p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})=\prod_{i=1}^{N}p(y_i|\boldsymbol{x}_i,\boldsymbol{w})$ である. 対数尤度函数を書き下すと,

$$\begin{aligned}\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) &= \sum_{i=1}^{N}\left\{-\frac{1}{2}\log 2\pi\sigma_y^2-\frac{(y_i-\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma_y^2}\right\} \\ &= -\frac{\sum_{i=1}^{N}(y_i-\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma_y^2}-\frac{N}{2}\log 2\pi\sigma_y^2 \qquad \text{(ag)}\end{aligned}$$

となるから尤度方程式 $\nabla_{\boldsymbol{w}}\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})=\boldsymbol{0}$ は

$$\nabla_{\boldsymbol{w}}\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})=-\frac{1}{\sigma_y^2}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_i))(-\boldsymbol{\phi}(\boldsymbol{x}_i))=\boldsymbol{0}$$

$$\iff \sum_{i=1}^{N}y_i\boldsymbol{\phi}(\boldsymbol{x}_i)=\sum_{i=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^\top\boldsymbol{w} \qquad \text{(ah)}$$

従って行列 $\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^\top$ が正則であれば式 (ah) は唯一つの解を持ちそれが最尤推定量となるから

$$\hat{\boldsymbol{w}}^{ML} = \left(\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^\top\right)^{-1} \left(\sum_{i=1}^{N} y_i \boldsymbol{\phi}(\boldsymbol{x}_i)\right) \qquad \text{(ai)}$$

となり求まる. 逆行列が存在する為の必要十分条件は以下で与えられる.

Theorem 1

$\mathbb{R}^n$ 上の N 個のベクトル $\{\boldsymbol{x}_i\}_{i=1}^{N}$ について以下の三条件は同値である. (証明は補足を参照)

1. $\mathbb{R}^n$ 上の基底を少なくとも一つ取れる
2. $\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正定値行列
3. $\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正則

最尤推定との比較実験の設定

ベイズ法と比較されることが多い最尤法との比較を実験的に行う. 実験の共通設定は以下である.

- 真の分布 $y = f(x) + \epsilon$ に於いて ϵ は平均 0 の正規分布に従う.
- 事前分布 $p(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. 但し次元が異なることがある.
- データセットは各実験で 3 セット,x は指定区間からの一様分布に従う.
- 真の分布はモデル函数で実現可能である. すなわちある $\boldsymbol{w}$ があって $f(x) = \boldsymbol{w}^\top \boldsymbol{\phi}(x)$ となる.

グラフの見方は以下のようである.

- 青い実線は真の分布の平均すなわち $y = f(x)$.
- 最左図は最尤法で, 赤い実線は最尤推定量の回帰線.
- 中央図はベイズ法で, 赤い実線は予測分布の平均すなわち $y = \mu_*(x)$.
- 中央図の薄水色の領域は予測平均から ± 予測標準偏差内になる領域すなわち $\{\mu_*(x) - \sigma_*(x) \le y \le \mu_*(x) + \sigma_*(x)\}$.
- 最右図はベイズ法で, 緑の実線 (10 本) は事後分布からの標本母数を用いた回帰曲線.

最尤推定との比較実験 1

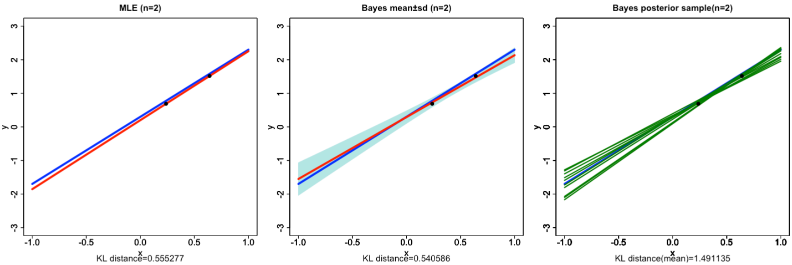

Figure 5: 真の分布の平均 (青い実線) は $y = 2x + 0.3$ の推定の比較. 真の分布の分散は 0.01 で $x \sim U([-1, 1])$ パラメータの数は $H_1 = 2$

データセット	MLE	Bayes	sample mean
1	0.03768442	0.03632561	0.07217360
2	0.04003825	0.03481267	0.06624334
3	0.04582605	0.04343713	0.08143445

Table 1: データセットごとにおける KL 情報量

最尤推定との比較実験 2

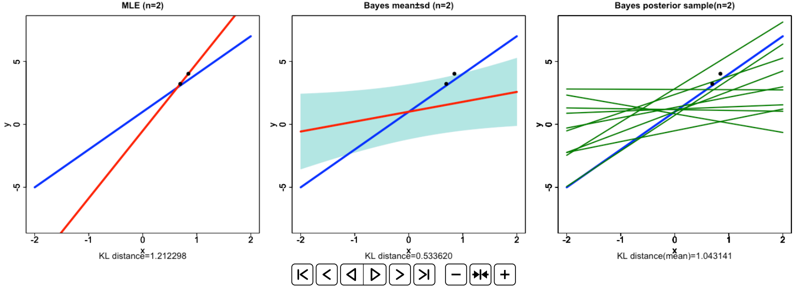

Figure 6: 真の分布の平均 (青い実線) は $y = 3x + 1$ の推定の比較. 真の分布の分散は 4 で $x \sim U([-2, 2])$ パラメータの数は $H_1 = 2$

データセット	MLE	Bayes	sample mean
1	0.04752186	0.02624517	0.06071060
2	0.02620013	0.05093969	0.09666167
3	0.02103184	0.06136713	0.12243690

Table 2: データセットごとにおける KL 情報量

最尤推定との比較実験 3

ベイズ深層学習

joeyoji

はじめに
ベイズ線形回帰
モデル
学習と予測
周辺尤度
逐次学習
能動学習への応用
ガウス過程との関係
最尤推定, MAP 推定との関係
最尤推定と誤差最小化
MAP 推定と正則化
分類モデルに対する誤差関数
補足
事後分布の存在
逆行列の正定値性
定理 1 の証明
正規分布の積分計算
参考文献

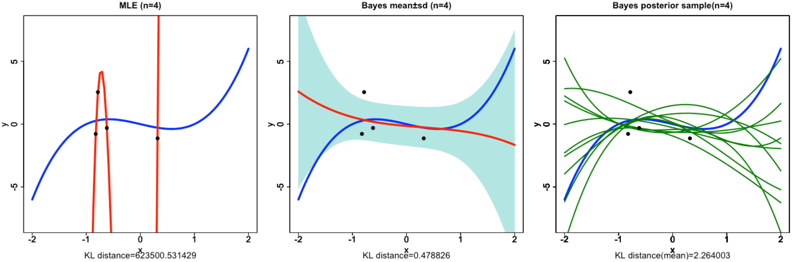

Figure 7: 真の分布の平均 (青い実線) は $y = x^3 - x$ の推定の比較. 真の分布の分散は 2 で $x \sim U([-2, 2])$ パラメータの数は $H_1 = 4$

データセット	MLE	Bayes	sample mean
1	0.12557937	0.09042780	0.13442643
2	0.01278180	0.01426971	0.05868668
3	0.02996302	0.03313227	0.06081361

Table 3: データセットごとにおける KL 情報量

実験結果の考察

- Bayes 予測分布に着目すると, データ数が増えるごとに予測の不確かさ $\simeq \phi^\top \hat{\Sigma}\phi$ が小さくなっていることが目に見える.
- KL 情報量の観点からすれば, 一概に MLE の方が優秀,Bayes の方が優秀というような結論は導けない.
- データ数が少ないとき,MLE は少ないデータにのみに依存して決まるので, データの出方 (特に分散が大きい場合) によってはかなり的外れになる.
- 一方,Bayes 予測の場合は安定した結果にはなるが, これはデータ数が少ないときには事前分布の影響が大きい為であるので, 安定しているから予測が当たっているということにはならない.

