階層ベイズ法とWAIC

渡辺澄夫
東京工業大学

概 要

階層ベイズ法における WAIC[1] の使いかた２通りを説明します。情報量規準は考察しているモデルの予測能力を測るものなので、予測するものが異なれば規準の作り方も変わります。

1 階層ベイズ法

例. 考察する問題を確認するため例をあげます。
ある高校に 1 年生のクラスが $m = 10$ クラスあり、全てのクラスで生徒の数は $n = 30$ であるとします。この高校で英語のテストを行い $mn = 300$ 人の点数が得られました。次のような統計モデルを考えました。
(1) 第 k クラスの平均点 w_k は、平均 μ 分散 1^2 の正規分布に従う。
(2) 第 k クラスの各生徒の点数 x_{ki} は、平均 w_k 分散 10^2 の正規分布に従う。
モデルは真の分布と同じとは限らないので、このモデルの適切さを測りたいと思いました。

考察する問題の統計的記述.
クラスは m 個。それぞれのクラスで n 人のデータが得られる。
(1) ハイパーパラメータを μ とする。
(2) 事前分布 $\varphi(w|\mu)$ からパラメータ $\{w_k\}_{k=1}^m$ が独立に生成される。
(3) 確率分布 $p(x|w_k)$ から第 k クラスのデータ $(x_k)^n \equiv \{x_{ki}\}_{i=1}^n$ が独立に生成される。

2 二つの予測

この問題には二つの異なる予測を考えることができます。
(第 1 の問題) 第 k クラスに新しい一個のデータが発生するときの予測です。
(第 2 の問題) クラス一個を新しく生成して n 人のデータを発生するときの予測です。
WAIC は予測を行うときの汎化損失を基礎とする規準ですから、予測するものが異なれば規準も異なります。(AIC もクロスバリデーションも同様です)。

3 第 1 の問題：各クラスの新しいデータの予測

まず第 1 の問題を考えます。全データ $\{(x_k)^n\}$ が与えられたときの $(w_1, w_2, ..., w_m)$ の事後分布は

$$p(w_1, w_2, ..., w_m|(x_1)^n, (x_2)^n, ..., (x_m)^n) \propto \prod_{k=1}^m \Big(\varphi(w_k|\mu) \prod_{i=1}^n p(x_{ki}|w_k)\Big)$$

となります。これは $(w_1, w_2, ..., w_m)$ について独立ですから、それぞれの事後分布は

$$p(w_k|x_k^n) \propto \varphi(w_k|\mu) \prod_{i=1}^{n} p(x_{ki}|w_k)$$

です。この事後分布による平均を $\mathbb{E}_{w_k}[\ ]$ 分散を $\mathbb{V}_{w_k}[\ ]$ と書くことにします。第一の問題では新しいデータ y の予測分布は $\mathbb{E}_{w_k}[p(y|w_k)]$ になるので、第 k クラスの WAIC は

$$WAIC_k \quad = \quad -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_{w_k}[p(x_{ik}|w_k)] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_{w_k}[\log p(x_{ik}|w_k)] \tag{1}$$

になります。全クラスでひとつずつデータが増えるときの予測を同時に行った場合の誤差の総和を知りたい場合は、これの和が求めるものになります。WAIC の最小化により μ を定めると、第 1 の問題の意味で適切にハイパーパラメータを定めたことになります。$p(x_{ik}|w_k)$ が正則であり真の分布を含んでいるときには、左辺第 2 項はおおよそ「w_k の次元$/n$」になります。その置き換えを行いさらに左辺第 1 項の $\mathbb{E}_{w_k}[\ ]$ を最尤推定量で置き換えれば、WAIC は AIC になります。なお、AIC や DIC の通常の定義とスケールを合わせたいときは、式 (1) を $2n$ 倍してください。

4　第 2 の問題：新しいクラスができるときの予測

第 2 の問題では、ひとつのサンプルに相当するものが $(x_k)^n$ であって、m 個のサンプルが得られていることになります。つまり、パラメトリックモデル

$$\mathcal{P}((x_k)^n|\mu) = \int \varphi(w|\mu) \prod_{i=1}^{n} p(x_{ki}|w) dw$$

を考えることになります。第 2 の問題を考える時にはハイパーパラメータがパラメータになります。ベイズ推測を行うためには μ についての事前分布 $\psi(\mu)$ を設定して事後分布

$$p(\mu|(x_1)^n, (x_2)^n, ..., (x_m)^n) \propto \psi(\mu) \prod_{k=1}^{m} \mathcal{P}(x_k^n|\mu)$$

を作ります。この事後分布についての平均操作を $\mathbb{E}_{\mu}[\ ]$ 分散を $\mathbb{V}_{\mu}[\ ]$ と書くことにします。第 2 の問題では新しいクラス全体のデータ y^n の予測分布は $\mathbb{E}_{\mu}[\mathcal{P}(y^n|\mu)]$ になるので、WAIC は

$$WAIC \quad = \quad -\frac{1}{m}\sum_{k=1}^{m} \log \mathbb{E}_{\mu}[\mathcal{P}((x_k)^n|\mu)] + \frac{1}{m}\sum_{k=1}^{m} \mathbb{V}_{\mu}[\log \mathcal{P}((x_k)^n|\mu)] \tag{2}$$

になります。第 2 の問題では μ は事後分布により推定されています。$\mathcal{P}((x_k)^n|\mu)$ が正則なモデルであり真の分布を含んでいるときには、左辺第 2 項はおおよそ「μ の次元$/m$」になります。その置き換えを行いさらに左辺第 1 項の $\mathbb{E}_{\mu}[\ ]$ を最尤推定量で置き換えれば、WAIC は AIC になります。AIC や DIC の通常の定義とスケールを合わせたいときは、式 (2) を $2m$ 倍してください。

注意. 第 1 、第 2 のどちらの問題とも異なる予測を考えたい時には、その予測に対応する WAIC を用いてください。AIC もクロスバリデーションも同様です。

参考文献

[1] 渡辺澄夫、ベイズ統計の理論と方法、コロナ社、2012