1　エビデンス近似

線形基底関数を完全にベイズ的に取り扱うためにはパラメタ $\boldsymbol{w}$ だけでなく、超パラメタ α, β についても事前分布を導入し、全ての変数について周辺化を行う必要がある。しかしながら、3 変数全てに対し周辺化の計算を解析的に実行することは難しいため、別のアプローチが必要となる[*1]。そのアプローチは次のような手順となる。

（1）パラメタ $\boldsymbol{w}$ について周辺化を行い周辺尤度を得る。

（2）（1）で得た周辺尤度を最大化するような超パラメタ α, β を定める。

この 2 段階を経て超パラメタを定める手法のことをエビデンス近似と呼ぶ。本節の以下 2 つのサブセクションではそれぞれ（1）と（2）について解析計算を通じて理解を深める。

まず、$\boldsymbol{w}$ について周辺化するために、準備としてパラメタ、超パラメタの両方に関する周辺化積分の表式を立てる。

$$p(t|\boldsymbol{t}) = \int\int\int p(t|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{t},\alpha,\beta)d\boldsymbol{w}p(\alpha,\beta|\boldsymbol{t})d\alpha d\beta \tag{1}$$

被積分関数については、初めの 2 因子がパラメタの周辺化[*2]を与え、お尻の因子が超パラメタの周辺化を与えるものとなっている。この後 $\boldsymbol{w}$ についてのみ積分を実行したい。そのために α, β は次の条件を満たす $\hat{\alpha}, \hat{\beta}$ で固定してしまう。

・事後分布 $p(\alpha, \beta|\boldsymbol{t})$ が $\hat{\alpha}, \hat{\beta}$ の周りで鋭い値を持つ

この仮定によって変数 α, β についてデルタ関数[*3]の積分を実行できる。$p(\alpha, \beta|\boldsymbol{t})$ をデルタ関数を用いて次のように表現する。

$$p(\alpha,\beta|\boldsymbol{t}) \simeq \delta(\alpha-\hat{\alpha})\delta(\beta-\hat{\beta}) \tag{2}$$

この表現によって式 1 は

$$\begin{aligned} p(t|\boldsymbol{t}) &\simeq \int\int_{\beta}\int_{\alpha} p(t|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{t},\alpha,\beta)d\boldsymbol{w}\delta(\alpha-\hat{\alpha})\delta(\beta-\hat{\beta})d\alpha d\beta \\ &= \int p(t|\boldsymbol{w},\hat{\beta})p(\boldsymbol{w}|\boldsymbol{t},\hat{\alpha},\hat{\beta})d\boldsymbol{w} \ (\alpha=\hat{\alpha},\beta=\hat{\beta}\text{の値が取り出される。}) \\ &= p(t|\boldsymbol{t},\hat{\alpha},\hat{\beta}) \end{aligned} \tag{3}$$

となり、これが新しい目標値の予測分布となる。したがい、我々は $\hat{\alpha}, \hat{\beta}$ を定める必要がある。その定め方としては超パラメタについての周辺尤度関数 $p(\boldsymbol{t}|\alpha, \beta)$ を最大にするような $\hat{\alpha}, \hat{\beta}$ を選んでやるというものとなる。本来は超パラメタについての事後分布 $p(\alpha, \beta|\boldsymbol{t})$ のピークを与える α, β を選んでやらねばならない[*4]。しかし、もし超パラメタについての事前分布が平坦（つまり定数）であるならば、式（3.76）より、事後分布のピークを与える α, β を取ってくることと、周辺尤度関数のピークを与える α, β を取ってくることは等価と考えて良い。したがって今後は尤度 $p(\boldsymbol{t}|\alpha, \beta)$ について最大化することを考えていく。

[*1] どういった点で難しいのかが正直理解できなかった。積分の結果が陽な結果で表せず、例えば演算子を挟み込みまくった形で表されてしまい、その後の理論的な考察が進まなくなってしまうということだろうか。

[*2] これは散々やってきたので大丈夫だと思う。

[*3] デルタ関数の取り扱いについては右記サイトを参照（https://eman-physics.net/math/delta_func.html）。

[*4] デルタ関数の尖った部分を与える α, β を取ってくるということである。

以降のサブセクションでは、周辺分布関数 $p(\boldsymbol{t}|\alpha,\beta)$ をエビデンス関数として、この関数を最大化[*5]する計算について手を動かして理解していくことにする。

1.1 エビデンス関数の評価

本サブセクションでは、エビデンスの値からどのようなモデルを選択すべきか、その判断方法について学ぶ。ポイントは以下 2 つである。

（i）上記（1）で述べた周辺尤度 $p(\boldsymbol{t}|\alpha,\beta)$ を積分の評価を実行することで導出する。すなわちエビデンス関数を陽な形で求める。

