CSCC11: Model Selection with BIC

November 15, 2015

Model selection

The problem of model selection has come up at a number of points in the course. How many dimensions in PCA? What degree of polynomial to select for regression? How many clusters should we use in a GMM?

One way to answer this question is to try ask the data. That is, select the model M with the highest probability given the data D. We'll assume that every model M has a unique set of parameters θ which would need to be estimate. Given those parameters, the likelihood is then $p(D|\theta, M)$ and there may be some prior on those parameters as $p(\theta|M)$. Now, using Bayes rule and marginalization we get:

$$\begin{aligned} p(M|D) &= \frac{p(D|M)p(M)}{p(D)} \\ &= \frac{p(M)\int p(D|M,\theta)p(\theta|M)d\theta}{p(D)} \end{aligned}$$

Unfortunately, as we've discussed elsewhere, there are very few cases where this the integral here is tractable. So how are we supposed to evaluate this? One approach which we'll look at is called BIC or the Bayesian Information Criterion.

Bayesian Information Criterion

To begin, lets consider the negative log of the model probability. That is.

$$-\log p(M|D) = \log p(D) - \log p(M) - \log \int p(D|M,\theta)p(\theta|M)d\theta$$

The first term doesn't depend on M and the second term is a prior which we'll leave for later. The hard term here is the integral. To work on that we're going to construct a Taylor series approximation to the log likelihood term. That is if

$$L_M(\theta) = \log p(D|M,\theta) + \log p(\theta)$$

then

$$L_M(\theta) \approx L_M(\hat{\theta}) + (\theta - \hat{\theta})^T \frac{\partial L_M(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial^2 L_M(\hat{\theta})}{\partial^2 \theta}(\theta - \hat{\theta})$$

If we select $\hat{\theta}$ to be a point where $L_M(\theta)$ is maximized (ie, the MAP solution to the problem) then $\frac{\partial L_M(\hat{\theta})}{\partial \theta} = 0$ and

$$\begin{aligned} L_M(\theta) &\approx L_M(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial^2 L_M(\hat{\theta})}{\partial^2 \theta}(\theta - \hat{\theta}) \\ &= L_M(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T (n\mathcal{F}(\hat{\theta}))(\theta - \hat{\theta}) \end{aligned}$$

where

$$\begin{aligned} \mathcal{F}(\hat{\theta}) &= -\frac{1}{N} \frac{\partial^2 L_M(\hat{\theta})}{\partial^2 \theta} \\ &= -\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial^2 \log p(D_i|M, \hat{\theta})}{\partial^2 \theta} + \frac{1}{N} \frac{\partial^2 \log p(\hat{\theta}|M)}{\partial^2 \theta} \right) \end{aligned}$$

is a matrix. An important property of this matrix is that it is basically an average over data observations so that, as N increases, it should converge to a constant.

Now, exponentiating our approximation, we get

$$p(D|M, \theta)p(\theta|M) \approx p(D|M, \hat{\theta})p(\hat{\theta}|M) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T (N\mathcal{I}(\hat{\theta}))(\theta - \hat{\theta})\right)$$

and plugging this into the integral above we get

$$\int p(D|M, \theta)p(\theta|M) d\theta \approx \int p(D|M, \hat{\theta})p(\hat{\theta}|M) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T (N\mathcal{F}(\hat{\theta}))(\theta - \hat{\theta})\right) d\theta$$

Now, we can go ahead and move the constants outside the integral and what's left inside is just an unnormalized Gaussian distribution. Then

$$\begin{aligned} \int p(D|M, \theta)p(\theta|M) d\theta &\approx p(D|M, \hat{\theta})p(\hat{\theta}|M) \int \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T (N\mathcal{F}(\hat{\theta}))(\theta - \hat{\theta})\right) d\theta \\ &= p(D|M, \hat{\theta})p(\hat{\theta}|M)(2\pi)^{\frac{K_M}{2}} \left| N\mathcal{F}(\hat{\theta}) \right|^{-\frac{1}{2}} \\ &= p(D|M, \hat{\theta})p(\hat{\theta}|M) \left(\frac{2\pi}{N}\right)^{\frac{K_M}{2}} \left| \mathcal{F}(\hat{\theta}) \right|^{-\frac{1}{2}} \end{aligned}$$

where K_M is the dimensionality of θ for model M. Now, plugging this integral into our equation above we get

$$\begin{aligned} -\log p(M|D) &\approx \log p(D) - \log p(M) - \log p(D|M, \hat{\theta})p(\hat{\theta}|M) + \frac{K_M}{2} \log \frac{N}{2\pi} + \frac{1}{2} \log \left| \mathcal{F}(\hat{\theta}) \right| \\ &= \log p(D) - \log p(M) - \frac{K_M}{2} \log 2\pi + \frac{1}{2} \log \left| \mathcal{F}(\hat{\theta}) \right| - \log p(D|M, \hat{\theta})p(\hat{\theta}|M) + \frac{K_M}{2} \log N \end{aligned}$$

Now, lets look at the limit of this as N increases to infinity.

