オッカムのカミソリ

Occam’s Razor

Carl Edward Rasmussen and Zoubin Ghahramani (2000)

Bayes パラダイムは，明らかに Occam の剃刀を生み出すときがあるだけで，非常に大きなモデルがうまく機能するときもある。 この 2 つの挙動について簡単な例を挙げる。 この 2 つの見解は，関数を実装するための機械ではなく，関数の複雑さを測定することで調和する。 ガウス過程と等価な線形パラメータモデルについて関数の複雑さを解析したところ，常にオッカムの剃刀が働いていることがわかった。

1. はじめに

オッカムの剃刀（Occam's Razor）は、一般的な科学的思考、特に統計的推測の問題において影響力を持つ「説明の正当性（parsimony of explanations）」の原理としてよく知られている。 本稿では、ベイズ統計モデルにおけるオッカムの剃刀の帰結を検討する。 より単純なモデルを明示的に優先するような事前分布を構築しなければならないと思うかもしれない。 しかし、これから述べるように、オッカムの剃刀はベイズ理論の応用において実際に具現化される。 この考え方は「自動オッカムの剃刀」[Smith&Spiegelhalter1980, MacKay1992, Jefferys&Berger1992] として知られている。

我々は，しばしばノンパラメトリックと呼ばれる，多数のパラメータを持つ複雑なモデルに焦点を当てている。 個々のパラメータが果たす役割が必ずしも分かっておらず，推論は主にパラメータそのものを対象とするのではなく，モデルによってなされる予測を対象とするモデルを指すために，この用語を使用する。 このようなタイプのモデルは，機械学習における応用例の典型的なものである。

非ベイズの観点から，過学習を避けるために，限られた訓練データを考慮してモデルの複雑さを調整することが提唱されている。 モデルの複雑さは，モデル内の自由パラメータの数を調整することで調整されることが多く (重み減衰のような) 正則化の使用によって複雑さがさらに制約されることもある。 モデルの複雑さが低すぎても高すぎても，独立したテストセットでの成績が低下し，特徴的なオッカムの丘が生じる。 通常，モデルの複雑さを制御するために，汎化誤差の推定量または独立した検証セットが使用される。

ベイズの観点からは，モデルの複雑さの問題に関して，著者は 2 つの相反する立場を取っているようである。 ひとつは，いくつかの異なるモデルサイズごとにモデルの確率を推測し，予測を行う際にこれらの確率を使用するというものである。 もう 1 つの見解は，単純に「十分に大きな」モデルを選択し，モデルサイズの選択の問題を回避することを提案している。 どちらの見解も，パラメータは平均化されていると仮定していることに注意してください。 例 オッカムの剃刀を使ってニューラルネットワークの最適な隠れユニット数を決めるべきか，それとも単純に計算上可能な限り多くの隠れユニットを使うべきか？ 次に，これら 2 つの見方について詳しく説明する。

1.1 視点1：モデルサイズの選択

ベイズ学習の中心的な量の 1 つはエビデンスであり，尤度パラメータ w に対する積分として計算されるモデル $P(Y|\mathcal{M}_i)$ を与えられたデータの確率である。 エビデンスはベイズ則によってモデルの確率 $P(\mathcal{M}_i|Y)$ と関係づけられる:

$$P(Y|\mathcal{M}_i) = \int p(Y|w, \mathcal{M}_i)P(w|\mathcal{M}_i)\, dw,$$
$$P(\mathcal{M}_i|Y) = \frac{P(Y|\mathcal{M}_i)P(\mathcal{M}_i)}{P(Y)}$$

ここで，$P(\mathcal{M}_i|Y)$ が証拠に比例するように，モデルの事前分布 $P(\mathcal{M}_i)$ が flat であることは珍しくない。 図 1 は，エビデンスが複雑すぎるモデルを抑制し (脚注 1)，最も確率の高いモデルを選択するのに使える理由を説明している。

