MAST90125: Bayesian Statistical learning

Lecture 7: Model comparison

Feng Liu and Guoqi Qian

Overview of model comparison

From previously taken statistics subjects, name some techniques that you are familiar with for comparing models?

- F-test based techniques in regression, such as extra sum of squares/forward selection/backward selection.
- Likelihood ratio tests.
  - Both of which imply models are nested.
- Akaike Information criterion (AIC)
- Cross-validation
  - Which do not require models to be nested.

Ideal properties of a measure of predictive accuracy

Given a set of candidate models M_i, what would you consider when trying to pick the 'best' model?

- Optimise the fit to data?
  - But to what data? Observed or hypothetical.
    - If we optimise the fit to observed data, which model would be picked? In regression, it would be the model M_c with the most parameters.
    - But would M_c provide a good fit to hypothetical data drawn from the same process? Not necessarily, if M_c is overly complex, it will likely overfit.
- Ideally, the best model will be the most parsimonious among those maximising model fit and generalisability while minimising model complexity.
  - But never forget, 'all models are wrong, but some are useful.'

Defining a measure of predictive accuracy

We already accept the idea that a good model should generalise. E.g. in lecture 6, we devoted considerable attention to generating replicate data to check model plausibility.

- This naturally suggests that a good choice for measuring predictive accuracy would be a function, say g, of the posterior predictive distribution

$$p(\tilde{y}_i|\mathbf{y}) = \int p(\tilde{y}_i|\theta)p(\theta|\mathbf{y})d\theta = E_{\theta|\mathbf{y}}(p(\tilde{y}_i|\theta)), \quad i = 1, \cdots, n,$$

where $\mathbf{y} = (y_1, \cdots, y_n)^\top$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)^\top$.

However in the case of real data, what can we say about all candidate models?

- All models are approximations of the true distribution, f, for $\mathbf{y}$. Hence ideally, rather than use $g(p(\tilde{\mathbf{y}}|\mathbf{y}))$ to measure model fit directly, we would find the expectation of $g(p(\tilde{\mathbf{y}}|\mathbf{y}))$ with respect to f.

Kullback-Leibler divergence

The Kullback-Leibler divergence is a measure of the discrepancy that occurs by using the distribution p as an approximation for the distribution f,

$$D_{KL}\{f(\mathbf{y})||p(\mathbf{y})\} = \int \{\log(f(\mathbf{y})) - \log(p(\mathbf{y}))\} f(\mathbf{y}) d\mathbf{y} = E_f(\log(f(\mathbf{y}))) - E_f(\log(p(\mathbf{y}))),$$

which is strictly non-negative such that $D_{KL}\{f(\mathbf{y})||p(\mathbf{y})\} \geq 0$.

- As already noted, we would never know the density f, however if we compare models with densities p_1, p_2 respectively, the difference in D_{KL} would be

$$D_{KL}\{f(\mathbf{y})||p_1(\mathbf{y})\} - D_{KL}\{f(\mathbf{y})||p_2(\mathbf{y})\} = E_f(\log(p_2(\mathbf{y}))) - E_f(\log(p_1(\mathbf{y}))),$$

suggesting we want to find an estimate of the expected log of the posterior predictive density, either for a single data-point, or for a set of data.

Estimating $E_f(\log(p(\tilde{\mathbf{y}}|\mathbf{y})))$

So the question now becomes how to estimate

$$E_f\left(\log\left(p(\tilde{\mathbf{y}}|\mathbf{y})\right)\right) = E_f\left(\log\left(E_{\theta|\mathbf{y}}[p(\tilde{\mathbf{y}}|\theta)]\right)\right) = E_f\left(\sum_{i=1}^{n}\log(E_{\theta|\mathbf{y}}[p(\tilde{y}_i|\theta)])\right).$$

- We already used simulation to generate $\mathbf{y}^{\text{rep}}$ from the posterior predictive distribution, so would $\sum_{i=1}^{n}\log\left(\frac{\sum_{s=1}^{S}p(y_i^{\text{rep}}|\theta^s)}{S}\right)$ be a good estimator?
- Note $\mathbf{y}^{\text{rep}} = (y_1^{\text{rep}}, \cdots, y_n^{\text{rep}})'$ is a randomly generated value of $\tilde{\mathbf{y}}$.
- Is $\mathbf{y}^{\text{rep}}$ a realisation of data with probability density f? No, but as we saw in lecture 6, $\mathbf{y}^{\text{rep}}$ is a useful measure of determining the level of discrepancy between the distribution implied by the model and the true distribution.