（ii）多項式回帰の例で、フィッティング性能とエビデンスの値の関係から選ぶべきモデルを特定する。

まず、（i）について教科書の説明に従って（3.77）を変形していく。（3.11）、（3.12）、（3.52）を用いると周辺尤度は次のように書ける（＞対応；演習 3.17）。

$$
\begin{aligned}
p(\boldsymbol{t}|\alpha,\beta) &= \int p(\boldsymbol{t}|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w} \\
&= \int exp[\frac{N}{2}ln\beta - \frac{N}{2}ln(2\pi) - \beta E_D(\boldsymbol{w})](\frac{\alpha}{2\pi})^{M/2}exp(\alpha\frac{\boldsymbol{w}^T\boldsymbol{w}}{2})d\boldsymbol{w} \ (\because (3.11),(3.52)) \\
&= (\frac{\alpha}{2\pi})^{M/2}\int exp(\frac{N}{2}ln\frac{\beta}{2\pi})exp[-\beta E_D(\boldsymbol{w}) - \alpha\frac{\boldsymbol{w}^T\boldsymbol{w}}{2}]d\boldsymbol{w} \\
&= (\frac{\alpha}{2\pi})^{M/2}\int exp(ln(\frac{\beta}{2\pi})^{N/2})exp[-(\beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{w}))]d\boldsymbol{w} \ (\because (3.25)) \\
&= (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\int exp(-E(\boldsymbol{w}))d\boldsymbol{w} \ (\because (3.79)) \qquad (4)
\end{aligned}
$$

式 4 内の $E_W(\boldsymbol{w})$ を $\boldsymbol{w}$ について平方完成することで式（3.80）が得られる。ここから式（3.85）の計算は（式（3.85）2 行目の積分部分）×（規格化係数）= 1 を利用することで簡単に実行できる。したがって、求めるべき対数エビデンス関数は、上式 4 の係数部分を考えることで、式（3.86）のように陽な形で表現できる。

次に、（ii）で述べたモデル選択について図 3.14 と図 1.4-5[*6]を用いて議論する。

最初に、エビデンスの振る舞いについて図 3.14 を用いて考える。そのために、（3.86）の $\frac{M}{2}ln\alpha$ の項に注目する。今、設定として $\alpha = 5\times10^{-3}$ とすれば、$ln\alpha < 0$ なので M →大とすると、$\frac{M}{2}ln\alpha$ →小となる。M=0 〜3 での振る舞いが、なぜ図 3.14 のようになるのかはわからないが、M=3 で最大となるらしい。$M > 3$ では $\frac{M}{2}ln\alpha$ が支配的となって、M →大とすると、$lnp(\boldsymbol{t}|\alpha,\beta)$ →小となっていく。

続いて、フィッティングの振る舞いについて図 1.5 を用いて考える。図 1.5 を見ると M=3〜8 で訓練集合のフィッティング性能はほとんど変化がないことがわかる。

以上、エビデンスが M=3 で最大、フィッティング性能が M=3〜8 で一定であることから、選ぶべきモデルは M=3 であると結論づけられる。

1.2 エビデンス関数の最大化

上記（2）で述べたエビデンス関数の最大化と、そのときの超パラメタの値を求める。これらの超パラメタは陽な形では得られず、実際に値を計算する際は数値解析法が必要となることを予め述べておこう。以下では最大化について

[*5] 正確には対数をとった対数エビデンスを最大化するのだが。

[*6] かなりページを遡るので注意されたい。p.6-7 を参照。

（ア）α について最大化
（イ）β について最大化
の 2 つ場合で、それぞれ別途に実行する[*7]。
まず（ア）から取り組む。準備として、以下のような固有方程式を考える。

$$(\beta\Phi^T\Phi)\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{5}$$

ここで式（3.81）より

$$(\beta\Phi^T\Phi) = A - \alpha I \tag{6}$$

これを固有方程式に代入すると

$$\begin{aligned}(A-\alpha I)\boldsymbol{u}_i &= \lambda_i \boldsymbol{u}_i \\ A\boldsymbol{u}_i &= (\alpha+\lambda_i)\boldsymbol{u}_i\end{aligned} \tag{7}$$

したがって A の固有値は $\alpha + \lambda_i$ となる。ここで行列式と固有値に関する次の定理[*8]を用いる。
定理
A の固有値が $\lambda_1, \ldots, \lambda_n$ であるとき

$$|A| = \lambda_1 \ldots \lambda_n \tag{8}$$

となる。
これを上式 7 に適用すると

$$\begin{aligned}|A| &= (\lambda_1+\alpha)\ldots(\lambda_M+\alpha) \\ &= \Pi_{i=1}(\lambda_i+\alpha)\end{aligned} \tag{9}$$