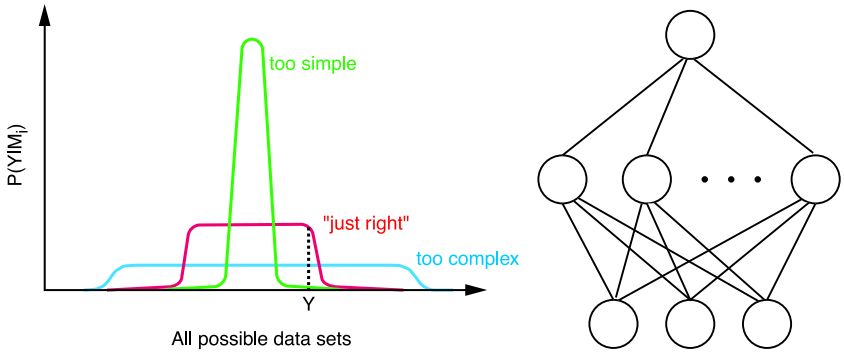


図 1左: すべての可能なデータセットの抽象的な一次元表現の関数としての証拠。 証拠は正規化しなければならないので，多くのデータセットを説明できる非常に複雑なモデルは，そこそこのエビデンスしか達成できない。 データセット Y が観測されたとき，証拠はモデルの複雑さを選択するために使われる。 このような選択は，尤度 $P(Y|w, \mathcal{M}_i)$ だけではできない。 右: 異なる数の隠れユニットを持つニューラルネットワークがモデルのファミリーを形成し，モデル選択問題を提起している。

1. 本当は，すべてのモデルの予測値を確率で加重平均すべきである。 しかし，エビデンスが強くピークを持つ場合，あるいは実際的な理由から，近似として 1 つを選択したい場合がある。

また，次のように解釈することで，エビデンスがいかに複雑すぎるモデルを抑制し，オッカムの剃刀を体現しているかを理解することができる。 エビデンスとは，モデルクラスからパラメータ値をランダムに選択した場合に，データセット Y が生成される確率のことである。 単純すぎるモデルは，その特定のデータセットを生成する可能性が非常に低いだろう。 一方，複雑すぎるモデルは，多くの可能性のあるデータセットを生成する可能性があるため，やはり，その特定のデータセットをランダムに生成する可能性は低いだろう。

1.2 視点2：大規模モデル

ノンパラメトリックベイズモデルでは、事前確率が信念を反映している限り，モデルを制約する統計的理由はない。 実際，モデルの次数 (つまりパラメータ数) をある小さな数に制限することは，真のデータ生成過程に関する事前確率としての信念と一致しないことが多いので，(どんなに多くのデータを持っていても) 大きなモデルを使い，可能であれば無限極限を追求することが理にかなっている (脚注 2)。 例えば、関数近似における基底関数の数を先験的に制限すべきではない。 なぜなら，データが少数の固定された基底関数から実際に生成されたとは考えていないからである。 したがって，計算上処理できる限り多くのパラメータを持つモデルを考えるべきである。

2. モデルによっては、パラメータ無限個の極限は扱いやすい単純なモデルである。 2 つの例として，ベイジアンニューラルネットワークのガウス過程極限 [Neal1996] とガウス混合モデルの無限極限 [Rasmussen2000] がある。

Neal1996 は，大量の隠れユニットを持つ多層パーセプトロンが，小さなデータセットでどのように優れた性能を達成するかを示した。 彼は高度な MCMC 手法を用いて，パラメータに対する平均化を実装した。 この考え方に従えば，モデルの複雑さを選択する課題はない: 我々はエビデンスを評価する必要がなく (これはしばしば困難である)，モデルのパラメータ数を制限するためにオッカムの剃刀を使う必要もないし，使いたいとも思わない。

