A Widely Applicable Bayesian Information Criterion

Ernest Nasyrov

MIPT

October 14, 2024

Problem

- Many statistical models are not regular, but singular. Such as ANNs, normal mixtures, Bayesian networks, HMMs.
- If a model is singular, then the likelihood function cannot be approximated by any normal distribution.
- So, neither AIC, BIC, nor MDL can be used in statistical model evaluation.

The proposal of the paper

WBIC has the same asymptotic behavior as the BIC. WBIC is an extension of BIC on singular models. WBIC can be calculated without any information about true distribution.

BIC

Let us define $p(x|w)$ — pdf of $x \in \mathbb{R}^N$, for a given parameter $w \in W \subset \mathbb{R}^d$.
Prior density function — $\phi(w)$ on W. $\mathbf{X}_1, \dots, \mathbf{X}_n \sim$ iid $q(x)$ — *true distribution*.
Log loss function:

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{X}_i|w)$$

Bayes free energy $\mathcal{F}$ can be unbserstood as *the minus logarithm of marginal likelihood of a model and a prior*:

$$\mathcal{F} \approx -\log \int \prod_{i=1}^{n} p(\mathbf{X}_i|w)\phi(w)dw$$

BIC

If a statistical model is *regular*, then the posterior distribution can be asymptotically approximated by a *normal distribution*, resulting that

$$\mathcal{F} = nL_n(\hat{w}) + \frac{d}{2}\log n + O(1) = BIC$$

In *singular* models the Bayes free energy cannot be approximated by BIC in general. But it was proved, that

$$\mathcal{F} = nL_n(w_0) + \lambda \log n + O(\log\log n)$$

Where w_0 — is a parameter, minimizing the $KL(q(x)||p(x|w))$ and $\lambda > 0$ called **real log canonical treshhold (RLCT)**.

There's a dog buried here.

WBIC

Authors propose to estimate $\mathcal{F}$ **without any information about a true distribution**. So, the WBIC is defined as

$$WBIC = \mathbb{E}_w^\beta \left[nL_n(w) \right], \quad \beta = \frac{1}{\log n}$$

Where $\mathbb{E}_w^\beta[]$ is the expectation over the *posteriori* distribution on W:

$$\mathbb{E}_w^\beta \left[f(w) \right] = \frac{\int f(w) \prod\limits_{i=1}^{n} p(\mathbf{X}_i|w)^\beta \phi(w) dw}{\int \prod\limits_{i=1}^{n} p(\mathbf{X}_i|w)^\beta \phi(w) dw}$$

Parameter $\beta > 0$ is called *inverse temperature*.

The main result of the paper

$$\mathcal{F} = WBIC + O(\sqrt{\log n})$$

Proof. Theorem 3

Theorem 3 (Unique Existence of the Optimal Parameter)

1. The value $\mathbb{E}_w^\beta\left[nL_n(w)\right]$ is a *decreasing* function of β.
2. There exists a unique $\beta^* \in (0,1)$ which satisfies

$$\mathcal{F} = \mathbb{E}_w^\beta\left[nL_n(w)\right]$$

And that β^* is a random variable, which satisfies convergence in probability: $\beta^* \log n \to_{n\to\infty} 1$

Proof. Theorem 4

Theorem 4 (Main Theorem)

Assume, that

$$\beta = \frac{\beta_0}{\log n}$$

Then there exists a random variable U_n:

$$\mathbb{E}_w^{\beta}\left[nL_n(w)\right] = nL_n(w_0) + \frac{\lambda \log n}{\beta_0} + U_n\sqrt{\frac{\lambda \log n}{2\beta_0}} + O(1)$$

So, WBIC has the same asymptotic behaviour as the Bayes free energy (substituting $\beta_0 = 1$):

$$WBIC = nL_n(w_0) + \lambda \log n + O(\sqrt{\log n})$$

Proof. Theorem 5

Theorem 5

If a statistical model is regular, then

$$WBIC = nL_n(\hat{w}) + \frac{d}{2}\log n + O(1)$$

Example

Example

Reduced rank regression model is studied:

$$p(x, y|w) = \frac{r(x)}{(2\pi\sigma^2)^{N/2}} exp\left(-\frac{1}{2\sigma^2}||y - BAx||^2\right)$$

$x \in \mathbb{R}^M, y \in \mathbb{R}^N, w = (A, B), A : H \times M, B : N \times H.$

H	1	2	3	4	5	6
WBIC_1 Ave.	17899.82	3088.90	71.11	78.21	83.23	87.58
WBIC_1 Std.	1081.30	226.94	3.67	3.78	3.97	4.09
WBIC_2 Ave.	17899.77	3089.03	71.18	75.43	82.54	86.83
WBIC_2 Std.	1081.30	226.97	3.54	3.89	4.03	4.08
BIC Ave.	17899.77	3089.03	71.18	83.47	91.86	94.87
BIC Std.	1081.30	226.97	3.54	3.89	4.03	4.08

Table 2: WBIC and BIC in Model Selection

Рис.:

Watanabe, Sumio (2013)
A widely applicable Bayesian information criterion. The Journal of Machine Learning Research, 14(1), 867-897.