Information criteria for non-normalized models

Takeru Matsuda
The University of Tokyo

Suppose we have N samples $x_1, \cdots, x_N$ from a parametric distribution

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta),$$

where θ is an unknown parameter and $Z(\theta)$ is the normalization constant. For several statistical models, only the non-normalized density $\widetilde{p}(x \mid \theta)$ is given and the calculation of $Z(\theta)$ is intractable. Thus, several methods have been developed to estimate θ without explicitly computing $Z(\theta)$. Here, we focus on noise contrastive estimation [3].

In noise contrastive estimation (NCE), the non-normalized model is rewritten as

$$\log p(x \mid \theta, c) = \log \widetilde{p}(x \mid \theta) + c,$$

where the scalar $c = -\log Z(\theta)$ is also viewed as an unknown parameter and estimated from data. In addition to data $x_1, \cdots, x_N$, we generate M noise samples $y_1, \cdots, y_M$ from a noise distribution $n(y)$. Then, the estimate of (θ, c) is defined by learning to discriminate between the data and the noise as accurately as possible:

$$(\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}) = \arg \max_{\theta, c} \hat{J}_{\mathrm{NCE}}(\theta, c),$$

where

$$\hat{J}_{\mathrm{NCE}}(\theta, c) = \sum_{t=1}^{N} \log \frac{N p(x_t \mid \theta, c)}{N p(x_t \mid \theta, c) + M n(x_t)} + \sum_{t=1}^{M} \log \frac{M n(y_t)}{N p(y_t \mid \theta, c) + M n(y_t)}.$$

The objective function $\hat{J}_{\mathrm{NCE}}$ is the log-likelihood of the logistic regression classifier. NCE has consistency and asymptotic normality under mild regularity conditions. Recently, NCE was extended to estimate a finite mixture of non-normalized models [4].

In this study, we derive information criteria for models estimated by NCE. Based on an observation that NCE is a projection with respect to a Bregman divergence [2], we develop an approximately unbiased estimator of model discrepancy induced by this Bregman divergence. Note that AIC [1] was derived as an approximately unbiased estimator of the Kullback-Leibler discrepancy. Experimental results demonstrate that the proposed criterion is useful for selection of non-normalized models. For example, it can be used for selecting the number of components in non-normalized mixture models.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723, 1974.

[2] M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

[3] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research* **13**, 307–361, 2012.

[4] T. Matsuda and A. Hyvärinen. Estimation of non-normalized mixture models and clustering using deep representation. arXiv:1805.07516.