2 パラメータに線形なモデル - 例：フーリエモデル

簡単のため，パラメータに関して線形な関数近似のクラスを考えよう。 このクラスには，多項式，スプライン，カーネル法などのよく知られたモデルが含まれる:

$$y(x) = \sum w_i \phi_i(x) \Leftrightarrow y = w^\top \Phi$$

ここで，y はスカラー出力，w はモデルの未知の重み (パラメータ)，$\phi_(x)$ は固定基底関数，$\Phi_{in} = \phi_i(x^{(n)}), x^{(n)}$ は例数 n の (スカラまたはベクトル) 入力である。 例えば，スカラ入力のフーリエモデルは次のような形になる:

$$y(x) = a_0 + \sum_{d=1}^{D} a_d \sin(d_x) + b_d \cos(dx),$$

ここで $w = \{a_0, a_1, b_1, \ldots, a_D, b_D\}$. 重みに独立なガウス事前分布を仮定する：

$$p(w|S, c) \propto \exp\left(-\frac{S}{2}\left[c_0 a_0^2 + \sum_{d=1}^{D} c_d \left(a_d^2 + b_d^2\right)\right]\right)$$

ここで S は全体のスケールであり，c_d は次数 (頻度 frequency) d の重みの事前精度 (逆分散) である。 重み上のガウス事前分布が，関数上のガウス過程事前分布を含意することを示すのは簡単である (脚注 3)。 対応するガウス過程事前分布の共分散関数は次のようになる:

$$K(x, x') = \left[\sum_{d=0}^{D} \cos\left(x - x'\right)/c_d\right]/S.$$

3. この事前分布の下では，任意の (有限の) 出力 y の集合の結合密度はガウシアンである。

2.1 フーリエモデルにおける推論

精度 τ の独立ガウス雑音があるデータ $\mathcal{D} = \{x^{(n)}, y^{(n)}|n = 1, \ldots, N\}$ が与えられたとき，尤度は次のようになる:

$$p\left(y|x, w, \tau\right) \propto \prod_{n=1}^{N} \exp\left(-\frac{\tau}{2}\left[y^{(n)} - \mathbf{w}^\top \mathbf{\Phi}_n r\right]^2\right).$$

分析の便宜上，事前分布のスケールをノイズの精度に比例する $S = C\tau$ とし，τ と C に曖昧なガンマ事前分布 (脚注 4) を置く:

$$p(\tau) \propto \tau^{\alpha-1} \exp\left(-\beta_1 \tau\right),$$
$$p(C) \propto C^{\alpha_2-1} \exp\left(-\beta_2 C\right),$$

そして，重みとノイズを積分して，事前ハイパーパラメータ C (全体スケール) と $\mathbf{c}$ (相対スケール) の関数としてのエビデンスを得ることができる:

$$E(C, \mathbf{c}) = \int\int p(y|x, w, \tau)\, p(w|C, \tau, \mathbf{c})\, p(\tau) p(C)\, d\tau\, d\mathbf{W} = \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2} \Gamma(\alpha_1 + N/2)}{(2\pi)^{N/2} \Gamma(\alpha_1) \Gamma(\alpha_2)}$$
$$\times |\mathbf{A}|^{1/2} \left[\beta_1 + \frac{1}{2}\mathbf{y}^\top \left(I - \Phi \mathbf{A}^{-1} \Phi^\top\right)\mathbf{y}\right]^{-\alpha_1 - N/2} C^{D+\alpha_2-1/2} \exp\left(-\beta_2 C\right) c_0^{1/2} \prod_{d=1}^{D} c_d,$$

4. ここで，全体を通して $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0.2$ として，曖昧な事前分布を選ぶ。

ここで，$\mathbf{A} = \Phi^\top \Phi + C\text{diag}(\tilde{c})$ であり，チルダは最初の成分以外の重複を示す。 我々は (例えばニュートン法を用いて) 重みの全体的なスケール C を最適化することができる (脚注 5)。 相対尺度 $\mathbf{c}$ はどのように選ぶのか。 この質問に対する答えは，ベイズ推論の 2 つの異なる見解に密接に関係していることがわかる。

2.2 例

このモデルの振る舞いを説明するために，分散 0.25 の独立加法ガウスノイズによって -1 から 1 に変化するステップ関数から生成されたデータを使用する。 真の関数は，現実的なモデリング状況では通常そうであるように，有限次数のモデルで正確に実装することができないことに注意 (真の関数は `realizable` ではなく，モデルは `incomplete` と言われる)。 入力点は 16 点と 8 点の 2 つの塊に配置され，段差は大きい方の真ん中で発生する (図 2参照)。

