P-6

モデル選択における情報量基準について
Informtion criteria in model selection

○椎名颯太 [1], 青柳美輝 [2]
Souta Shiina, Miki Aoyagi

Abstract: In recent studies, real data associated with, for example, image or speech recognition, psychology and economics have been analyzed by the learning system. For the purpose, a lot of learning models have been proposed and then the need of appropriate model selection methods is increasing. In this paper, we show information criteria for model selection methods in Bayesian estimation.

1. 機械学習理論

機械学習とは、得られた何らかのデータに対してその規則性や構造を抽出することにより、未知の現象に対する予測やそれに基づく判断を行うための計算技術の総称である。近年、機械学習は画像認識や音声認識、心理学、経済学など様々な分野に応用されている。また、そういった背景に合わせ機械学習理論では多様な数理モデルが提案されており、得られたデータに対して適切なモデル選択を行う必要がある。本稿では、ベイズ推定の立場に立ち、データを得ることでどのように学習がなされるのか、また適切なモデル選択のための指標とは何か、そしてその近似である情報量基準について述べていく。

2. ベイズ推定

以下, 確率分布 $q(x)dx$ 上の確率変数 $X_1, \cdots, X_n$ は独立であるとする。確率モデルを $p(x|w)$、パラメータ w の事前分布を $\varphi(w)$ とすると、 確率変数の集合 $X^n = \{X_1, \cdots, X_n\}$ に対して, パラメータ w の事後分布は、

$$p(w|X^n) = \frac{1}{Z_n(\beta)}\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta},$$

と定義される。ここで、 β は逆温度と呼ばれるパラメータであり、ベイズ推定では $\beta = 1$ が用いられる。また、$Z_n(\beta)$ は正規化定数とよばれ、

$$Z_n(\beta) = \int dw\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta},$$

を満たす。逆温度 β の事後分布での期待値を $E_w^{\beta}[\cdot]$ と書くことにする。事後分布 $p(w|X^n)$ を用いて、予測分布は、

$$p(x|D_n) = \int dwp(x|w)p(w|X^n) = E_w^{\beta}[p(X|w)],$$

と定義される。

3. ベイズ線形回帰による学習例

平均 $\boldsymbol{m} \in \mathbb{R}^M$、共分散 $\boldsymbol{\Lambda}^{-1} \in \mathbb{R}^{M\times M}$ をもつ M 次元正規分布を $N = (\,\cdot\,|\boldsymbol{m}, \boldsymbol{\Lambda}^{-1})$ とする。機械学習の代表的なタスクに回帰がある。M 次元入力 $\boldsymbol{\phi}(x) = (\phi_0(x), \cdots, \phi_{M-1}(x)) \in \mathbb{R}^M$ から出力 $y \in \mathbb{R}$ を予測するような関数 $y = f(\phi(x))$ をデータから求めるタスクである。例えば関数 ϕ_j は $\phi_j(x) = x^j$ を満たすもので、基底関数と呼ばれる。n 個の入力データを $\boldsymbol{X} = \{\boldsymbol{\phi}(x_1), \cdots, \boldsymbol{\phi}(x_n)\}$、対応する n 個の実数値出力データを $\boldsymbol{Y} = \{y_1, \cdots, y_n\}$ とする。関数の形状を決めるための M 次元パラメータを $\boldsymbol{w} \in \mathbb{R}^M$ とすれば、次のようにモデル化できる。

$$\text{多項式回帰}\quad y_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(x_i) + \varepsilon_i.$$

ここで，ε_i は平均 0、分散 $\lambda^{-1} \in \mathbb{R}$ の 1 次元正規分布 $N(\varepsilon_i|0, \lambda^{-1})$ に従う正規ノイズとする。

よって y_i の確率密度関数は次のようになる。

$$p(y_i|x_i, \boldsymbol{w}) = N(y_i|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(x_i), \lambda^{-1}).$$

また、$\boldsymbol{w}$ の事前分布を次のように定める：

$$\varphi(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{\Lambda}^{-1}).$$

以上よりパラメータ $\boldsymbol{w}$ に関する事後分布は次のように求められる。

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{Y}) &= \frac{\varphi(\boldsymbol{w})\prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{w})}{p(\boldsymbol{Y}|\boldsymbol{X})} \\ &= N(\boldsymbol{w}|\hat{\boldsymbol{m}}, \hat{\boldsymbol{\Lambda}}^{-1}). \end{aligned}$$

$$\begin{aligned} \text{ただし、}\ \hat{\boldsymbol{\Lambda}} &= \lambda\sum_{i=1}^{n}\boldsymbol{\phi}(x_i)\boldsymbol{\phi}(x_i)^{\mathrm{T}} + \boldsymbol{\Lambda}, \\ \hat{\boldsymbol{m}} &= \hat{\boldsymbol{\Lambda}}^{-1}(\lambda\sum_{i=1}^{n} y_i\phi(x_i) + \boldsymbol{\Lambda}\boldsymbol{m}). \end{aligned}$$

パラメータの事前分布を決めるハイパーパラメータ $\boldsymbol{m}, \boldsymbol{\Lambda}$ に入出力の組 $(\boldsymbol{X}, \boldsymbol{Y})$ の情報を加え、元の正規分布の形状を更新した形になっている (ベイズ更新という)。したがって、

$$\mu_* = \hat{\boldsymbol{m}}^{\mathrm{T}}\boldsymbol{\phi}(x),\ \ {\lambda_*}^{-1} = \lambda^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\phi}(x),$$