周辺尤度の定義

Definition 3

データ $\boldsymbol{X} \times \boldsymbol{Y} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ が与えられたもとで, 同時分布 *(3.66)* からパラメータ $\boldsymbol{w}$ を積分除去した値, あるいは事後分布の定義式 *(3.69)* の分母に当たる正規化定数

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} \tag{aj}$$

をベイズ線形回帰の周辺尤度 *(marginal likehood)* 或いはエビデンス *(evidence)* という.

周辺尤度の計算

ここでは周辺尤度の具体的な函数形は既に (i) として求めており, 変形して再掲すると,

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \exp\left\{-\frac{1}{2}\left(-\log\det\hat{\boldsymbol{\Sigma}} + N\log\sigma_y^2 + N\log 2\pi + H_1\log\sigma_w^2 - \hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + \sigma_y^{-2}\sum_{i=1}^N y_i^2\right)\right\} \tag{3.78}$$

である. 一般には周辺尤度は解析的に求めるのは困難だったり, 式を解析的に表せたとしても具体的に数値を求めるのに膨大な時間がかかることがある.

周辺尤度の意味

定義 (aj) を見直すと, 周辺尤度は確率モデル函数 $p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})$ と事前分布 $p(\boldsymbol{w})$ の与え方によって変わる, データの (条件付き) 確率密度函数である. このことから周辺尤度は, モデル $p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})$ と事前分布 $p(\boldsymbol{w})$ の組がどのくらいデータを説明するのに相応しいか表している量であると考えることができる [2].

ベイズ因子

Definition 4

二つの異なるモデルと事前分布の組 $M_1 = (p(x|w), p(w)), M_2 = (q(x|w'), q(w'))$ に対してそれらの周辺尤度の比

$$B = \frac{MLH(M_1)}{MLH(M_2)} = \frac{\int p(X|w)p(w)dw}{\int q(X|w')q(w')dw'}$$

をベイズ因子 (Bayes factor) という [5].

$B > 1$ であれば M_2 より M_1 の方が, $B < 1$ であれば M_1 より M_2 の方が与えられたデータセット X を説明するのが相応しいと考えられる.

モデルの比較実験の設定

ここでは実際に周辺尤度を基準として用いたモデル選択を行う. 実験の共通設定は以下である.

- 取り扱うモデルは全て線形回帰モデルとし, それぞれ特徴量函数が異なるものとする. 特徴量函数は 1 次から 5 次までの多項式のものである.
- 事前分布は $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ とする. 但しモデルごとに次元が違う.
- 真の分布 $y = f(x) + \epsilon$ に於いて $\epsilon \sim \mathcal{N}(0, 0.01)$ とした.
- x は指定区間からの一様分布に従う.
- 各実験ではデータ数を増やして変遷の過程も追うが, データセットは上乗せではなく, データ数を増やすごとに全て取り替えている.

グラフの見方

グラフの見方は以下のようである.

- それぞれ色の異なる折れ線は 1 つのデータセットに対する周辺尤度五つ組を表す.
- 各コマごとに折れ線は 10 本ある, つまり異なるデータセットが 10 個ある.
- 横軸は特徴量函数の次元であり, 縦軸は各データセットに対する周辺尤度の比を表す. 但し周辺尤度が最大のものが 1 となるように正規化している.
- N は各データセットのデータ数を表す.di はあるデータセットに対する周辺尤度が次数 i で最大になったとき 1 加算される.$d1 + d2 + d3 + d4 + d5 = 10$ である.

モデルの比較実験 (1)

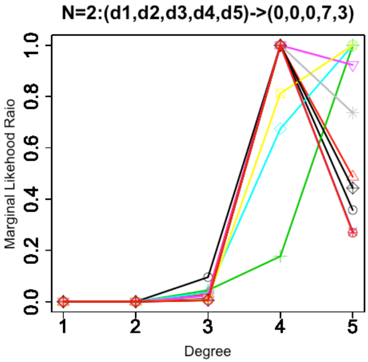

Figure 8: 真の分布の平均は $f(x) = (4x^2 - 1)x$, $x \sim U(-3, 3)$

モデルの比較実験 (2)

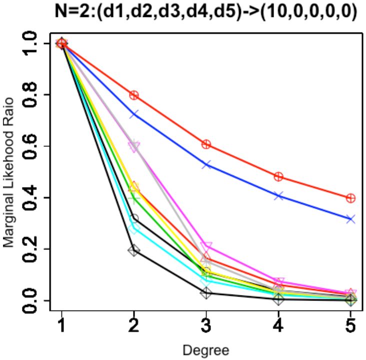

Figure 9: 真の分布の平均は $f(x) = \cos x$, $x \sim U(-\pi, \pi)$

モデルの比較実験 (3)

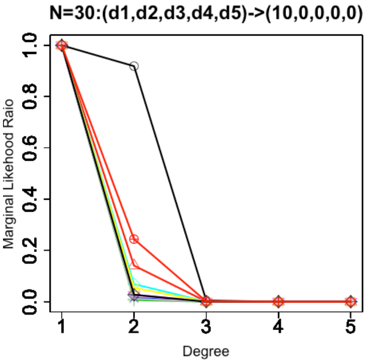

Figure 10: 真の分布の平均は $f(x) = 0.01x^2 + 2x + 3$, $x \sim U(-3, 3)$

モデルの比較実験 (4)

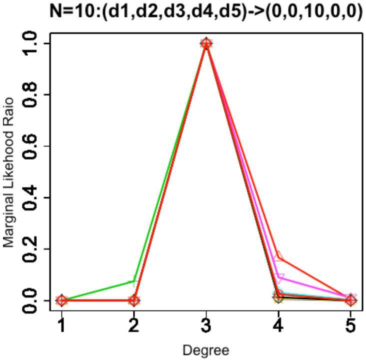

Figure 11: 真の分布の平均は $f(x) = \sin x$, $x \sim U(-\pi, \pi)$

実験結果の考察

4 つの実験の結果から以下のことが分かった.

- モデルの次数が高いからといって, 低次のモデルより必ず勝るわけではない.
- 誤差程度に微小ながらも構造があれば, データ数を十分増やすことでその構造が捉えられる.
- 逆に本当の誤差ならば, データ数を十分増やすことで情報が相殺される.

逐次学習

ベイズ推論によるモデルの学習ではあるデータセットに対して学習して得た事後分布を, 別のデータセットを学習する為の事前分布として用いて学習することができる. これを繰り返し事後分布 (或いは事前分布) を更新していく学習方法を逐次学習 (sequential learning) 或いはオンライン学習 (online learning) という.

