10-801: Advanced Topics in Graphcal Models 10-801, Spring 2007

Model Selection

Lecturer: Eric P. Xing *Scribes: Wilson Tam*

Two model selection approaches are covered: (1) Information criteria, and (2) Bayesian model criteria

Recall that the minimum description length (MDL) is motivated by the coding theory. We need to send the data and the model parameter which describes the data distribution:

$$L(x) = -log(x) + K/2 \cdot \log N \tag{1}$$

In this lecture, we go at a different angle for model selection. Assuming that we have some iid training data $Y = \{y_1, y_2, ..., y_n\}$ drawn from an unknown true distribution $y \sim f(.)$. We choose some tractable distribution to approximate f(.) by minimizing a distance measure $D(f(.)||g(.|\hat{\theta}(Y)))$ between f(.) and g(.):

$$D_g(.) = E_Y[D(f(x)||g(x|\hat{\theta}(Y))] \tag{2}$$

One criteria which fits into this framework is AIC (by Akaike) which employs the Kullback Leibler (KL) divergence between f(.) and g(.).

$$D(.) = E_X[\log \frac{f(x)}{g(x|\hat{\theta}(Y)}] \tag{3}$$

Minimizing D(.) is equivalent to maximizing $E_X[\log g(x|\hat{\theta}(Y)]$. The challenge here is to perform the expectation. One technique to approximate the expectation is Laplace approximation. The idea is to approximate $\log g(x|\theta)$ around the optimal solution $\theta_0 = argmax_\theta E_X[\log g(x|\theta)]$ using the Taylor expansion up to the 2nd-order term:

$$\log g(x|\hat{\theta}_y) = \log g(x|\theta_0) + \frac{\partial}{\partial\theta} \log g(x|\theta) \cdot (\hat{\theta}_y - \theta_0) + 0.5(\hat{\theta}_y - \theta_0)^t \cdot I(\theta_0) \cdot (\hat{\theta}_y - \theta_0) \tag{4}$$

where $I(\theta)$ is the Hessian w.r.t. θ: $I(\theta) = \frac{\partial}{\partial\theta\partial\theta^t} \log g(x|\theta)$ By definition, the 2nd term in the expansion vanishes since the gradient at the optimal value θ_0 is zero.

By noticing that the 2nd-order term in the expansion is a scalar, the trace of a scalar is the scalar itself. We apply the trace trick to compute $E_Y[.]$ on this term. Recall the trace trick: $Tr(ABC) = Tr(C^tAB)$. Therefore,

$$E_Y[Tr((\theta_y - \theta_0)^t \cdot I(\theta_0) \cdot (\theta_y - \theta_0)] = E_Y[Tr(I(\theta_0) \cdot (\theta_y - \theta_0)(\theta_y - \theta_0)^t)] \tag{5}$$

$$= Tr(I(\theta_0) \cdot E_Y[(\theta_y - \theta_0) \cdot (\theta_y - \theta_0)^t]) \tag{6}$$

After some manipulation, we arrive at the AIC criterion: $E_Y[E_X[D(.)]] = \log g(x|\hat{\theta}_y)$ - K

Now, let's take a look at the Bayesian Information Criterion (BIC): $\theta = argmaxp(D|m)$, i.e. choose the parameter which maximizes the marginal likelihood of the data given some model m.

The BIC provides a principled framework to compare different model complexity in constract to the hypothesis testing approach which may be hard to reject the null hypothesis due to the potentially large dimensionality of the parameter space.

By Bayes rule, we see that $p(D, \theta|m) \propto p(\theta|D, m) \cdot p(D|m)$. We first apply Laplace approximation to approximate $\log p(\theta|D, m)$:

$$\log p(\theta|D, m) \quad = \quad \log p(\theta_{map}|D, m) - 0.5(\theta - \theta_{map})^t H(\theta - \theta_{map}) \tag{7}$$

where H denotes the -ve Hessian wrt θ at θ_{map}.

Now, by exponentiating both sides of the above equation and plug it into the Bayes rule, we get the approximation of $p(D, \theta|m)$:

$$p(D, \theta|m) \quad \approx \quad p(D, \theta_{map}|m) \cdot exp(-0.5(\theta - \theta_{map})^t H(\theta - \theta_{map})) \tag{8}$$

Integrating out θ on both sides of the equation is now easy since the exp(.) term is actually Gaussian-like and the result of the integration is $(2\pi)^{K/2} \cdot |H^{-1}(\theta_{map})|^{1/2}$ where the Hessian is evaluated at $\theta = \theta_{map}$.

So the BIC criterion can be approximated as the log of the approximated marginal likelihood:

$$BIC \quad = \quad \log \left\{ p(D|\theta_{map}, m) \cdot p(\theta_{map}|m) \cdot (2\pi)^{K/2} \cdot |H^{-1}(\theta_{map})|^{1/2} \right\} \tag{9}$$

I am sorry that I missed the part of approximating the inverse of the Hessian.