発表予定日：2019 年 11 月 4 日
慶應義塾大学理工学部物理学科
岡崎健人

卒研ゼミ「深層学習」

5　機械学習の基礎

5.6　ベイズ統計

データ $D_m := \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\right\}$ が与えられたとき、データによって推定されるパラメータが $\boldsymbol{\theta}$ という値をとりうる確率はベイズの定理より

$$p(\boldsymbol{\theta} \mid D_m) = \frac{p(D_m \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D_m)} \propto p(D_m \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1}$$

となる。式 (1) の $p(\boldsymbol{\theta})$ を事前確率（事前分布）、$p(\boldsymbol{\theta} \mid D_m)$ を事後確率（事後分布）、$p(D_m \mid \boldsymbol{\theta})$ を尤度（尤度関数）という。事前確率はデータを手に入れる前に想定していた確率であるのに対して、事後確率はデータを得たあとに事前確率を修正（ベイズ修正）したものである。式 (1) からわかるように、ベイズ修正とは事前確率に尤度をかけて規格化することで、よりもっともらしい分布に修正することである。

例：ベイズ線形回帰

データ $X,\ \boldsymbol{y}$ が得られたもとでパラメータが $\boldsymbol{w}$ となる条件付き確率 $p(\boldsymbol{w} \mid X, \boldsymbol{y})$ を求めたい。

$$p(\boldsymbol{w} \mid X, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid X, \boldsymbol{w})p(X \mid \boldsymbol{w})p(\boldsymbol{w})}{p(X, \boldsymbol{y})} \propto p(\boldsymbol{y} \mid X, \boldsymbol{w})p(\boldsymbol{w}). \tag{2}$$

尤度 $p(\boldsymbol{y} \mid X, \boldsymbol{w})$ が正規分布に従うとき、式 (2) の事後確率を具体的に計算する例をみせる。

$$p(\boldsymbol{y} \mid X, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; X\boldsymbol{w}, I) \propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^{\mathsf{T}}(\boldsymbol{y} - X\boldsymbol{w})\right). \tag{3}$$

式 (3) は、観測値 $\boldsymbol{y} = [y_1, \ldots, y_m]^{\mathsf{T}}$ の平均からのずれ具合が正規分布に従うと仮定していることを意味している。事前確率はデータを得る前に想定していた確率のことであるから、過去の経験にもとづいた、想定していた分布を設定すればよい。ここでは計算を容易にするため、事前分布は尤度と同じく正規分布と設定する。

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_0, \Lambda_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^{\mathsf{T}}\Lambda_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\right). \tag{4}$$

実用上、$\Lambda_0 = \mathrm{diag}(\boldsymbol{\lambda}_0)$ などとすれば計算はさらに簡単になる。

事後確率は尤度と事前確率の積に比例するので、式 (2)、式 (3)、式 (4) より

$$p(\boldsymbol{w} \mid X, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^{\mathsf{T}}(\boldsymbol{y} - X\boldsymbol{w})\right) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^{\mathsf{T}}\Lambda_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\right).$$

指数関数の肩の部分を展開して、$\boldsymbol{w}$ の二次形式 $-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_m)^{\mathsf{T}}\Lambda_m^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_m)$ の形にまとめることを試みる。まず Λ_0^{-1} は対称行列であり、$\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y}$ と $\boldsymbol{w}^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{\mu_0}$ はスカラーであるから、$\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y} = \left(\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y}\right)^{\mathsf{T}} =$