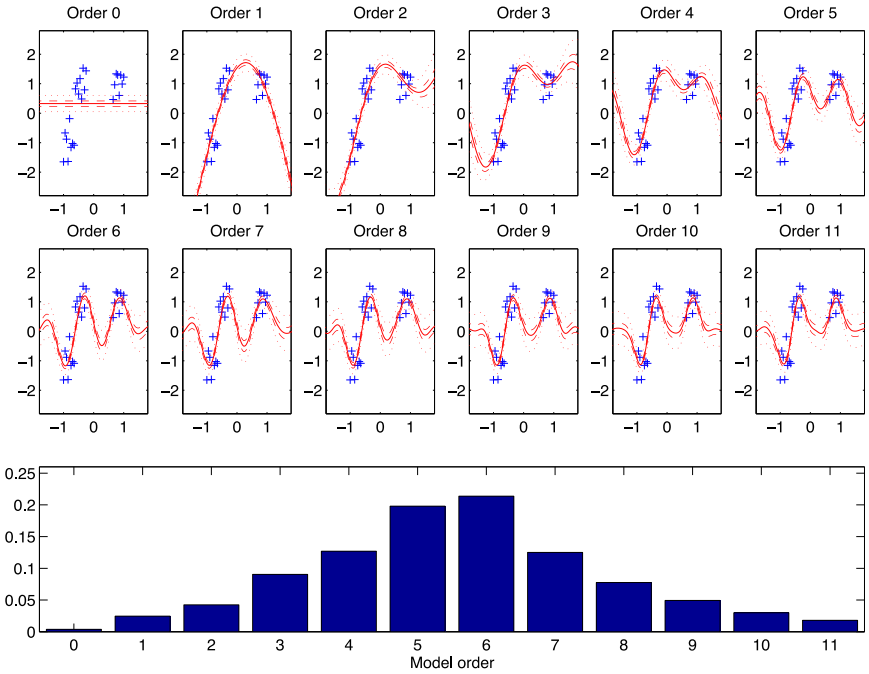


図 2: 上: スケールなし モデルの 12 種類のモデル次数: $c_d \propto 1$. 破線と点線は，予測分布 (student-t である) の 50% と 95% の中心質量 mass を制限する。 下: モデルの事後確率，12 個のモデルで正規化． モデルの確率はオッカムの丘を示し， 小さすぎる または 大きすぎる モデルを抑制している。

スケーリング精度を寄与の頻度に依存しないように選ぶと，$c_d \mapsto 1$（逆精度の和を正規化する）となり，図 2 のような予測が得られる。 明らかにオッカムの剃刀のような振る舞いが見られる。 D=6 程度のモデル次数が望ましい。 限られたデータでは，これより複雑なモデルはサポートされないと言えるかもしれない。 このことを理解する一つの方法は，モデル次数が大きくなるにつれて，事前パラメータ量は大きくなるが，相対的な事後パラメータ量は小さくなることに注目することである。 事前体積と事後体積の比がオッカム係数であり，これはパラメータをフィッティングするために支払うペナルティと解釈できる。

本モデルでは，係数の値を事前分布から引くだけで，関数を事前分布からランダムに引くことは簡単である。 図 3 の左のパネルは，D=6 と D=500 の場合の前の例の事前分布からのサンプルを示している。 次数が高くなるにつれて，関数は高周波数成分によって支配されるようになる。 しかし，ほとんどのモデリング・応用では，我々は滑らかさについてある程度の事前期待値を持っている。 精度係数 c_d をスケーリングすることで，D が無限大になるにつれて，関数に対する事前期待値が特定の特性を持つ関数に収束するようにすることができる。 ここでは，スケーリング指数である γ の異なる値について，$c_d = d^\gamma$ の形のスケーリングに注目する。 例として，スケーリング $c_d = d^3$ を選択した場合，モデルの次数に関してオッカムの剃刀は得られない (図 4)。 次数が十分大きい限り，予測値とその誤差バーはモデルの次数にほとんど依存しないことに注意してください。 また，これらの大きなモデルのエラーバーは，図 2 の D=6 の場合 (2 つのデータの塊の間にスプリアスな ディップ が高い信頼性で予測される) よりも合理的に見えることに注意。 このスケーリングの選択では，「大きなモデル」という見方が適切であると思われる。