Proposition 3

尤度函数に対して共役事前分布を用いた学習を考える. データは (時系列に関して) 独立に生成されるとする. 同じデータセットに対して, それらを分割して逐次学習をさせ最終的に得られる事後分布と, 一度に全て学習させて得られる事後分布は一致する.

共役事前分布を用いた学習

Proof.

モデル函数を $p(x|w)$ とし, それに対する共役事前分布を $p(w|\lambda)$ とする. 共役事前分布は事前分布と事後分布の函数形が一致し超母数 λ のみが異なり得る. よって逐次学習において $i \in \{1, 2, \ldots, m\}$ 回目の学習で得られる事後分布の超母数を λ_i とおき, 初期事前分布の超母数を λ_0 とおく. また一括学習の事後分布の超母数を λ_* とおく. 更に全データセット X に対し, それらを m 分割した部分集合を X_i とおく [1]. このとき

$$
\begin{aligned}
p(w|\lambda_m) &\propto p(w|\lambda_{m-1})p(X_m|w) \propto p(w|\lambda_{m-2})p(X_{m-1}|w)p(X_m|w) \propto \cdots \\
&\propto p(w|\lambda_0)\prod_{i=1}^{m} p(X_i|w) \propto p(w|\lambda_0)\prod_{i=1}^{m}\prod_{x \in X_i} p(x|w) \\
&\propto p(w|\lambda_0)\prod_{x \in X} p(x|w) \propto p(w|\lambda_0)p(X|w) \propto p(w|\lambda_*)
\end{aligned}
$$

よって $p(w|\lambda_m) \propto p(w|\lambda_*)$ となり両辺はともに確率密度函数或いは確率質量函数なので一致する. □

[1] 任意の i に対して $X_i \neq \emptyset$ であり, 異なる j, k に対して $X_j \cap X_k = \emptyset$ であり, かつ $\bigcup_{i=1}^{m} X_i = X$

ベイズ線形回帰における逐次学習

先まで扱っていたベイズ線形回帰について逐次学習を考える. 今回はモデル函数に対して共役事前分布を用いており, データは独立に生成されることを仮定していたので命題 3 が成り立つ. 証明にも記述した通り, 超母数の変遷を辿れば学習過程を追うことができる. 簡単の為, 一回の学習につき一個のデータを読み込ませることを考える.
i 回目の学習の結果得られる超母数を $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ で表す. 初期事前分布の超母数は定義 (3.68) により $(\mathbf{0}, \sigma_w^2 \boldsymbol{I})$ である. はじめ求めた事後分布や予測分布と同様に, 指数の中身を $\boldsymbol{w}$ について整理すればよい.

ベイズ線形回帰における逐次学習 (2)

$i-1$ 回目の学習で得られる事後分布の対数をとると,

$$\log p(\boldsymbol{w}|\boldsymbol{\mu}_{i-1},\boldsymbol{\Sigma}_{i-1}) = -\frac{1}{2}\left\{\boldsymbol{w}^\top\boldsymbol{\Sigma}_{i-1}^{-1}\boldsymbol{w} - 2(\boldsymbol{\Sigma}_{i-1}^{-1}\boldsymbol{\mu}_{i-1})^\top\boldsymbol{w}\right\} + \mathrm{Const.} \tag{ak}$$

i 番目に学習させるデータの尤度函数の対数をとると,

$$\log p(y_i|\boldsymbol{x}_i,\boldsymbol{w}) = -\frac{1}{2}\left\{\boldsymbol{w}^\top\left(\sigma_y^{-2}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right)\boldsymbol{w} - 2\left(\sigma_y^{-2}y_i\boldsymbol{\phi}_i\right)^\top\boldsymbol{w}\right\} + \mathrm{Const.} \tag{al}$$

これと式 (c) により,i 回目の学習で得られる事後分布の超母数は,

$$\begin{cases} \boldsymbol{\Sigma}_i^{-1} &= \boldsymbol{\Sigma}_{i-1}^{-1} + \sigma_y^{-2}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ \boldsymbol{\mu}_i &= \boldsymbol{\Sigma}_i\left(\boldsymbol{\Sigma}_{i-1}^{-1}\boldsymbol{\mu}_{i-1} + \sigma_y^{-2}y_i\boldsymbol{\phi}_i\right) \end{cases} \tag{am}$$

となり, 超母数の漸化式が構成できる. $\boldsymbol{\Sigma}_0^{-1} = \sigma_w^{-2}\boldsymbol{I}$ は正定値行列なので帰納的に全ての $\boldsymbol{\Sigma}_i^{-1}$ は正定値故に逆行列 $\boldsymbol{\Sigma}_i$ も存在する.

逐次学習の利点, 注意点

利点 命題 3 が成り立つと考えられる場合, 逐次学習をすることで計算量を抑えられることができる. 新しいデータを取得した度に一括学習をすると学習の総計算量はデータ数を n として $O(n^2)$ となるが, 逐次学習であれば総計算量は $O(n)$ となる. また学習用のデータを蓄える必要性がなくなる.

注意点 ニューラルネットワークなどの複雑なモデルを用いて学習する場合, 逐次学習の各更新で事後分布が解析的に計算できなくなる. そこでモーメントマッチング (moment matching) と呼ばれる手法を用いて近似的に事後分布を更新したりする.

教師あり学習を行う上での現実的な問題点

線形回帰などの教師あり学習は様々な問題に適応でき汎用性が高い手法である. しかし現実的には, 入力データは大量に手に入るにも関わらず, 対応するラベルデータを収集するにはコストがかかる為, 満足に手に入らないということが多々起こる.

Example 1

画像認識では, ウェブ上を探せば画像データは大量に見つかるが, それに対応するラベルデータは同じ画像でも問題ごとに適切なものを与えないと意味がない. ラベルを与えるには (現状) タスクを理解した人手 (アノテーターという.) が画像ごとに適切にラベル付けする必要があり, これには大きなコストがかかる.

能動学習

以上のような問題を解決する為に, 少ない学習データで効率的に学習することを目指す必要がある.
ベイズ線形回帰では予測分布を計算することにより, 予測対称の不確実性を分散などの指標によって定量的に測れることが出来る.
そこでラベルのついてない入力集合から, 適切と思われる指標を用いて入力データ点を選択し, アノテーターにラベルを質問するという学習手法が考えられる. このような手法のことを**能動学習** (active learning) という.

能動学習のイメージ

Figure 12: 能動学習のアニメ
画像のキャラクタの出典はイラストや

能動学習で使われる指標

能動学習で使われる指標は様々なものが考えられる. 直感的には入力集合 $\boldsymbol{X}_{\text{pool}}$ の中で最も予測に自信がないものを選ぶのが効率が良さそうである. 例として分散最大選択やエントロピー最大選択などがある. エントロピーが最大となる入力 $\boldsymbol{x}_q$ は

$$\boldsymbol{x}_q = \operatorname*{arg\,max}_{\boldsymbol{x}_* \in \boldsymbol{X}_{\text{pool}}} \{-\mathbb{E}_{p^*}[\log p^*]\} \tag{3.81}$$

但し $p^* = p(y_*|\boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X})$ を表す. エントロピーは $-\mathbb{E}_{p^*}[\log p^*]$ である. 分布のエントロピーとは, 直感的には分布の曖昧さを表す指標である. 分布が一様に近ければエントロピーも大きい.

