作成日：2020 年 1 月 17 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 $D_m := \left\{x^{(1)}, \dots, x^{(m)}\right\}$ が与えられたとき、データによって決まるパラメータが $\boldsymbol{\theta}$ 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。中辺の分母は

$$p(D_m) = \int p(D_m \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

で計算でき、これは $\boldsymbol{\theta}$ によらず、また $D_m = \left\{x^{(1)}, \dots, x^{(m)}\right\}$ の値は既知のものであるから、定数として扱ってよい。式 (1) の $p(\boldsymbol{\theta})$ を事前確率（分布）、$p(\boldsymbol{\theta} \mid D_m)$ を事後確率（分布）、$p(D_m \mid \boldsymbol{\theta})$ を尤度という。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し、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

m 個の訓練データ $\left(X^{(\mathrm{train})}, \boldsymbol{y}^{(\mathrm{train})}\right) = (X, \boldsymbol{y})$ が与えられたとする（右肩の (train) は省略する）。ここに $\boldsymbol{y} = [y_1, \dots, y_m]^{\mathsf{T}}$ であり、基底関数を $\phi_k(x),\ k = 0, \dots, n$ とすれば計画行列は $(X)_{ij} = \phi_j(x_i)$ である（推定する関数を n 次多項式と仮定したときは $\phi_k(x) = x^k$ であった）。$\boldsymbol{x} = [\phi_0(x), \dots, \phi_n(x)]^{\mathsf{T}},\ \boldsymbol{w} = [w_0, w_1, \dots, w_n]^{\mathsf{T}}$ とすれば、訓練データによる y の推定は $\hat{y} = \sum_{k=0}^{n} w_k \phi_k(x) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}$ と表現できる。また各訓練データ $x_i\ (i = 1, \dots, m)$ に対する y の推定を計算した $\hat{y}_i = \sum_{k=0}^{n} w_k \phi_k(x_i)$ を並べたベクトルを $\hat{\boldsymbol{y}} = \left[\hat{y}_1, \dots, \hat{y}_m\right]^{\mathsf{T}}$ と書けば、$\hat{\boldsymbol{y}} = X\boldsymbol{w}$ である。

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

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

尤度を $p(\boldsymbol{y} \mid X, \boldsymbol{w})$ としたのは理由がある。ベイズの定理を考えれば、尤度としての候補は $p(X, \boldsymbol{y} \mid \boldsymbol{w}),\ p(X \mid \boldsymbol{y}, \boldsymbol{w}),\ p(\boldsymbol{y} \mid X, \boldsymbol{w})$ の 3 つが挙げられる。しかし推定したい関数 $\hat{y}$ は X と $\boldsymbol{w}$ が決まれば決まるというものなので、尤度としては $p(\boldsymbol{y} \mid X, \boldsymbol{w})$ が適当である。また $\boldsymbol{w}$ の値は $X,\ \boldsymbol{y}$ の値によって決められる不確かさを持っているが、X の値は観測された既知のデータで固定されているため、$\boldsymbol{w}$ の値がわかっている条件のもと

でデータの値が X である条件付き確率 $p(X \mid \boldsymbol{w})$ は、$\boldsymbol{w}$ の値によらず常に 1 である。また式 (2) の分母は $\boldsymbol{w}$ について積分をしたものなので、$\boldsymbol{w}$ によらない定数と考えることができる。

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。

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

なお多変量正規分布は

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \sqrt{\frac{1}{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{4}$$

であった（教科書の式 3.23)。この $\boldsymbol{\mu}$ は $\boldsymbol{x}$ の各成分の平均を並べたベクトルで、Σ は分散共分散行列である。ただし式 (3) では各 y_i の分散 s_i^2 は 1 で、$i \neq j$ に対して y_i と y_j の共分散 s_{ij} はゼロ、すなわち相関はないと仮定している。式 (3) は、観測値 $\boldsymbol{y} = [y_1, \ldots, y_m]^{\mathsf{T}}$ 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 [1]。

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。つまり予想とか「直感的信頼度（the degree of belief)」でよい。たとえば何も情報を得ておらず、どんな分布かもわからない場合は、ただ単に定数とすることがある（無情報事前分布、non-informative prior という [2])。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。

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_0, \Lambda_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^{\mathsf{T}}\Lambda_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\right). \tag{5}$$

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 (5) より

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

これは $e^x e^y = e^{x+y}$ のようにひとつの指数関数にまとめることができる。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目指す。まず Λ_0^{-1} は対称行列であり、$\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y}$ と $w^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{\mu_0}$ はスカラーであるから、$\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y} = \left(\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}\boldsymbol{y}\right)^{\mathsf{T}} = \boldsymbol{y}^{\mathsf{T}}X\boldsymbol{w}$, $\boldsymbol{w}^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{\mu_0} = \left(\boldsymbol{w}^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{\mu_0}\right)^{\mathsf{T}} = \boldsymbol{\mu}_0^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{w}$ のように、項の左側にある $\boldsymbol{w}^{\mathsf{T}}$ を項の右側に移すことができる。

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

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

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

$$\begin{aligned}
p(\boldsymbol{w} \mid X, \boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\left(-2\boldsymbol{y}^{\mathsf{T}}X\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{w} - 2\boldsymbol{\mu}_0^{\mathsf{T}}\Lambda_0^{-1}\boldsymbol{w}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left[-2\left(\boldsymbol{y}^{\mathsf{T}}X + \boldsymbol{\mu}_0^{\mathsf{T}}\Lambda_0^{-1}\right)\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\left(X^{\mathsf{T}}X + \Lambda_0^{-1}\right)\boldsymbol{w}\right]\right).
\end{aligned} \tag{6}$$

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 (6) 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、事前分布が正規分布に従うとき事後確率は

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

となり、これもまた正規分布に従っていることがわかる。規格化をするには式 (4) を参照すればよい。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 [2]。

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 $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ とかけば、

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

である。ただしここでは関数 $f(\boldsymbol{\theta})$ の最大点と $\log f(\boldsymbol{\theta})$ の最大点が一致することと、$p(\boldsymbol{\theta} \mid D_m) \propto p(D_m \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ であることを用いている。

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 $\mathcal{N}\big(\boldsymbol{w}; \boldsymbol{0}, \frac{1}{\lambda}I\big)$ 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、

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

である。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。なお式に定数が足されているが、最大・最小となる点はそれによって影響されないので、やっていることは正則化最小二乗法と同じである。

As with full Bayesian inference, MAP Bayesian inference has the advantage of leveraging information that is brought by the prior and cannot be found in the training data. This additional information helps to reduce the variance in the MAP point estimate (in comparison to the ML estimate). However, it does so at the price of increased bias. なぜ？

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ので、MAP 推定は有用である。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することができるが、そのときバイアスの増大が伴う。

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 $\boldsymbol{\theta}$ のみによる確率密度関数であるため、「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 推定によって正則化項を設計することができる。

参考文献

[1] 中谷秀洋. 「第 12 回　ベイズ線形回帰［前編］」. gihyo.jp. https://gihyo.jp/dev/serial/01/machine-learning/0012. 最終閲覧 2020 年 1 月 17 日.

[2] 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.