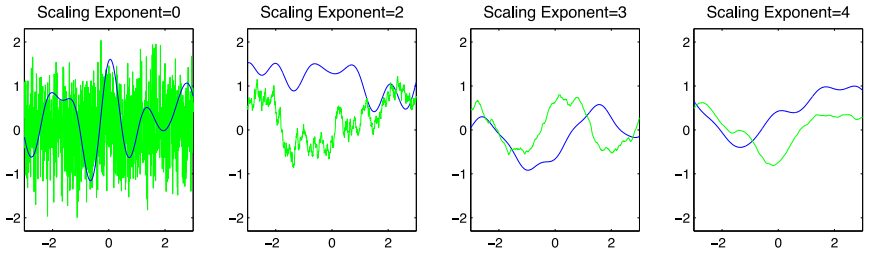


図 3: 4 つの異なるスケーリングについて，次数 D=6 (暗) と D=500 (明) のフーリエモデルからランダムに引き出された関数; 左から右へ: 不連続，ブラウン運動，円滑境界 borderline smooth，滑らか smooth。

5. Of course, we ought to integrate over C, but unfortunately that is difficult.

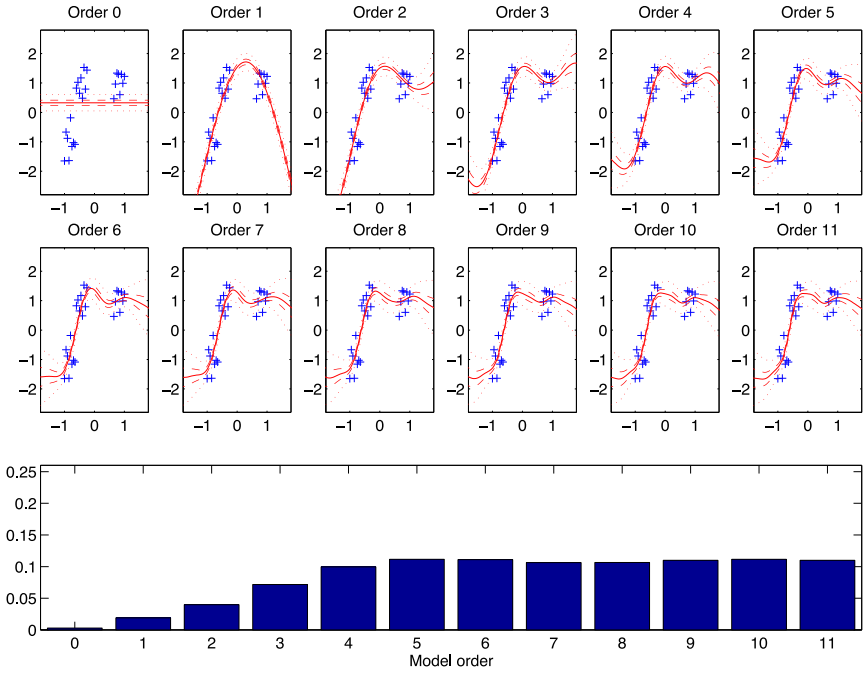


図 4: $c_d = d^3$ のスケーリングが使われ，$D \mapsto \infty$ として滑らかな関数に収束する事前分布を導いた以外は，図 2 と同じ。 オッカムの剃刀 (Occam's Razor) はなく，モデルが十分複雑である限り，証拠は平坦であることがわかる。 また，D が十分大きい限り，モデルの予測密度は変わらない。