線形回帰モデルでの能動学習

先に求めた線形回帰モデルの予測分布のエントロピーを求める. 予測分布は正規分布であったので, 一般に正規分布のエントロピーを求める.$X \sim \mathcal{N}(\mu, \sigma^2)$ として

$$\begin{aligned}\mathbb{E}[\log p(X)] &= \int \left\{ -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} \right\} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} dx \\ &= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{\mathrm{Var}(X)}{2\sigma^2} = -\frac{1}{2}(1 + \log 2\pi\sigma^2) \qquad \text{(an)}\end{aligned}$$

従って予測分布 $p(y_* | \boldsymbol{x}_*, \boldsymbol{Y}, \boldsymbol{X})$ のエントロピーは式 (an) に代入して

$$-\mathbb{E}_{p^*}[\log p^*] = \frac{1}{2}(1 + \log 2\pi\sigma_*^2) \qquad (3.83)$$

よって予測分布が正規分布となる場合はエントロピー最大選択と分散最大選択は等価になる.

能動学習の実験の設定

具体的に能動学習のイメージを掴むために実験をいくつか行う．実験の共通設定は以下である．

- データ選択手法はエントロピー最大選択とする．
- 特徴量関数は $\boldsymbol{\phi}(x) = (x^3, x^2, x, 1)^\top$, 事前分布は $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ とする．
- 真の分布 $y = f(x) + \epsilon$ に置いて誤差 ϵ は平均 0 の正規分布に従う．
- x は指定区間からの一様分布に従う．

グラフの見方は以下のようである．

- 青実線は真の分布の平均 $y = f(x)$
- 赤実線は予測分布の平均 $y = \boldsymbol{\mu}_*^\top \boldsymbol{\phi}(x)$
- 薄水色の領域は予測平均から $\pm$ 予測標準偏差となる領域
- 黒丸印は既習データ, 罰印は未学習データ, 緑丸印は予測分布のエントロピーが最大となるデータ点である．

能動学習の実験 (1)

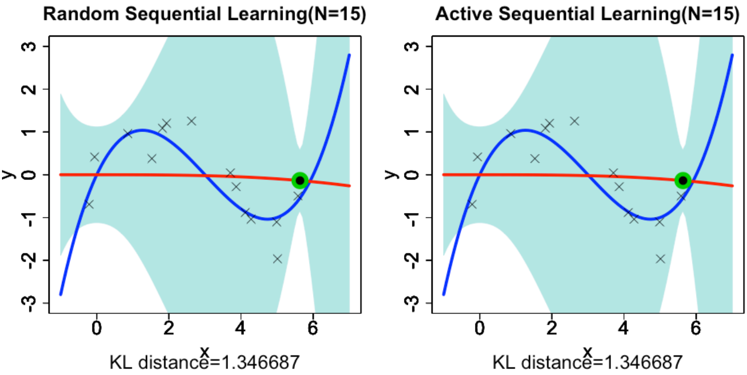

Figure 13: 真の分布の平均は $f(x) = 0.1x(x-3)(x-6)$
誤差の分散は 0.25 , また $x \sim U(-1, 7)$

能動学習の実験 (2)

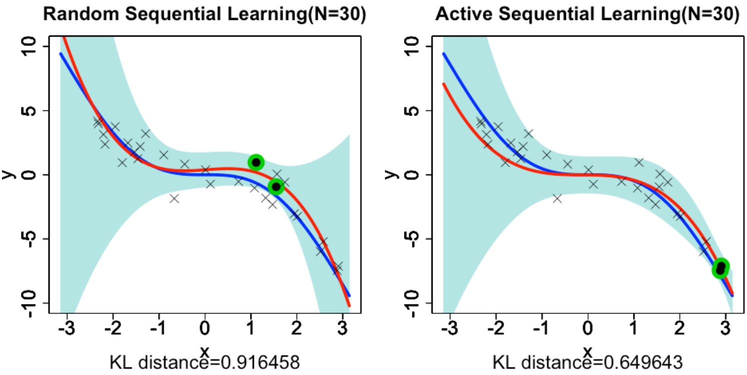

Figure 14: 真の分布の平均は $f(x) = 3\sin x - 3x$
誤差の分散は 1 , また $x \sim U(-\pi, \pi)$

能動学習の実験 (3)

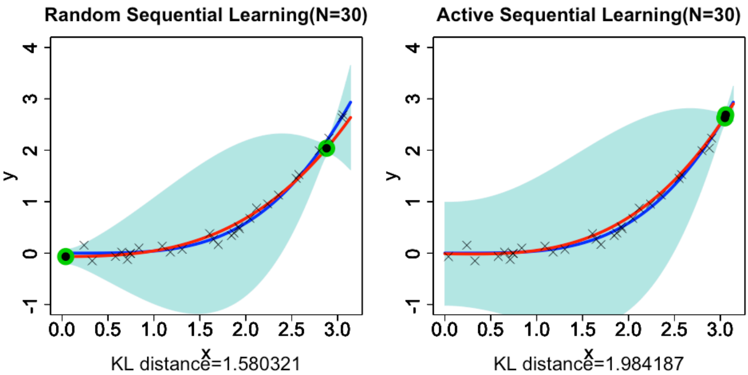

Figure 15: 真の分布の平均は $f(x) = -1 + 0.5x^2 + \cos x$
誤差の分散は 0.01 ，また $x \sim U(0, \pi)$

実験結果の考察

3 つの実験の結果から以下のことが分かった.

- データが独立に生成されるとき, 同じデータセットに対して最終的に得られる予測分布が一致することが確かめられた. (事後分布が一致するので予測分布も一致する.)
- ランダムにデータ点を学習するよりも, 戦略を使う方が KL 情報量の観点から見ても, 学習初期の段階でも誤差を抑えやすい傾向がある.
- 戦略を使ったからといって常にランダム選択に対して優位であるわけではない.

ベイズ最適化

予測の不確実性を利用した能動学習の枠組みは, 未知の函数の最大値の探索などにも利用される. 最適化の分野に於いて, このような方法はベイズ最適化 (Bayesian optimization) と呼ばれる. ベイズ最適化では, 予測対称に関して弱い仮定を設定できるガウス過程 (Gaussian process) を用いるのが一般的である.

事後分布のパラメータ (3.72),(3.73) は

$$\boldsymbol{\Phi} = \left[\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_N)\right],\ \boldsymbol{\Lambda} = \sigma_w^{-2}\boldsymbol{I}, \boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$ を用いて次

のように書き直せる.

$$\hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Lambda} + \sigma_y^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top \quad \text{(ao)}$$

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}}\sigma_y^{-2}\boldsymbol{\Phi}\boldsymbol{Y} \quad \text{(ap)}$$

これらをを式 (3.76),(3.77) に代入して, 予測分布のパラメータをより厳密に計算する.

