Model Selection — Bayesian Information Criterion

PkGu

10/29/2021

Viewpoint from Homo-Bayesianis

The backdrop: We have a bunch of alternative models: $\mathcal{M}_i$, and each model gives a parameter space Θ and a setting for generation of data $p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)$. For comparing fidelity of different models, under Bayesian principle, we should use $p(\mathcal{D}|\mathcal{M}_i) \propto p(\mathcal{M}_i|\mathcal{D})$, assuming the prior for different $\mathcal{M}_i$ are equal. The previous $p(\mathcal{M}_i|\mathcal{D})$ is called *model evidence*. By Bayes' formula, the evidence is obtained by:

$$\begin{aligned} p(\mathcal{D}|\mathcal{M}_i) &= \int p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)\pi(\boldsymbol{\theta}_i|\mathcal{M}_i))d\boldsymbol{\theta} \\ &=: \int f_{\mathcal{M}_i;\mathcal{D}}(\boldsymbol{\theta}_i)d\boldsymbol{\theta} \end{aligned}$$

So if the integration is hard to compute, it's reasonable to assume that $f_{\mathcal{M}_i;\mathcal{D}}(\boldsymbol{\theta}_i)$ as a function of $\boldsymbol{\theta}_i$ is close to a p.d.f. of a normal distribution [a homo-frequentitus will explain it by asymptotic normality], so by Laplace approximation near the MAP point $\hat{\boldsymbol{\theta}}_i$, the integration is approximately decided by the Hessian matrix of $log\ f_{\mathcal{M}_i;\mathcal{D}}$ at the MAP, as follows:

$$log\ p(\mathcal{D}|\mathcal{M}_i) = log\ p(\mathcal{D}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i) + log\ \pi(\hat{\boldsymbol{\theta}}_i|\mathcal{M}_i) + \frac{k}{2}log(2\pi) - \frac{1}{2}log|\mathbf{A}|$$

where k represents the dimension of the i-th parameter space and A is the negative Hessian matrix at $\hat{\boldsymbol{\theta}}_i$:

$$\begin{aligned} \mathbf{A} &= -\nabla^2\Big|_{\hat{\boldsymbol{\theta}}_i} log\ p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)\pi(\boldsymbol{\theta}_i|\mathcal{M}_i)) \\ &= -\nabla^2\Big|_{\hat{\boldsymbol{\theta}}_i} \Big[log\ p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i) + log\ \pi(\boldsymbol{\theta}_i|\mathcal{M}_i))\Big] \end{aligned}$$

The last 3 terms of RHS are comprised as the penalization term against complexity, called "Occam factor", while the first term of RHS describes how well can a prediction under the model fit the given data.As N increases far larger than k: the first 2 terms of Occam factor $log\ \pi(\hat{\boldsymbol{\theta}}_i|\mathcal{M}_i) + \frac{k}{2}log(2\pi)$ can be relatively ignored and, for the last term:

$$\begin{aligned} |det(\mathbf{A})|^{\frac{1}{2}} &\sim \Big| -\nabla^2|_{\hat{\boldsymbol{\theta}}_i} log\ p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)\Big|^{\frac{1}{2}} \\ &\sim \Big| -\nabla^2|_{\hat{\boldsymbol{\theta}}_i} N \cdot log\ p(x_{typical}|\boldsymbol{\theta}_i, \mathcal{M}_i)\Big|^{\frac{1}{2}} \\ &\sim O(N^{\frac{k}{2}}) \end{aligned}$$

So we can approximately estimate $log\ p(\mathcal{D}|\mathcal{M}_i)$ by

$$log\ p(\mathcal{D}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i) - \frac{k}{2} logN$$

which we should maximize among all models.

The BIC (which we should minimize, like AIC) is formally defined as

$$BIC = K\ logN - 2\ log\ p(\mathcal{D}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i)$$

However, though we've already seen that BIC penalizes heavier than AIC especially when N is large, but the reason of this difference is intriguing. Recall that the motivation of the AIC and its approach is to minimize the KL-distance of the distribution subordinate to the MLE restricted in the limited sub-model to the "real" distribution, *i.e.* optimization of the prediction effect. But the approach of BIC requires the optimization of the model evidence, the "marginal likelihood", or in another word, it pursues the most reasonable model for generating the given "training" data. As N going large, the concentration of posterior distribution of parameter at the MAP rises, that leads to the rising of the penalization term.