CS281B/Stat241B: Advanced Topics in Learning & Decision Making

TIC/AIC

Lecturer: Michael I. Jordan *Scribes: Francis Bach, Xiaofeng Ren*

In this lecture we discuss TIC/AIC, an information-theoretic criterion for model selection. AIC stands for *An Information Criterion* or *Akaike Information Criterion*. AIC was one of the earliest attempts in the 1960s to move beyond Fisher's *Maximum Likelihood* framework. It has been widely used in practice (e.g., in time series analysis).

Let $f(x)$ denote the truth, the underlying distribution of the data y. Let $g(x, \theta)$ denote the model family we are evaluating. We do not necessarily assume that the truth $f(x)$ is in the model family. We define the optimal θ_0 for a given family to be that which minimizes the KL divergence to the truth:

$$\theta_o = \text{argmin}_\theta D(f \| g(\cdot|\theta))$$

More complex families of models are likely to have smaller values of this minimum KL divergence. However, we don't have knowledge of the truth $f(x)$, and θ has to be estimated from the data. Hence, we need to take variance into account. Suppose we have data y and an estimator $\hat{\theta}$ of θ, say, the maximum likelihood estimator $\hat{\theta}_{ML}(y)$; the KL divergence between f and $g(\cdot|\hat{\theta})$

$$D(f \| g(\cdot|\hat{\theta}_{ML})) = \int f(x) \log \frac{f(x)}{g(x|\hat{\theta}_{ML}(y))} dx$$

is now a random variable which depends on the data y. We seek to minimize the expectation of this divergence:

$$\begin{aligned} &\min \left\{ E_y[D(f \| \log g(x|\hat{\theta}_{ML}(y)))] \right\} \\ = &\min \left\{ \int f(x) \log(f(x)) dx - E_y \int f(x) \log g(x|\hat{\theta}_{ML}(y) dx \right\} \end{aligned}$$

which is equivalent to

$$\max \left\{ E_y E_x \log g(x|\hat{\theta}_{ML}(y)) \right\}$$

Notice that there is a flavor of cross-validation here. The parameter θ is estimated from the data y, yet it's tested on another independent set of data x, not the same data set y. Also, Akaike used the ML estimator for historical reasons; in the definition of the criterion any $1/\sqrt{N}$ estimator will do (although later we'll have reasons to use the ML estimator).

Now we have the criterion. We don't know the truth $f(x)$, however; so we can not calculate the expectation analytically. We will make a sequence of approximations, mainly using the Taylor expansions, which eventually lead to the AIC:

Step 1 Taylor expansion of order 2 of $\log(g(x|\hat{\theta}(y)))$ around θ_0.

$$\begin{aligned} \log g(x|\hat{\theta}(y)) \approx &\log g(x|\theta_0) + \left(\frac{\partial \log g(x|\theta_0)}{\partial \theta} \right)^T (\hat{\theta}(y) - \theta_0) \\ &+ \frac{1}{2} (\hat{\theta}(y) - \theta_0)^T \frac{\partial^2 \log g(x|\theta_0)}{\partial \theta \partial \theta^T} (\hat{\theta}(y) - \theta_0) \end{aligned}$$

By the optimality of θ_0, we can show that $E_x\left(\frac{\partial \log g(x|\theta_0)}{\partial\theta}\right) = 0$:

$$\begin{aligned}
&\theta_0 = \text{argmin}_\theta D(f\|g(x|\theta)) \\
\Rightarrow &-\frac{\partial}{\partial\theta}\int f(x)\log g(x|\theta)dx = 0 \\
\Rightarrow &\frac{\partial}{\partial\theta}E_x \log g(x|\theta_0) = 0
\end{aligned}$$

Let $I(\theta_0)$ denote $-E_x\frac{\partial^2 \log g(x|\theta_0)}{\partial\theta\partial\theta^T}$. Taking the expectation of the log likelihood above, with respect to x:

$$E_x(\log g(x|\hat{\theta}(y))) \approx E_x(\log g(x|\theta_0)) - \frac{1}{2}(\hat{\theta}(y) - \theta_0)^T I(\theta_0)(\hat{\theta}(y) - \theta_0)$$

Taking the expectation with respect to y and using the trace trick ($\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$ and if z is a vector $z^T Az = \text{tr}(z^T Az) = \text{tr}(Azz^T)$):

$$E_y(E_x(\log g(x|\hat{\theta}(y))))) = E_x(\log g(x|\theta_0)) - \frac{1}{2}\text{tr}(I(\theta_0)\Sigma)$$

where we let Σ denote $E_y(\hat{\theta}(y) - \theta_0)(\hat{\theta}(y) - \theta_0)^T$.

Step 2 Taylor expansion of order 2 of $\log g(x|\theta_0)$ around $\hat{\theta}(x)$.

$$\begin{aligned}
\log g(x|\theta_0) \approx \log g(x|\hat{\theta}(x)) &+ \left(\frac{\partial \log g(x|\hat{\theta}(x))}{\partial\theta}\right)^T (\theta_0 - \hat{\theta}(x)) \\
&+ \frac{1}{2}(\theta_0 - \hat{\theta}(x))^T \frac{\partial^2 \log g(x|\hat{\theta}(x))}{\partial\theta\partial\theta^T}(\theta_0 - \hat{\theta}(x))
\end{aligned}$$

Since $\hat{\theta}(x)$ maximizes $g(x|\theta)$ for a given x, the derivative $\frac{\partial \log g(x|\hat{\theta}(x))}{\partial\theta}$ is equal to 0 and, taking the expectation with respect to x:

$$\begin{aligned}
E_x(\log g(x|\theta_0)) &\approx E_x(\log g(x|\hat{\theta}(x))) \\
&\quad - \frac{1}{2}\text{tr}(E_x(\hat{I}(\hat{\theta}(x))(\theta_0 - \hat{\theta}(x))^T(\theta_0 - \hat{\theta}(x))))
\end{aligned}$$

where

$$\hat{I}(\hat{\theta}(x))) = -\frac{\partial^2 \log g(x|\hat{\theta}(x))}{\partial\theta\partial\theta^T}$$

In the last expression, $\hat{I}(\hat{\theta}(x))$ fluctuates around its mean like $1/n$ whereas the other term inside the expectation fluctuates as $1/\sqrt{n}$, so the first term can be considered as constant and equal to $I(\theta_0)$. We then get:

$$E_x(\log g(x|\theta_0)) \approx E_x(\log g(x|\hat{\theta}(x))) - \frac{1}{2}\text{tr}(I(\theta_0)\Sigma)$$

Step 3 Putting everything together, the criterion $T = E_y(E_x(\log(g(x|\hat{\theta}(y)))))$ can be written as

$$E_y(E_x(\log g(x|\hat{\theta}(y)))) \approx E_x(\log g(x|\hat{\theta}(x))) - \text{tr}(I(\theta_0)\Sigma)$$

To proceed from here, we will use $\log(g(x|\hat{\theta}(x)))$ as an unbiased estimator of $E_x(\log g(x|\hat{\theta}(x)))$. For the second term, we make the approximation:

$$\mathrm{tr}(I(\theta_0)\Sigma) \approx k,$$

where k is the number of parameters in the model. This approximation can be verified for linear regression and is true more generally. We obtain the AIC criterion:

$$A = \log g(x|\hat{\theta}(x)) - k.$$

More generally we have the TIC criterion:

$$A = \log g(x|\hat{\theta}(x)) - \mathrm{tr}(I(\theta_0)\Sigma),$$

where we can approximate the trace term in various ways (e.g., using the bootstrap).