予測分布のパラメータ計算 (2)

まず Woodbury の公式 (2) を用いて

$$\hat{\mathbf{\Sigma}} = (\mathbf{\Lambda} + \mathbf{\Phi}\sigma_y^{-2}\boldsymbol{I}\mathbf{\Phi}^\top)^{-1} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\mathbf{\Phi}^\top\mathbf{\Lambda}^{-1} \quad \text{(aq)}$$

但し,$\boldsymbol{K} = \mathbf{\Phi}^\top\mathbf{\Lambda}^{-1}\mathbf{\Phi}$ とした (グラム行列という). これを用いて

$$\begin{aligned}
\mu_* &= \boldsymbol{\phi}_*^\top\hat{\boldsymbol{\mu}} = \boldsymbol{\phi}_*^\top\hat{\mathbf{\Sigma}}\sigma_y^{-2}\mathbf{\Phi}\boldsymbol{Y} \\
&= \boldsymbol{\phi}_*^\top\left\{\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\mathbf{\Phi}^\top\mathbf{\Lambda}^{-1}\right\}\sigma_y^{-2}\mathbf{\Phi}\boldsymbol{Y} \\
&= \boldsymbol{\phi}_*^\top\mathbf{\Lambda}^{-1}\mathbf{\Phi}\sigma_y^{-2}\boldsymbol{Y} - \boldsymbol{\phi}_*^\top\mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{K}\sigma_y^{-2}\boldsymbol{Y} \\
&= \boldsymbol{\phi}_*^\top\mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\left\{(\sigma_y^2\boldsymbol{I} + \boldsymbol{K}) - \boldsymbol{K}\right\}\sigma_y^{-2}\boldsymbol{Y} \\
&= \boldsymbol{\phi}_*^\top\mathbf{\Lambda}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{Y} \qquad (3.86) \\
\sigma_*^2 &= \sigma_y^2 + \boldsymbol{\phi}_*^\top\hat{\mathbf{\Sigma}}\boldsymbol{\phi}_* = \sigma_y^2 + \boldsymbol{\phi}_*^\top\left\{\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\mathbf{\Phi}^\top\mathbf{\Lambda}^{-1}\right\}\boldsymbol{\phi}_* \\
&= \sigma_y^2 + \boldsymbol{\phi}_*^\top\mathbf{\Lambda}^{-1}\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top\mathbf{\Lambda}^{-1}\mathbf{\Phi}(\sigma_y^2\boldsymbol{I} + \boldsymbol{K})^{-1}\mathbf{\Phi}^\top\mathbf{\Lambda}^{-1}\boldsymbol{\phi}_* \qquad (3.87)
\end{aligned}$$

と表せる.

カーネル函数

式 (3.86),(3.87) について特徴量函数に注目して見てみると, 特徴量函数は以下のようにまとまっていることがわかる.

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}_i \boldsymbol{\Lambda}^{-1} \boldsymbol{\phi}_j \tag{3.88}$$

これをカーネル函数 (kernel function) 或いは共分散函数 (covariance function) という. 式 (3.86),(3.87) に含まれるカーネル函数を明示的にすると以下のように表せる.

$$\mu_* = \begin{bmatrix} k_{*1}, & \cdots & , k_{*N} \end{bmatrix} \left(\sigma_y^2 \boldsymbol{I} + \begin{bmatrix} k_{11} & \cdots & k_{1N} \\ \vdots & \ddots & \vdots \\ k_{N1} & \cdots & k_{NN} \end{bmatrix} \right)^{-1} \boldsymbol{Y} \tag{ar}$$

$$\sigma_*^2 = \sigma_y^2 + k_{**} - \begin{bmatrix} k_{*1}, & \cdots & , k_{*N} \end{bmatrix} \left(\sigma_y^2 \boldsymbol{I} + \begin{bmatrix} k_{11} & \cdots & k_{1N} \\ \vdots & \ddots & \vdots \\ k_{N1} & \cdots & k_{NN} \end{bmatrix} \right)^{-1} \begin{bmatrix} k_{1*} \\ \vdots \\ k_{N*} \end{bmatrix} \tag{as}$$

カーネルトリック

ガウス過程を用いることによって, カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}')$ を直接設計し回帰を行うことができる. この計算テクニックは**カーネルトリック**と呼ばれる.
カーネル関数は二つの入力点に関する相関を規定していると考えられ, ある意味異なるデータ間の類似度や近さを設計しているとも言える.

一般の回帰モデル

ここでは線形に限らない一般の回帰モデルのパラメータ推定に於いて, 最尤推定と最小二乗法による誤差最小化が等価であることを示す. 一般の回帰モデルは次のように表される.

$$y = f(\boldsymbol{x}; \boldsymbol{w}) + \epsilon \tag{3.90}$$

観測誤差 ϵ は平均 0 分散 σ_y^2 を持つ正規分布に従うとする.

$$\epsilon \sim \mathcal{N}(0, \sigma_y^2) \tag{3.91}$$

式 (3.90) 及び (3.91) から次が得られる.

$$y \mid \boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(f(\boldsymbol{x}; \boldsymbol{w}), \sigma_y^2) \tag{3.92}$$

最尤推定量の導出

データ $\boldsymbol{X}\times\boldsymbol{Y}=\{(\boldsymbol{x}_1,y_1),\dots,(\boldsymbol{x}_N,y_N)\}$ が独立に生成されたとすると，以上のモデルの尤度函数は，

$$p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})=\prod_{i=1}^{N}p(y_i|\boldsymbol{x}_i,\boldsymbol{w})=\frac{1}{\sqrt{2\pi\sigma_y^2}^N}\exp\left\{-\frac{\sum_{i=1}^{N}(y_i-f(\boldsymbol{x}_i;\boldsymbol{w}))^2}{2\sigma_y^2}\right\} \tag{3.93}$$

となる. 最尤推定量は

$$\hat{\boldsymbol{w}}^{ML}=\arg\max_{\boldsymbol{w}}\{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})\}=\arg\max_{\boldsymbol{w}}\{\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})\} \tag{3.94}$$

で与えられる. ここで対数尤度函数は

$$\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w})=-\frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\{y_i-f(\boldsymbol{x}_i;\boldsymbol{w})\}^2-\frac{N}{2}\log 2\pi\sigma_y^2 \tag{3.95}$$

従って以上をまとめると

$$\hat{\boldsymbol{w}}^{ML}=\arg\min_{\boldsymbol{w}}\left\{\sum_{i=1}^{N}(y_i-f(\boldsymbol{x}_i;\boldsymbol{w}))^2\right\} \tag{at}$$

誤差最小化パラメータの導出

一方, 最小二乗法による誤差最小化では誤差函数を計算し, それを最小にするパラメータを求める. 与えられたデータに関する誤差函数は,

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 \tag{au}$$

従って求める最適解は

$$\hat{\boldsymbol{w}}^{LS} = \underset{\boldsymbol{w}}{\arg\min}\left\{\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i;\boldsymbol{w}))^2\right\} \tag{av}$$

従って式 (at),(av) から $\hat{\boldsymbol{w}}^{ML} = \hat{\boldsymbol{w}}^{LS}$ であることが分かる. すなわち回帰モデルに於いて最尤法と最小二乗法は等価であることが示された.

MAP 推定量の定義

ベイズ法に於いて, 事後分布を最大にするパラメータを最大事後確率推定量 (maximum a posteriori estimator) 或いは英字の頭文字をとって MAP 推定量という.MAP 推定量は以下のように与えられる.

$$\hat{\boldsymbol{w}}^{MAP} = \arg\max_{\boldsymbol{w}}\{p(\boldsymbol{w}|\boldsymbol{Y},\boldsymbol{X})\} = \arg\max_{\boldsymbol{w}}\{\log p(\boldsymbol{w}|\boldsymbol{Y},\boldsymbol{X})\} \tag{3.99}$$

但し以下では尤度函数は回帰モデルのもの (3.93) を考える.