周辺尤度（3.86）を α について最大化することを考えているから、上式 9 を（3.86）に代入し、α についての項のみを取り出すと

$$\begin{aligned}lnp(\boldsymbol{t}|\alpha,\beta) &= \frac{M}{2}ln\alpha - E(\boldsymbol{m}_N) - \frac{1}{2}ln\Pi_{i=1}(\lambda_i+\alpha) \\ &= \frac{M}{2}ln\alpha - E(\boldsymbol{m}_N) - \frac{1}{2}\Sigma_{i=1}ln(\lambda_i+\alpha)\end{aligned} \tag{10}$$

周辺尤度についての停留条件は、これを α について偏微分して、イコール 0 とすることに他ならないので

$$\begin{aligned}\frac{\partial}{\partial\alpha}lnp(\boldsymbol{t}|\alpha,\beta) &= \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N - \frac{1}{2}\Sigma_{i=1}\frac{1}{\lambda_i+\alpha} \ (\because (3.82)) \\ &= 0\end{aligned} \tag{11}$$

この式を以下のように変形していく。

$$\begin{aligned}M - \alpha\boldsymbol{m}_N^T\boldsymbol{m}_N - \Sigma\frac{\alpha}{\lambda_i+\alpha} &= 0 \\ \alpha\boldsymbol{m}_N^T\boldsymbol{m}_N &= M - \Sigma\frac{\alpha}{\lambda_i+\alpha}(=\gamma)\end{aligned} \tag{12}$$

[*7] 気になるのは α, β について、同時に最大化することはできないのか、ということ。α, β は後に示すように、γ を介して表されるため、互いに独立ではないはず。別個に求めた $\hat{\alpha}, \hat{\beta}$ を再代入したときに、エビデンス関数の真の最大値が得られるのかが不明。
[*8] 証明は右記サイトを参照（> https://risalc.info/src/determinant-eigenvalue-product.html）。

式 12 第 1 辺と γ の関係式は

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T \boldsymbol{m}_N} \tag{13}$$

であり、式 12 第 2 辺と γ の関係式は

$$\begin{aligned} M - \Sigma \frac{\alpha}{\lambda_i + \alpha} &= \gamma \\ M - \Sigma \frac{\lambda_i + \alpha - \lambda_i}{\lambda_i + \alpha} &= \gamma \\ M - \Sigma (1 - \frac{\lambda_i}{\lambda_i + \alpha}) &= \gamma \\ M - M - \Sigma \frac{\lambda_i}{\lambda_i + \alpha} &= \gamma \\ \gamma &= \Sigma \frac{\lambda_i}{\lambda_i + \alpha} \end{aligned} \tag{14}$$

となる。これらの結果を用いて周辺尤度を最大化するような α を iterative に計算することができる。ただし、iterative method の詳細な説明は教科書に任せる。

次に（イ）について取り組む。ほとんど（ア）と同じ手続きを踏めば良いのだが、$ln|A|$ を β で偏微分する際に、1 つ工夫が必要になる。その工夫は $d\lambda_i/d\beta = \lambda_i/\beta$ という関係式[*9]を式（3.94）の第二等号において用いることである。その後の式変形については、式（3.82）を使うところなど（ア）の時と同様である。初等的な式変形によって式（3.95）が導かれ、やはり β について self-consistent な方程式が得られるので、α の時と同じように iterative に計算を実行してやれば良い。

1.3 有効パラメタ数

事前パラメタ α と固有値 λ_i との大小関係から、パラメタ w_i と比 $\frac{\lambda_i}{\lambda_i+\alpha}$ についての解釈を与える。$\lambda_i >> \alpha$ のとき、つまりパラメタの値がデータの影響を受けるときは、w_i は最尤推定値に近くなり、かつ、$\frac{\lambda_i}{\lambda_i+\alpha}$ も 1 に近くなる。逆に、$\lambda_i << \alpha$ のとき、つまりパラメタの値がデータの影響を受けず、初期値から変化しないときは、w_i は 0 に近くなり、かつ、$\frac{\lambda_i}{\lambda_i+\alpha}$ も 0 に近くなる。最尤推定値に近いパラメタはモデルに組み込まれ、0 に近いパラメタはモデルに組み込まれない。前者が $\gamma \simeq M$、後者が $\gamma \simeq 0$ に相当するため、γ の大きさがデータに反応するパラメタの数を支配していると言える。

β の方は何言いたいかわからん。

近似式 2 つについては $N >> M$ であることを用いて、(3.92) と (3.25)、そして (3.95) と (3.26) との対応関係を考えれば容易に導出できる。

2 固定された基底関数の限界

ここまで学んできた固定された非線形基底関数を線形結合したモデルについて、利点と欠点を振り返る。ここは読み合わせ[*10]。

[*9] これを書いている時点で、なぜこの関係式が成立するのかわかっていない。

[*10] 読んでいて気になったことが 1 つ。「ベイズ推定の計算が簡単になる利点があった」とは、(3.98-99) の形が適用できるとき行列のスペクトル分解を必要としなくなるというところについての言及であろうか？それとも「基底関数を線形結合で表すモデル自体が単純なモデルである」という広い意味での主張をしているのだろうか？