- The first three terms are constants in terms of N
- The fourth term $\frac{1}{2}\log\left|\mathcal{F}(\hat{\theta})\right|$ will converge to a constant as N increases
- The fifth and sixth terms, $-\log p(D|M,\hat{\theta})p(\hat{\theta}|M)+\frac{K_M}{2}\log N$ will grow as the amount of data increases

So asymptotically, the only two terms we're left with are the last two. Together, this gives us what is know as the Bayes Information Criterion or BIC.

$$\begin{aligned} -\log p(M|D) &\sim -\log p(D|M,\hat{\theta})p(\hat{\theta}|M)+\frac{K_M}{2}\log N \\ &= BIC(M|D) \end{aligned}$$

Uses of BIC

- First and foremost, BIC gives us a way to choose between two different models with different numbers of parameters by selecting the one which gives us the lowest BIC score.
- More complex models are almost always likely to fit the data better (and therefore have a lower value of $-\log p(D|M,\hat{\theta})$. BIC gives us a relatively principled way to penalize these extra parameters in the form of the term $\frac{K_M}{2}\log N$. Note that this term doesn't just penalize more parameters, it also says that if you have more data, you expect those extra parameters to help you that much more.
- Beyond basic model selection, BIC can give us some clue as to whether the differences between models are meaningful. For instance, define $\Delta = BIC(M_1|D) - BIC(M_2|D)$. If Δ is positive, then M_2 is better than M_1 but how much better? Roughly speaking,

$$|\Delta| = \begin{cases} 0-1 & \text{insignificant} \\ 1-3 & \text{meaningful} \\ 3-5 & \text{strong} \\ 5+ & \text{very strong} \end{cases}$$

- Finally, since $BIC(M|D)$ can be thought of as a surrogate for $-\log p(M|D)$ we can use it to do model averaging. Specifically, consider making a prediction for some value y^{new} given the data we've observed

$$\begin{aligned} p(y^{new}|D) &= \sum_M p(y^{new},M|D) \\ &= \sum_M p(y^{new}|M,D)p(M|D) \\ &\approx \sum_M p(y^{new}|M,D)w_M \end{aligned}$$

where

$$w_M = \frac{\exp(-BIC(M|D))}{\sum_{M'} \exp(-BIC(M'|D))}$$

In other words, instead of picking the single model with the lowest BIC, we can use the BIC to define weights and combine together all the models we evaluated. This is generally found to work better. Plus, we already went to all the trouble of evaluating all the models in the first place to compute the BIC for all of them, might as well use them!

Other forms of Model Selection

BIC is but one of many model selection criterias. Some others that you should be aware of include

- BIC is traditionally defined slightly differently. Specifically, it is assumed that the prior is relatively weak around the the mode $\hat{\theta}$ and there is an arbitrary scaling constant involved. Thus, the traditional BIC measure is

  $$BIC_{trad}(M|D) = -2 \log p(D|M, \hat{\theta}) + K_M \log N$$

- A similar, but related, measure is called AIC (Akaike Information Criterion) and is derived using information theory. It has a similar form to BIC

  $$AIC(M|D) = -2 \log p(D|M, \hat{\theta}) + 2K_M$$

  and there are many variations of AIC which can be derived to improve the quality of it's estimate. Looking at this you can see that BIC will generally penalize complex models more strongly than AIC whenever $N > e^2$.

- BIC and AIC both can be considered an instance of a general technique known as *penalized likelihood*, where a penalty term is added to the negative log likelihood which penalizes more complex models. This is not unlike adding a prior. However, it is often difficult to come up with good, general purpose priors for things like model complexity which AIC and BIC.

- N-fold cross validation is another form of model selection which takes a very different approach to BIC and, as a consequence, has somewhat different properties. Specifically, BIC or AIC *cannot differentiate between two models with the same number of parameters.* For instance, consider trying to choose between a set of polynomial or sinusoidal basis functions in basis function regression. In this case, they both simply devolve into Maximum Likelihood. In contrast, N-fold cross validation can diffentiate between different models with the same number of parameters.