事前分布が正規分布のときの MAP 推定

事前分布として (3.68) を用意したときの MAP 推定量を考える．対数事後分布は

$$\begin{aligned}\log p(\boldsymbol{w}|\boldsymbol{Y},\boldsymbol{X}) &= \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) + \log p(\boldsymbol{w}) - p(\boldsymbol{Y}|\boldsymbol{X}) \\ &= -\frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 - \frac{N}{2}\log 2\pi\sigma_y^2 \\ &\quad - \frac{1}{2\sigma_w^2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{1}{2}\log 2\pi\sigma_w^2 - p(\boldsymbol{Y}|\boldsymbol{X}) \quad (3.100)\end{aligned}$$

従って (3.99),(3.100) から

$$\hat{\boldsymbol{w}}^{MAP} = \arg\min_{\boldsymbol{w}}\left\{\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 + \frac{\sigma_y^2}{\sigma_w^2}\boldsymbol{w}^\top\boldsymbol{w}\right\} \quad \text{(aw)}$$

を得る．

L2 正則化による最適パラメータの導出

一方, 過剰適合を防ぐための L2 正則化を施した新たなコスト関数は,

$$J(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda\Omega_{L2}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 + \frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} \tag{ax}$$

但し $\lambda > 0$ である. 以上によりコスト関数を最小化するパラメータは

$$\hat{\boldsymbol{w}}^{L2} = \arg\min_{\boldsymbol{w}}\left\{\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 + \lambda\boldsymbol{w}^\top\boldsymbol{w}\right\} \tag{ay}$$

従って $\lambda = \frac{\sigma_y^2}{\sigma_w^2}$ とすれば (aw),(ay) により $\hat{\boldsymbol{w}}^{MAP} = \hat{\boldsymbol{w}}^{L2}$ となることが分かる.

事前分布がラプラス分布のときの MAP 推定

次に事前分布として以下で定義される平均 0 の Laplace 分布を用意した場合を考える.

$$Lap(\boldsymbol{w}|\boldsymbol{0}, b) = \frac{1}{(2b)^{H_1}} \exp\left(-\frac{\|\boldsymbol{w}\|_1}{b}\right) \tag{3.101}$$

但し $b > 0, \|\boldsymbol{w}\|_1 = |w_1| + \cdots + |w_{H_1}|$ である. このとき対数事後分布は

$$\begin{aligned}\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}) &= -\frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i; \boldsymbol{w})\}^2 - \frac{N}{2}\log 2\pi\sigma_y^2 \\ &\quad - \frac{\|\boldsymbol{w}\|_1}{b} - H_1 \log 2b - p(\boldsymbol{Y}|\boldsymbol{X})\end{aligned} \tag{az}$$

であるから式 (3.99),(az) により

$$\hat{\boldsymbol{w}}^{MAP} = \arg\min_{\boldsymbol{w}}\left\{\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i; \boldsymbol{w})\}^2 + \frac{\sigma_y^2}{b}\|\boldsymbol{w}\|_1\right\} \tag{ba}$$

L1 正則化による最適パラメータの導出

一方,L1 正則化を施した新たなコスト関数は,

$$J(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda\Omega_{L1}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 + \lambda\|\boldsymbol{w}\|_1 \tag{bb}$$

但し $\lambda > 0$ である. 以上によりコスト関数を最小化するパラメータは

$$\hat{\boldsymbol{w}}^{L1} = \underset{\boldsymbol{w}}{\arg\min}\left\{\sum_{i=1}^{N}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{w})\}^2 + 2\lambda\|\boldsymbol{w}\|_1\right\} \tag{bc}$$

従って $\lambda = \frac{\sigma_y^2}{2b}$ とすれば (ba),(bc) により $\hat{\boldsymbol{w}}^{MAP} = \hat{\boldsymbol{w}}^{L1}$ となることが分かる.

最尤推定,MAP 推定とベイズ推論の相違点

最尤推定 (或いは誤差最小化) や MAP 推定 (正則化) はパラメータを未知の定数だと見做して, 一つのパラメータ点を推定する点推定 (point estimation) である.
一方でベイズ推論ではパラメータを確率変数と見做して, その分布を推定する.

二値分類モデルへの拡張

実数地を予測する回帰モデルから有限個への分類モデルへの拡張法を考える. まずラベル y が二値である場合を考える. このとき y は Bernoulli 分布に従うと考えられる.

$$y \sim \mathrm{Bern}(\mu) \tag{3.102}$$

さらにパラメータ $\mu \in (0,1)$ は回帰モデルの連続値出力のシグモイド変換によって得られると仮定する. すなわち

$$\mathrm{Sig}^{-1}(\mu) \mid \boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(f(\boldsymbol{x}; \boldsymbol{w}), \sigma^2) \tag{3.104}$$

特に線形回帰の場合にはロジスティック回帰モデル (logistic regression model) と呼ばれる.

ベルヌーイ分布の最尤推定と交差エントロピー誤差

ベルヌーイ分布の最尤推定を考える. 一回一回の試行が独立であるとき対数尤度函数は,

$$\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) = \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i,\boldsymbol{w}) = \sum_{i=1}^{N} \{y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\} \tag{3.106}$$

となる. これは交差エントロピー誤差函数

$$E(\boldsymbol{w}) = -\sum_{i=1}^{N} \{y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\} \tag{2.60}$$

に対して $\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) = -E(\boldsymbol{w})$ という関係が成り立っている. 従って尤度函数の最大化は交差エントロピーの最小化と等価である.

多値分類モデルへの拡張

次にラベルがベクトルでパラメータ $\boldsymbol{\pi}$ のカテゴリ分布に従う場合を考える.

$$\boldsymbol{y} \sim \mathrm{Cat}(\boldsymbol{\pi}) \tag{3.107}$$

さらにパラメータ $\boldsymbol{\pi}$ は回帰モデルの連続値出力のソフトマックス変換によって得られると仮定する. すなわち

$$\pi_n = \frac{\exp \eta_n}{\sum_{k=1}^{D} \exp \eta_k} \quad (n \in \{1, \ldots, D\}) \tag{3.108}$$

$$\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}), \sigma^2 \boldsymbol{I}) \tag{3.109}$$

但し $\boldsymbol{f} : \mathbb{R}^{H_0} \to \mathbb{R}^D$ である.

カテゴリ分布の最尤推定と交差エントロピー誤差

カテゴリ分布の最尤推定を考える. 一回一回の試行が独立であるとき対数尤度函数は

$$\begin{aligned} \log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) &= \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i,\boldsymbol{w}) \\ &= \sum_{i=1}^{N}\sum_{n=1}^{D} y_{i,n}\log \pi_{i,n} \end{aligned} \tag{3.110}$$

これは交差エントロピー誤差函数

$$E(\boldsymbol{w}) = -\sum_{i=1}^{N}\sum_{n=1}^{D} y_{i,n}\log \pi_{i,n} \tag{2.62}$$

に対して $\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w}) = -E(\boldsymbol{w})$ という関係が成り立っている. 従って尤度関数の最大化は交差エントロピーの最小化と等価である.

