AIC for Conditional Model Selection

Thijs van Ommen THIJS.VAN.OMMEN@CWI.NL
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands

Keywords: model selection, prediction, supervised learning, covariate shift, AIC

In supervised learning applications, AIC and many other popular model selection methods are biased because they implicitly assume that the inputs (covariates) X in the training set take the same values as the inputs X' in the test set. Based on a novel, unbiased expression for KL divergence, we propose FAIC, a *focused* version of AIC that takes the value of X' on the test set into account. Our experiments suggest that if X' substantially differs from X, then FAIC predictively outperforms AIC, BIC and several other methods including Bayesian model averaging.

We introduce FAIC as an adaptation of AIC to supervised learning problems. The aim of AIC (Akaike, 1973) and many other model selection methods is to use the data to find the model g which minimizes

$$-2\,\mathrm{E}_{\mathbf{U}}\,\mathrm{E}_{\mathbf{V}} \log g(\mathbf{V} \mid \hat{\theta}(\mathbf{U})), \tag{1}$$

where $\hat{\theta}$ represents the maximum likelihood estimator in that model, and both random variables are independent samples of n data points each, both following the true distribution of the data. This quantity can be seen as representing that we first estimate the model's parameters using a random sample $\mathbf{U}$, then judge the quality of this estimate by looking at its performance on an independent, identically distributed sample $\mathbf{V}$.

In supervised learning problems such as regression and classification, the data points consist of two parts $u_i = (x_i, y_i)$, and the models are sets of distributions on the *output variable* $\mathbf{y}$ conditional on the *input variable* x (which may or may not be random). We call these *conditional* models. Then (1) can be adapted in two ways: as the extra-sample error

$$-2\,\mathrm{E}_{\mathbf{Y}|X}\,\mathrm{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' \mid X', \hat{\theta}(X, \mathbf{Y})), \tag{2}$$

and, replacing both X and X' by a single variable X, as the in-sample error

$$-2\,\mathrm{E}_{\mathbf{Y}|X}\,\mathrm{E}_{\mathbf{Y}'|X} \log g(\mathbf{Y}' \mid X, \hat{\theta}(X, \mathbf{Y})). \tag{3}$$

The standard expression behind AIC (1) makes no reference to X or X', so that all known versions of AIC end up estimating the in-sample error. However, the extra-sample error (2) is more appropriate as a measure of the expected performance on new data.

To get an estimator for (2), we do not make any assumptions about the processes generating X and X' (so we can deal with covariate shift and with nonrandom inputs) but treat these values as given. A derivation similar to AIC's leads to a penalty term of $k + \kappa_{X'}$ in place of AIC's $2k$; in the case of linear regression,

$$\kappa_{X'} = \frac{n}{n'} \operatorname{tr}\left[X'^{\top} X' (X^{\top} X)^{-1} \right],$$

where X, X' represent design matrices and n, n' their respective numbers of data points. Similarly, a small sample corrected version analogous to $\mathrm{AIC_C}$ (Hurvich & Tsai, 1989) can be derived and has penalty

$$k + \kappa_{X'} + \frac{(k + \kappa_{X'})(k+1)}{n - k - 1}.$$

If our goal is prediction, then X corresponds to the training data, and X' may be replaced by a single point x for which we need to predict the corresponding $\mathbf{y}$. We name this method *Focused AIC*. Note that FAIC may select different models for different values of x. Alternatively, X' may be chosen using (an estimate of) all test inputs if a single choice of model is desired.

Acknowledgments

I thank Peter Grünwald and Steven de Rooij for their valuable comments and encouragement.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (pp. 267–281).

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.