に対し，

1: 日大理工・院 (前)・数学　2: 日大理工・教員・数学

予測分布　$p(y|x,\boldsymbol{X},\boldsymbol{Y}) = E^1_{\boldsymbol{w}}[p(X|\boldsymbol{w})] = N(|\mu_*, {\lambda_*}^{-1}),$

が得られる。

4. 情報量基準

ベイズ線形回帰において入力 $x \in \mathbb{R}$ を変換 $\boldsymbol{\phi}(x)$ を用いて M 次元入力ベクトルに変換した。モデルの次数 M が小さいと、単純なグラフしか描けず得られたサンプルを十分に説明できているとは言い難い。大きすぎると、モデルのサンプルに対する表現力は増すが、過学習という予測能力が低い状態に陥ってしまう。このようなことが起きないモデルを選ぶための指標、つまり真の分布 $q(x)$ に対して最適なモデル $p(x|w)$ と事前分布 $\varphi(w)$ を選ぶ指標は主に汎化誤差と自由エネルギーの二つが挙げられる。

定義 1 (ベイズ汎化誤差)

$$\begin{aligned} B_g(n) &= \int q(x)\log\frac{q(x)}{p(x|X^n)}dx \\ &= -\int q(x)\log p(x|X^n)dx + \int q(x)\log q(x)dx \\ &= B_gL - S, \end{aligned}$$

を汎化誤差といい、真の分布 $q(x)$ と予測分布 $p(x|X^n)$ のカルバック情報量で定義される。さらに、汎化誤差は真の分布に依存するエントロピー S と汎化損失 B_gL と呼ばれる部分に分解できる。よって、汎化損失が小さいほど真の分布と予測分布との違いが小さくなり、そのような組 $(p(x|w), \varphi(w))$ を選ぶための指標である。

定義 2 (ベイズ自由エネルギー) ベイズ自由エネルギーは、

$$F_n(1) = -\log Z_n(1),$$

と定義される。$Z_n(1)$ は周辺尤度と呼ばれ、得られたデータサンプル X^n に対する組（$p(x|w), \varphi(w)$）の尤もらしさを表している。真の分布 $q(x)$ と周辺尤度 $Z_n(1)$ のカルバック情報量は、自由エネルギーを真の分布で期待値をとった項を含んでいる。よって、自由エネルギーが小さいほど最適な組であるといえる。

通常得られるのはデータ X^n だけであり、その分布 $q(x)$ は知ることは出来ないので、汎化誤差、平均自由エネルギーは計算が出来ない。また、自由エネルギー自体も複雑なモデルを用いると解析的な計算が困難になる。そこで、サンプルのみからこれら 2 つの量を近似するために作られたのが、情報量基準といわれるものである。

定義 3 (WAIC) 経験損失 T_n、汎関数分散 V_n はそれぞれ

$$T_n = -\frac{1}{n}\sum_{i=1}^{n} E^\beta_w[\log p(X_i|w)],$$

$$V_n = \sum_{i=1}^{n}\{E^\beta_w[(\log p(X_i|w))^2] - E^\beta_w[\log p(X_i|w)]^2\},$$

と定義される。このとき、広く使える情報量基準（WAIC）は

$$\mathrm{WAIC}(n) = T_n + \frac{\beta V_n}{n},$$

と定義される。

定義 4 (WBIC) 経験対数損失関数を

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i|w),$$

と定義する。このとき広く使えるベイズ情報量基準（WBIC）は

$$\mathrm{WBIC}(n) = E^\beta_w[nL_n(w)],$$
$$\beta = \frac{1}{\log n},$$

と定義される。

定義 5 (LOOCV、CV) サンプル X^n からあるデータ X_i を除いたサンプルを $X^n \backslash X_i$ と表す。このとき、クロスバリデーション損失 (LOOCV) は

$$\mathrm{CV}(n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i|X^n\backslash X_i),$$

と定義される。

定理 6 $E[\cdot]$ はサンプル X_n 全体での期待値を表す。このとき、

$$E[B_gL(n)] = E[\mathrm{WAIC}(n)] + o(\frac{1}{n}),$$
$$E[B_gL(n-1)] = E[\mathrm{CV}(n)],$$

が成立する。

定理 7 任意の逆温度 $0 < \beta < \infty$ に対して

$$\mathrm{CV}(n) = \mathrm{WAIC}(n) + O_p(\frac{1}{n^{\frac{3}{2}}}).$$

特に $\beta = 1$ のとき

$$\mathrm{CV}(n) = \mathrm{WAIC}(n) + O_p(\frac{1}{n^2}).$$

定理 8

$$F_n(1) = \mathrm{WBIC}(n) + O_p(\sqrt{\log n}).$$

5. 参考文献

[1] S. Watanabe:“ Mathematical Theory of Bayesian Statistics ”, CRC Press, New York, USA, 2018.

[2] S. Watanabe:“ A Widely Applicable Bayesian Infomation Criterion ”, Journal of Machine Learning Research 14 867-897, 2013

[3] 須山敦志:“ ベイズ推論による機械学習入門 ”, 講談社,2018