補足:逆行列の存在

Proposition 4

(3.72) のように $\hat{\mathbf{\Sigma}}^{-1}$ を定めたとき逆行列 $\hat{\mathbf{\Sigma}}$ は存在する.

Proof.

定義から明らかに $\hat{\mathbf{\Sigma}}^{-1}$ は対称行列である. $\boldsymbol{y} \in \mathbb{R}^{H_1} \backslash \{\mathbf{0}\}$ を任意とする. このとき $\hat{\mathbf{\Sigma}}^{-1}$ の二次形式は,

$$\boldsymbol{y}^\top \hat{\mathbf{\Sigma}}^{-1} \boldsymbol{y} = \sigma_y^{-2} \sum_{i=1}^{N} \left(\boldsymbol{y}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) \right)^2 + \boldsymbol{y}^\top (\sigma_w^{-2} \boldsymbol{I}) \boldsymbol{y} \geq \boldsymbol{y}^\top (\sigma_w^{-2} \boldsymbol{I}) \boldsymbol{y} > 0$$

最後は $\sigma_w^{-2}\boldsymbol{I}$ が正定値行列であることから従う. よって $\hat{\mathbf{\Sigma}}^{-1}$ は正定値行列である. 正定値であれば正則である. 何故ならば,

$$\text{正則} \iff \text{斉次方程式の解が } \mathbf{0} \text{ のみ}$$

であるので [4] 斉次方程式 $\hat{\mathbf{\Sigma}}^{-1}\boldsymbol{z} = \mathbf{0}$ を考えたとき両辺に $\boldsymbol{z}^\top$ を左からかけると $\boldsymbol{z}^\top \hat{\mathbf{\Sigma}}^{-1} \boldsymbol{z} = 0$ で正定値性から $\boldsymbol{z} = \mathbf{0}$ のみだとわかる為. □

補足:逆行列の正定値性

Proposition 5

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ が正定値行列ならば, その逆行列 $\boldsymbol{A}^{-1}$ も正定値行列となる.

Proof.

$\boldsymbol{A}$ が正定値のときの逆行列の存在は命題 4 で既に示した. まず $\boldsymbol{A}^{-1}$ は対称行列である. 何故ならば,

$$(\boldsymbol{A}^{-1})^\top \boldsymbol{A} = (\boldsymbol{A}^{-1})^\top \boldsymbol{A}^\top = (\boldsymbol{A}\boldsymbol{A}^{-1})^\top = \boldsymbol{I}$$

より逆元の一意性から $\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^\top$ が得られる為である. 続いて正定値性を示す.$\boldsymbol{A}^{-1}$ は正則なので任意の $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ に対して $\boldsymbol{A}^{-1}\boldsymbol{x} \neq \boldsymbol{0}$ である. よって $\boldsymbol{A}$ の正定値性から

$$0 < (\boldsymbol{A}^{-1}\boldsymbol{x})^\top \boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x}$$

$\boldsymbol{x}$ は零ベクトルでない任意の元であったので,$\boldsymbol{A}^{-1}$ が正定値行列であることが示された. $\square$

補足:逆行列の存在-弐 (0)

まずは定理を再掲する.

Theorem 1

$\mathbb{R}^n$ 上の N 個のベクトル $\{\boldsymbol{x}_i\}_{i=1}^N$ について以下の三条件は同値である.

1. $\mathbb{R}^n$ 上の基底を少なくとも一つ取れる
2. $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正定値行列
3. $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正則

補足:逆行列の存在-弐 (1)

Proof.

$(1 \Longrightarrow 2)$
条件より $n \leq N$ となることに注意する. $\mathbb{R}^n$ の基底を $< \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n >$ とおいても一般性は失われない. 基底をその順のまま並べて出来る行列を基底行列と呼び以降 $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ と書く. このとき基底行列は正則である. 何故ならば, 基底は線形独立なので

$$a_1\boldsymbol{x}_1 + \cdots + a_n\boldsymbol{x}_n = \mathbf{0} \Longleftrightarrow \boldsymbol{a} = (a_1, \ldots, a_n)^\top = \mathbf{0}$$

である. 左辺は書き換えると $\boldsymbol{X}\boldsymbol{a} = \mathbf{0}$ であり, これは $\boldsymbol{a}$ に関する斉次方程式である. $\boldsymbol{X}$ が正則であることと, 斉次方程式の解が $\mathbf{0}$ のみであることは同値であるから, $\boldsymbol{X}$ が正則であることがわかった. 次に $\boldsymbol{X}\boldsymbol{X}^\top$ が正定値であることを示す. 変形すると

$$\boldsymbol{X}\boldsymbol{X}^\top = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} = \sum_{i=1}^{N} \boldsymbol{x}_i\boldsymbol{x}_i^\top$$

であるから $\boldsymbol{y} \in \mathbb{R}^n$ を任意とすると,

$$\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{y}^\top\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\boldsymbol{y} = \sum_{i=1}^{n} (\boldsymbol{y}^\top\boldsymbol{x}_i)^2 \geq 0$$

よって半正定値であることが示された. 等号成立は $\boldsymbol{y}^\top\boldsymbol{x}_i = 0$ が任意の $i \in \{1, \ldots, n\}$ について成り立つときである.

補足:逆行列の存在-弐 (2)

Proof.

$< \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n >$ は $\mathbb{R}^n$ の基底であるので任意の $\boldsymbol{y}$ に対して

$$\boldsymbol{y} = b_1\boldsymbol{x}_1 + \cdots + b_n\boldsymbol{x}_n = \boldsymbol{X}\boldsymbol{b}$$

となる $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$ をただ一つとることができる. 任意の i に対して $\boldsymbol{y}^\top\boldsymbol{x}_i = 0$ が成り立つときとは $\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{y}^\top\boldsymbol{x}_n \end{bmatrix} = \boldsymbol{0}$ となるときである. この左辺を変形すると

$$\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{y}^\top\boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top\boldsymbol{X}\boldsymbol{b} \\ \vdots \\ \boldsymbol{x}_n^\top\boldsymbol{X}\boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b}$$

$\boldsymbol{X}$ は正則であったので行列式が等しい $\boldsymbol{X}^\top$ も正則でそれらの積 $\boldsymbol{X}^\top\boldsymbol{X}$ も正則となる. よって $\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{y}^\top\boldsymbol{x}_n \end{bmatrix} = (\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{b} = \boldsymbol{0}$ は斉次方程式でありその解は $\boldsymbol{b} = \boldsymbol{0}$ のみである. 従って $\boldsymbol{X}\boldsymbol{X}^\top$ は正定値行列であることが示された.$N = n$ のときはここで証明終了である.

補足:逆行列の存在-弐 (3)

Proof.

$N > n$ のときは $N - n = m$ として

$$\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \sum_{i=1}^{m} \boldsymbol{x}_{n+i} \boldsymbol{x}_{n+i}^\top = \boldsymbol{X}\boldsymbol{X}^\top + \sum_{i=1}^{m} \boldsymbol{x}_{n+i} \boldsymbol{x}_{n+i}^\top$$

と表すことができる.$\boldsymbol{y} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ を任意にとってくると

$$\boldsymbol{y}^\top \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} + \sum_{i=1}^{m} (\boldsymbol{y}^\top \boldsymbol{x}_{n+i})^2 \geq \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y} > 0$$

最後は $\boldsymbol{X}\boldsymbol{X}^\top$ が正定値であることから従う. よって $\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正定値行列であることが示された.
$(2 \implies 3)$ 正定値行列は正則行列である. これは命題 4 で既に示した.
$(3 \implies 1)$
対偶証明を行う. すなわち

$$\mathbb{R}^n \text{の基底を一組も取れない} \implies \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \text{は正則ではない}$$

ことを証明する. 次の補題を利用する.

