Bayesian Information Criterion

Konstantin Yakovlev

MIPT, 2022

October 4, 2022

Definition

Definition of BIC

Given a finite set of observed data $\mathbf{X} = \{X_i\}_{i=1}^n$. Given a model with its parameters $\hat{\theta}$ that maximizes $p(\mathbf{X}|\theta)$ w.r.t. θ. Then BIC is defined as follows:

$$\mathrm{BIC} = \dim\hat{\theta} \log n - 2 \log p(\mathbf{X}|\hat{\theta}).$$

Derivation 1

Problem statement

Given finite set of models $\{M_i\}_{i=1}^m$ with its parameters $\{\boldsymbol{\theta}_i\}_{i=1}^m$ and likelihoods $\{p(\mathbf{X}|\boldsymbol{\theta}_i, M_i)\}_{i=1}^m$. The task is to perform model selection that maximizes posterior probability $p(M_i|\mathbf{X})$.

Write down the posterior probability of M_i:

$$p(M_i|\mathbf{X}) = \frac{p(\mathbf{X}|M_i)p(M_i)}{p(\mathbf{X})}.$$

Write down the likelihood:

$$p(\mathbf{X}|M_i) = \int p(\mathbf{X}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i = \int \exp(\log(p(\mathbf{X}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i)))d\boldsymbol{\theta}_i.$$

Derivation 2

Using taylor's formula, expand $\log(p(\mathbf{X}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i))$ around MAP estimate $\boldsymbol{\theta}_i^{MAP}$:

$$\underbrace{\log(p(\mathbf{X}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i))}_{Q(\boldsymbol{\theta}_i)} \approx \log(p(\mathbf{X}|\boldsymbol{\theta}_i^{MAP}, M_i)p(\boldsymbol{\theta}_i^{MAP})) +$$

$$\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{MAP})^\top \mathbf{H}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{MAP}),$$

where $\mathbf{H}_{\boldsymbol{\theta}_i} = \frac{\partial^2 Q(\boldsymbol{\theta}_i^{MAP})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^\top}$. Let $p(M_i) = \frac{1}{m}\ \forall\ i = \overline{1, m}$. Then $\boldsymbol{\theta}_i^{MAP} = \hat{\boldsymbol{\theta}}_i$.

Now consider:

$$-H_{kl} = -\frac{\partial^2 \log(\prod_{j=1}^n p(x_j|\hat{\boldsymbol{\theta}}_i))}{\partial(\theta_i)_k \partial(\theta_i)_l} = -\frac{1}{n}\sum_{j=1}^n \underbrace{\frac{\partial^2 n \log p(x_j|\hat{\boldsymbol{\theta}}_i)}{\partial(\theta_i)_k \partial(\theta_i)_l}}_{\xi_j}.$$

Using the weak law of large numbers:

$$\frac{1}{n}\sum_{j=1}^n \xi_j \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\xi_1.$$

Derivation 3

Then:

$$-H_{kl} \xrightarrow[n\to\infty]{\mathbb{P}} -n\mathbb{E}_x \frac{\partial^2 \log p(x|\hat{\boldsymbol{\theta}}_i)}{\partial(\theta_i)_k \partial(\theta_i)_l} = nI(\hat{\boldsymbol{\theta}}_i)_{kl}.$$

Laplace approximation of $p(\mathbf{X}|M_i)$ for $n \gg \dim \boldsymbol{\theta}_i$ gives:

$$\log p(\mathbf{X}|M_i) \approx \log p(\mathbf{X}|\hat{\boldsymbol{\theta}}_i) + \log p(\hat{\boldsymbol{\theta}}_i) + \frac{\dim \boldsymbol{\theta}_i (\log 2\pi - \log n) - \log |\mathbf{I}(\hat{\boldsymbol{\theta}}_i)|}{2}.$$

For $n \gg \dim \boldsymbol{\theta}_i$ ignore terms that does not depend on n:

$$\log p(\mathbf{X}|M_i) \approx \log p(\mathbf{X}|\hat{\boldsymbol{\theta}}_i) - \frac{\dim \theta_i}{2} \log n = -\frac{1}{2}\mathrm{BIC}_i.$$

Model Selection

Model selection

Let $n \gg \dim \boldsymbol{\theta}_i$, $i = \overline{1, m}$. Let also $p(M_i) = \frac{1}{m}$. Then

$$\arg\max_i p(M_i|\mathbf{X}) = \arg\min_i \mathrm{BIC}_i.$$

Remark: In the construction of BIC, the effect of priors are ignored since we are working on the limiting regime but we still use the Bayesian evidence as a model selection criterion. We are selecting the model with the highest evidence. When the data is indeed generated from one of the model in the collection of models we are choosing from, the posterior will concentrate on this correct model. So BIC would eventually be able to select this model.

Literature

1. **Derivation** On the derivation of the Bayesian Information Criterion.
2. **Remark** Lecture 7: Model Selection and Prediction.