$\boldsymbol{y}^\mathsf{T} X\boldsymbol{w}$, $\boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{\mu_0} = \big(\boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{\mu_0}\big)^\mathsf{T} = \boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w}$ のように、項の左側にある $\boldsymbol{w}^\mathsf{T}$ を $\boldsymbol{w}$ として項の右側に移すことができる。ゆえに

$$\begin{aligned}-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{w})^\mathsf{T}(\boldsymbol{y}-X\boldsymbol{w}) &= -\frac{1}{2}\big(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}X\boldsymbol{w} - \boldsymbol{w}^\mathsf{T}X^\mathsf{T}\boldsymbol{y} + \boldsymbol{w}^\mathsf{T}X^\mathsf{T}X\boldsymbol{w}\big)\\ &= -\frac{1}{2}\big(\text{const.} - 2\boldsymbol{y}^\mathsf{T}X\boldsymbol{w} + \boldsymbol{w}^\mathsf{T}X^\mathsf{T}X\boldsymbol{w}\big),\end{aligned}$$

$$\begin{aligned}-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_0)^\mathsf{T}\Lambda_0^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_0) &= -\frac{1}{2}\big(\boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w} - \boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w} + \boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\boldsymbol{\mu}_0\big)\\ &= -\frac{1}{2}\big(\boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w} + \text{const.}\big).\end{aligned}$$

ただし $\boldsymbol{w}$ によらない項は定数 const. とした。ゆえに事後確率は

$$\begin{aligned}p(\boldsymbol{w}\mid X,\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\big(-2\boldsymbol{y}^\mathsf{T}X\boldsymbol{w} + \boldsymbol{w}^\mathsf{T}X^\mathsf{T}X\boldsymbol{w} + \boldsymbol{w}^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\boldsymbol{w}\big)\right)\\ &= \exp\left(-\frac{1}{2}\big[-2\big(\boldsymbol{y}^\mathsf{T}X + \boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}\big)\boldsymbol{w} + \boldsymbol{w}^\mathsf{T}\big(X^\mathsf{T}X + \Lambda_0^{-1}\big)\boldsymbol{w}\big]\right).\end{aligned} \tag{5}$$

ここで、目指している二次形式を展開すると

$$-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_m)^\mathsf{T}\Lambda_m^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_m) = -\frac{1}{2}\big(-2\boldsymbol{\mu}_m^\mathsf{T}\Lambda_m^{-1}\boldsymbol{w} + \boldsymbol{w}^\mathsf{T}\Lambda_m^{-1}\boldsymbol{w}\big) + \text{const.}$$

であるから、これと式 (5) の引数部分を比較して、$\Lambda_m^{-1} = X^\mathsf{T}X + \Lambda_0^{-1}$, $\boldsymbol{\mu}_m^\mathsf{T}\Lambda_m^{-1} = \boldsymbol{y}^\mathsf{T}X + \boldsymbol{\mu}_0^\mathsf{T}\Lambda_0^{-1}$ を得る。すなわち

$$\Lambda_m = \big(X^\mathsf{T}X + \Lambda_0^{-1}\big)^{-1}, \quad \boldsymbol{\mu}_m = \Lambda_m\big(X^\mathsf{T}\boldsymbol{y} + \Lambda_0^{-1}\boldsymbol{\mu}_0\big)$$

として、事後確率は

$$p(\boldsymbol{w}\mid X,\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_m)^\mathsf{T}\Lambda_m^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_m)\right)$$

となり、これもまた正規分布に従っている。この計算結果は、はじめ $\boldsymbol{w}$ は平均 $\boldsymbol{\mu}_0$, 分散共分散行列 Λ_0 の正規分布に従うと想定していたものが、データ X, $\boldsymbol{y}$ を得たことにより、平均 $\boldsymbol{\mu}_m$, 分散共分散行列 Λ_m の正規分布に従うように修正されたことを表している。ただし Λ_m が正則になるように Λ_0 を設定する必要はある。この例のように、事前分布と事後分布が同じ分布族になるとき、これを共役自然分布という [1]。

事前確率において $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\Lambda_0 = \frac{1}{\lambda}I$ と設定した場合、事後確率において $\Lambda_m^{-1} = X^\mathsf{T}X + \lambda I$, $\boldsymbol{\mu}_m = \Lambda_m X^\mathsf{T}\boldsymbol{y}$ となるので、$\boldsymbol{w}$ の真の値 $\boldsymbol{\mu}_m$ についての方程式

$$\Lambda_m^{-1}\boldsymbol{\mu}_m = \big(X^\mathsf{T}X + \lambda I\big)\boldsymbol{\mu}_m = X^\mathsf{T}\boldsymbol{y}.$$

