Subspace Information Criterion
— Unbiased Generalization Error Estimator for Linear Regression

Masashi Sugiyama   Hidemitsu Ogawa

Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.

sugi@og.cs.titech.ac.jp

We will report a new method for generalization error estimation: the *subspace information criterion* [5].

Let us consider a regression problem. The learning target and learning result functions are denoted by $f(x)$ and $\hat{f}(x)$, respectively, and given training examples are denoted by $\{(x_m, y_m) \mid y_m = f(x_m) + \epsilon_m\}_{m=1}^{M}$, where ϵ_m is an additive noise. The generalization error is measured by using the probability distribution $p(x)$ of *future* (*unlabeled*) input points as

$$J_G = E_\epsilon \int \left(\hat{f}(x) - f(x)\right)^2 p(x)dx. \qquad (1)$$

Here E_ϵ denotes the expectation over the noise and the integral with respect to x means the expectation over future input points.

The methods for generalization error estimation devised so far can be roughly classified into three approaches. The first approach is that $p(x)$ is replaced with the empirical distribution of $\{x_m\}_{m=1}^{M}$. This approach includes C_P [3], generalized cross-validation [2], and also traditional cross-validation. These methods often work well in practice, but the theoretical mechanism for small sample cases is not still clear. For the second approach, the asymptotic approximation is used for estimating the expectation of J_G over training input points $\{x_m\}_{m=1}^{M}$. This approach includes Akaike's information criterion (AIC) [1] and its various derivatives. Although it is theoretically shown that AIC gives an asymptotic unbiased estimate of the generalization error, it does not work well with small samples. For the third approach, e.g. VC-theory [6], a probabilistic upper bound of the generalization error is evaluated. Although it does not depend on the number of training examples, the tightness of the bound is not sure.

In this talk, we will give a method for generalization error estimation called the *subspace information criterion (SIC)*. The feature of SIC is that it directly evaluates the generalization error J_G itself and it gives an exact unbiased estimate of J_G. Therefore, SIC works well with small samples. Moreover, SIC does not impose any restriction on models while AIC-type criteria and VC-theory are valid only for nested models.

For achieving the above advantages, SIC assumes some conditions. Here we consider the simplest case for describing the main idea of SIC.

1. $f(x)$ belongs to a specified, finite dimensional functional Hilbert space H. The norm in H is defined as
$$\|f\|^2 = \int (f(x))^2 p(x)dx, \qquad (2)$$
where $p(x)$ is the distribution of future input points, which is assumed to be known. In this case, the generalization error J_G is expressed by using the norm as
$$J_G = E_\epsilon \|\hat{f} - f\|^2. \qquad (3)$$

2. A learning result function $\hat{f}(x)$ is obtained by
$$\hat{f} = Xy, \qquad (4)$$
where X is a linear operator from $\mathbf{R}^M$ to H and $y = (y_1, y_2, \ldots, y_M)^\top$. Eq.(4) includes usual *least mean squares (LMS) learning* and *regularization learning with quadratic regularizers* for linear regression models expressed as
$$\hat{f}(x) = \sum_{p=1}^{\mathrm{rank}X} \theta_p \varphi_p(x), \qquad (5)$$
where $\varphi_p(x)$'s are prefixed basis functions in the range of X.

3. The number M of training examples is larger than the dimension of H.

4. A linear operator X_u which gives an unbiased estimate $\hat{f}_u$ of f is available:
$$E_\epsilon \hat{f}_u = f \quad : \quad \hat{f}_u = X_u y. \qquad (6)$$

Usually, X_u can be obtained by LMS learning in H.

5. The mean noise is zero and the noise covariance matrix is in the form $\sigma^2 I$ with $\sigma > 0$, where I is the identity matrix on $\mathbf{R}^M$.

The generalization error J_G can be decomposed into the bias and variance:

$$J_G = \|E_\epsilon \hat{f} - f\|^2 + \sigma^2 \mathrm{trace} X X^\top. \tag{7}$$

Although the variance can be easily estimated by $\hat{\sigma}^2 \mathrm{trace} X X^\top$ with

$$\hat{\sigma}^2 = \frac{\sum_{m=1}^{M} (\hat{f}_H(x_m) - y_m)^2}{M - \dim H}, \tag{8}$$

the bias is generally difficult to estimate since $E_\epsilon \hat{f}$ and f are both unknown. Our idea is that unbiased estimates $\hat{f}$ and $\hat{f}_u$ of $E_\epsilon \hat{f}$ and f are respectively available, so the bias is at first roughly approximated by $\|\hat{f} - \hat{f}_u\|^2$ and it is then corrected. Based on the idea, SIC is given as

$$\begin{aligned} \mathrm{SIC}[X] \; = \; & \|\hat{f} - \hat{f}_u\|^2 \\ & - \hat{\sigma}^2 \mathrm{trace}(X - X_H)(X - X_H)^\top \\ & + \hat{\sigma}^2 \mathrm{trace} X X^\top, \end{aligned} \tag{9}$$

where the second term is the bias correction term. The goodness of SIC as a generalization error estimator is theoretically substantiated by the fact that under the explained conditions, it is an unbiased estimate of J_G for any X:

$$E_\epsilon \mathrm{SIC} = J_G. \tag{10}$$

By using SIC, model selection for LMS learning is performed as follows.

1. A set $\mathcal{M}$ of subspaces $\{S\}$ of H is prepared.
2. $\mathrm{SIC}[X_S]$ is calculated for all $S \in \mathcal{M}$, where $X_S y$ is the LMS learning result function in S.
3. The model $\hat{S}$ which minimizes SIC is selected.

We experimentally compared SIC with the network information criterion (NIC) [4], which is a generalization of AIC. In this simulation, $H = S_{40}$ and $\mathcal{M} = \{S_{15}, S_{20}, S_{25}, S_{30}, S_{35}, S_{40}\}$, where S_n is spanned by polynomials $\{x^i\}_{i=0}^n$. The distribution of the generalization error by 1000 trials is displayed in Fig.1. This shows that SIC works as well as NIC when M is large, and it outperforms NIC when M is small.

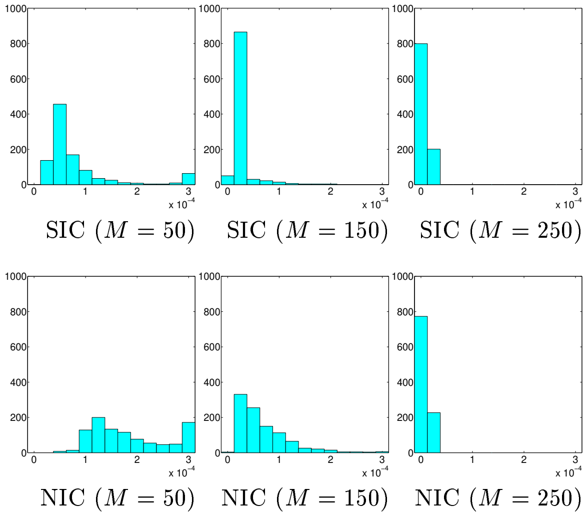

Figure 1: Simulation results.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19(6)*, 716–723.

[2] Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik, 31*, 377–403.

[3] Mallows, C. L. (1973). Some comments on C_P. *Technometrics, 15(4)*, 661–675.

[4] Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion—Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks, 5(6)*, 865–872.

[5] Sugiyama, M., & Ogawa, H. (2000). Subspace information criterion for model selection. *Neural Computation.* (to appear)

[6] Vapnik, V. N. (1995). *The nature of statistical learning theory.* Berlin: Springer-Verlag.