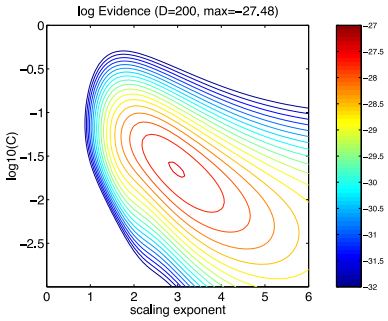
3. 考察

前例では，パラメータに対する事前分布のスケーリング特性によって，オッカムの剃刀の見方と大きなモデルの見方の両方が適切と思われることを見た。 しかし，スケーリング指数 γ の選び方が明らかでないため，この例は満足のいくものではなかった。 D が大きい極限において事前分布から引き出される関数の特性を分析することで，γ の意味についてより深い洞察を得ることができる。 Δ だけ離れた近傍の入力に対応する出力の期待二乗差を考えることは有用である：

$\Delta \mapsto 0$ の極限において，

$$G(\Delta) = \mathbb{E}\left[(f(x) - f(x+\delta))^2\right],$$

図 5 の表では，これらの関数の特徴とともに，様々な γ の値についてこれらの極限を計算した。 例えば，滑らかな関数の性質として $G(\Delta) \propto \Delta^2$ がる。 このような情報を利用することで，実際の応用において γ の良い値を選ぶことができるかもしれない。 実際，我々はデータから「関数の特性」γ を推論することができる。 図 5 では，大きな次数 (D=200) のモデルについて，証拠が γ と全体のスケール C にどのように依存するかを示している。 $\gamma = 3$ 付近で証拠が最大になることがわかる。 実際，我々はオッカムの剃刀を再び見ている！ 今回は，モデルの次元ではなく，異なる γ の値によって暗示される事前分布の下での関数の複雑さの観点からである。 γ の値が大きいと，単純な関数の確率質量が最も大きい事前分布に対応し，γ の値が小さいと，より複雑な関数を許容する事前分布に対応する。 $\gamma = 3$ という 最適な 設定は，まさに図 4 で使用したモデルであることに注意されたい。



γ	$\lim_{\Delta\to 0} G(\Delta)$	properties
≤ 1	1	discontinuous
2	Δ	Brownian
3	$\Delta^2(1-\ln\Delta)$	borderline smooth
>3	Δ^2	smooth

図 5 左: スケーリング指数 γ と全体スケール C の関数としての証拠は，$\gamma = 3$ で最大となる。 表は，γ の異なる値に対する関数の特性を示している。 これらの関数の例を図 3 に示す。

4. 結論

ベイズモデルにおける自動オッカムの剃刀をレビューし，パラメータ数に必ずしもペナルティを課さない一方で，この処理過程が関数の複雑さの点でいかに有効であるかを見てきた。 ここでは単純化された例のみを示したが，動作の説明は一般的に適用可能な非常に基本的な原理に依存している。 2 つの異なるベイズ的見解のうち，どちらが最も魅力的であるかは状況に依存する：時には大きなモデルの限界は計算上厳しいかもしれない。 一方，ノンパラメトリックモデルの典型的な応用では，「大規模モデル」見解が最も便利な事前分布を表現する方法かもしれない。 さらに，連続ハイパーパラメータ上で最適化 (または積分) することは，モデルサイズの離散空間上で最適化するよりも簡単かもしれない。 結局のところ，どのような見解を取るにせよ，オッカムの剃刀は常に働いており，複雑すぎるモデルを抑制しているのである。

Acknowledgements

This work was supported by the Danish Research Councils through the Computational Neural Network Center (CONNECT) and the THOR Center for Neuroinformatics. Thanks to Geoff Hinton for asking a puzzling question which stimulated the writing of this paper.

References

- Jefferys, W. H. & Berger, J. O. (1992) Ockham's Razor and Bayesian Analysis. Amer. Sci., 80:64–72.
- MacKay, D. J. C. (1992) Bayesian Interpolation. Neural Computation, 4(3):415–447.
- Neal, R. M. (1996) Bayesian Learning for Neural Networks, Lecture Notes in Statistics No. 118, New York: Springer-Verlag.
- Rasmussen, C. E. (2000) The Infinite Gaussian Mixture Model, in S. A. Solla, T. K. Leen and K.-R. M¨uller (editors.), Adv. Neur. Inf. Proc. Sys. 12, MIT Press, pp. 554–560.
- Smith, A. F. M. & Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. J. Roy. Stat. Soc., 42:213–220.