補足:逆行列の存在-弐 (4)

Proof.

$A \in \mathbb{R}^{l\times m}, B \in \mathbb{R}^{m\times n}$ のとき

$$\mathrm{rank}(AB) \le \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$$

(証明)
$r_A = \mathrm{rank}(A), r_B = \mathrm{rank}(B)$ とする．階数の定義から，ある基本行列の積 P_A, Q_A, P_B, Q_B があって

$$P_A A Q_A = \begin{bmatrix} E_{r_A} & O_{r_A, m-r_A} \\ O_{l-r_A, r_A} & O_{l-r_A, m-r_A} \end{bmatrix}, P_B B Q_B = \begin{bmatrix} E_{r_B} & O_{r_B, n-r_B} \\ O_{m-r_B, r_B} & O_{m-r_B, n-r_B} \end{bmatrix}$$

となる．この両辺にそれぞれ Q_A^{-1} を右から,P_B^{-1} を左からかけると，

$$P_A A = \begin{bmatrix} E_{r_A} & O_{r_A, m-r_A} \\ O_{l-r_A, r_A} & O_{l-r_A, m-r_A} \end{bmatrix} \begin{bmatrix} Q_{r_A} \\ Q_{m-r_A} \end{bmatrix} = \begin{bmatrix} Q_{r_A} \\ O_{l-r_A, m} \end{bmatrix}$$

$$BQ_B = \begin{bmatrix} P_{r_B} & P_{m-r_B} \end{bmatrix} \begin{bmatrix} E_{r_B} & O_{r_B, n-r_B} \\ O_{m-r_B, r_B} & O_{m-r_B, n-r_B} \end{bmatrix} = \begin{bmatrix} P_{r_B} & O_{m, n-r_B} \end{bmatrix}$$

但し $Q_A^{-1} = \begin{bmatrix} Q_{r_A} \\ Q_{m-r_A} \end{bmatrix}, P_B^{-1} = \begin{bmatrix} P_{r_B} & P_{m-r_B} \end{bmatrix}$ である．

補足:逆行列の存在-弐 (5)

Proof.

従ってこの二つから

$$P_A(AB)Q_B = \begin{bmatrix} Q_{r_A} \\ O_{l-r_A,m} \end{bmatrix} \begin{bmatrix} P_{r_B} & O_{m,n-r_B} \end{bmatrix} = \begin{bmatrix} Q_{r_A}P_{r_B} & O_{r_A,n-r_B} \\ O_{l-r_A,r_B} & O_{l-r_A,n-r_B} \end{bmatrix}$$

行列の階数は基本変形によって変わらないので $\mathrm{rank}(AB) = \mathrm{rank}(P_A ABQ_B)$ である. また $P_A ABQ_B$ の構成を見ればさらに $\mathrm{rank}(P_A ABQ_B) = \mathrm{rank}(Q_{r_A}P_{r_B})$ となり行列の階数は行数および列数を超えないから $\mathrm{rank} \leq \min\{r_A, r_B\}$ を得る. これらをまとめて $\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$ を得る. □

補題に於いて $A = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$,$B = A^\top$ としてとってくると

$$AB = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N] \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

となるから $\mathrm{rank}(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top) \leq \mathrm{rank}(A)$ を得る. 転置行列の階数が元の行列の階数に一致するのは階数の定義から明らかであるので省いた.

補足:逆行列の存在-弐 (6)

Proof.

あとは $\mathrm{rank}(A) < n$ であることを示せばよい. $<\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m>$ を $\{\boldsymbol{x}_i\}_{i=1}^N$ の極大線形独立系 [4] とする. すなわち

1. $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m$ は線形独立である
2. $\{\boldsymbol{x}_i\}_{i=1}^N$ の任意のベクトルは $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m$ の線型結合として表せる

を満たすとする. 条件より $\{\boldsymbol{x}_i\}_{i=1}^N$ からは $\mathbb{R}^n$ の基底を一組も取れないから $m < n$ である. このとき, 明らかに A に基本変形を施して $\left[\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m, O_{N-m}\right]$ を得るので $\mathrm{rank}(A) = m$ が従う. よって

$$\mathrm{rank}\left(\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \leq \mathrm{rank}\left(\left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right]\right) = m < n$$

となるから $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^\top$ は正則ではないことが示された. □

補足:正規分布の積分計算

Theorem 2

$\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ は正定値行列だとする．このとき

$$I \triangleq \int \cdots \int_{\mathbb{R}^n} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} d\boldsymbol{x} = \sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}$$

Proof.

$\boldsymbol{\Sigma}$ は正定値なのでその逆行列 $\boldsymbol{\Sigma}^{-1}$ も正定値である (命題 5)．正定値行列は唯一つの Cholesky 分解を持つ [3]. すなわち対角成分がいずれも正の実数であるような唯一つの下三角行列 L があり $\boldsymbol{\Sigma}^{-1} = LL^\top$ が成り立つ．このことから $\det L = \det L^\top > 0$ であり従って $\det \boldsymbol{\Sigma}^{-1} > 0$ また $\det \boldsymbol{\Sigma} > 0$ が成り立つ.$L^\top$ は正則であるから変数変換 $\boldsymbol{z} = L^\top(\boldsymbol{x}-\boldsymbol{\mu})$ は全単射である．このとき

$$\boldsymbol{z}^\top \boldsymbol{z} = (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \ , \ \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \det L^\top$$

である. 但し $\partial \boldsymbol{z}/\partial \boldsymbol{x}$ はヤコビアンである.

補足:正規分布の積分計算 (2)

Proof.

積分の変数変換公式から

$$I = \int_{\mathbb{R}^n} \exp\left\{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right\}\left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}\right| d\boldsymbol{z} = \prod_{i=1}^{n}\left(\int_{\mathbb{R}} \exp(-\frac{1}{2}z_i^2)dz_i\right) |\det L^\top|^{-1}$$

ここでガウス積分の公式と $\det L \det L^\top = \det \boldsymbol{\Sigma}^{-1} > 0$ から

$$I = (\sqrt{2\pi})^n |\sqrt{\det \boldsymbol{\Sigma}^{-1}}|^{-1} = \sqrt{(2\pi)^n}\sqrt{\frac{1}{\det \boldsymbol{\Sigma}}}^{-1} = \sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}$$

□

参考文献

[1] 須山敦志 (2019)『ベイズ深層学習』(機械学習プロフェッショナルシリーズ) 講談社.

[2] 渡邊澄夫 (2015)「初めてのベイズ学習」,[online] http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/Bayestheory.html (参照 2020/5/1)

[3] 山本哲郎 (2003)『数値解析入門 [増訂版]』(サイエンスライブラリ現代数学への入門=14) サイエンス社.

[4] 齋藤正彦 (1966)『線形代数入門』(基礎数学 1) 東京大学出版社.

[5] 久保川達也 (2017)『現代数理統計学の基礎』(共立講座 数学の魅力 11) 共立出版株式会社.