が得られる。ここで X と $\boldsymbol{y}$ は有限個の観測されたデータであるため、この方程式を解いて得られる $\boldsymbol{\mu}_m$ の値は実際には $\boldsymbol{w}$ の推定値 $\hat{\boldsymbol{\mu}}_m$ である。この方程式は、正則化項を重み減衰 $\lambda\|\boldsymbol{w}\|_2^2$ として正則化最小二乗法により得られた方程式 $\big(X^\mathsf{T}X + \lambda I\big)\boldsymbol{w} = X^\mathsf{T}\boldsymbol{y}$ に一致している。ただし $\lambda = 0$ としてしまうと、これは $\boldsymbol{w}$ の分散がはじめ無限大であったことを表しているが、Λ_0, Λ_0^{-1} を定義できないため、ベイズ推定ではこの場合を取り扱うことができない。さらに重要な違いとしては、正則化最小二乗法では $\boldsymbol{w}$ の推定値のみ得られていたが、ベイズ推定ではそれに加えて $\boldsymbol{w}$ の分散共分散行列 Λ_m も得ることができる。

5.6.1 MAP 推定

事後確率 $p(\boldsymbol{\theta} \mid D_m)$ が最大になるときの $\boldsymbol{\theta}$ の値をその推定値とする方法を、MAP 推定（最大事後確率推定、maximum a posteriori estimation）という。すなわちその推定値を $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ とかけば、

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid D_m) = \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid D_m) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left[\log p(D_m \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right].$$

例として、線形回帰モデルにおいて事前分布 $p(\boldsymbol{w})$ が正規分布 $\mathcal{N}\left(\boldsymbol{w}; \mathbf{0}, \frac{1}{\lambda} I\right)$ に従うと設定する。すると

$$p(\boldsymbol{w}) \propto \exp\left(-\frac{1}{2}\lambda \boldsymbol{w}^{\mathsf{T}} \boldsymbol{w}\right), \quad \log p(\boldsymbol{w}) = -\frac{\lambda}{2} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{w} + \text{const.}$$

であるから、

$$\log p(\boldsymbol{w} \mid X, \boldsymbol{y}) = \log p(\boldsymbol{y} \mid X, \boldsymbol{w}) + \log p(\boldsymbol{w}) + \text{const.} = -\mathrm{MSE}(\boldsymbol{w}) - \frac{\lambda}{2} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{w} + \text{const.}$$

これは最尤推定法で最大化する関数（式 5.65）に重み減衰を付加したものになっていることがわかる。符号がマイナスとなっているから、これを最大化することは、正則化最小二乗法における $\mathrm{MSE}(\boldsymbol{w}) + \frac{\lambda}{2} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{w}$ を最小化することと等価である。

データからのみでは得られないような情報を事前確率から得ることができるため、MAP 推定は有用である。ただしそれによって推定量のバリアンスを最尤推定に比べて小さくできるが、バイアスは増大する。

正則化を含む推定方法の多くは、裏で MAP 推定を行なっていると考えることができ、その正則化項は $\log p(\boldsymbol{\theta})$ に対応している。ただしすべての正則化項が $\log p(\boldsymbol{\theta})$ に対応しているわけではない。例えば正則化項がデータを含むとき、$p(\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ のみの関数であるため、「正則化項は $\log p(\boldsymbol{\theta})$ に対応している」と考えることはできない。逆に MAP 推定によって機械的に正則化項を設計できる。たとえば事前確率 $p(\boldsymbol{\theta})$ を正規分布の線形結合と設定すれば、より複雑な正則化項を作り出すことができる。

参考文献

[1] 安道知寛，「ベイズ統計モデリング」，株式会社朝倉書店，2010 年．pp. 28–41.