Estimating $E_f(\log(p(\tilde{\mathbf{y}}|\mathbf{y})))$

- The only quantities that you can be sure follow the distribution with density f are the observed data y_i, $i = 1, \ldots, n$. This means in order to estimate $\log(E_{\theta|\mathbf{y}}[p(\tilde{\mathbf{y}}|\theta)]) = \sum_{i=1}^{n} \log(E_{\theta|\mathbf{y}}[p(\tilde{y}_i|\theta)])$ by simulation, we have to consider

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} p(y_i|\theta^s)}{S}\right),$$

- However, as already discussed, using the observed data to both estimate model parameters and assess model fit leads to an overly optimistic estimate of predictive ability. Hence we need to correct for this bias.

Measures of fit: AIC

$$AIC = -2\log(p(\mathbf{y}|\hat{\theta}_{MLE})) + 2k,$$

where k is the number of model parameters.

- You likely have encountered AIC before. It also turns out to be related to $E_f(\log(p(\tilde{\mathbf{y}}|\mathbf{y})))$, can you see how?
  - Let $n \to \infty$. Then $\theta|\mathbf{y} \sim \mathcal{N}(\hat{\theta}_{MLE}, V_0/n)$ in the limit, and $p(\theta|\mathbf{y}) \to p(\mathbf{y}|\theta)$ when prior $p(\theta) \propto 1$. Hence

$$\log(p(\mathbf{y}|\theta)) = c(\mathbf{y}) - \frac{k\log(2\pi) + \log|V_0/n|}{2} - \frac{n(\theta - \hat{\theta}_{MLE})' V_0^{-1} (\theta - \hat{\theta}_{MLE})}{2}$$

and $E_{\theta|\mathbf{y}}(\log(p(\mathbf{y}|\theta))) = \log(p(\mathbf{y}|\hat{\theta}_{MLE})) - k/2$ as $n(\theta - \hat{\theta}_{MLE})' V_0^{-1} (\theta - \hat{\theta}_{MLE}) \sim \chi^2_k$

Measures of fit: DIC

One of the criticisms of AIC is that the number of estimated parameters, k, is not the same as the effective number of parameters, especially if informative priors are used and/or the models considered are highly hierarchical.

- One method to account for this is the deviance information criterion (DIC)

$$DIC = -2\log(p(\mathbf{y}|\hat{\theta}_{\text{Bayes}})) + 2p_{\text{DIC}},$$

where $\hat{\theta}_{\text{Bayes}} = E(\theta|\mathbf{y})$ and p_{DIC} is either

  - $p_{\text{DIC}} = 2\{\log(p(\mathbf{y}|\hat{\theta}_{\text{Bayes}})) - E_{\theta|\mathbf{y}}(\log(p(\mathbf{y}|\theta)))\}$ or
  - $p_{\text{DIC}} = 2\text{Var}_{\theta|\mathbf{y}}(\log(p(\mathbf{y}|\theta)))$.

- It turns out that both expressions for p_{DIC} are justified in the large sample scenarios we looked at when trying to justify AIC.

Measures of fit: WAIC

The Watanabe-Akaike or widely available information criterion (WAIC) is a measure of prediction accuracy similar to DIC and AIC, but which does not need point estimates.

- The formula for WAIC is,

$$WAIC = -2 \times \widehat{lppd} + 2p_{\text{WAIC}},$$

where p_{WAIC} is either

  - $p_{\text{WAIC}} = 2\{\log(E_{\theta|\mathbf{y}}(p(\mathbf{y}|\theta))) - E_{\theta|\mathbf{y}}(\log(p(\mathbf{y}|\theta)))\}$ or
  - $p_{\text{WAIC}} = \text{Var}_{\theta|\mathbf{y}}(\log(p(\mathbf{y}|\theta)))$.

- In practice, the quantities required to estimate *WAIC* are estimated based on simulation.

Measures of fit: Leave one out cross validation

The measures of fit we have looked at so far used the posterior based on all data available. However, similar statistics can be constructed in a cross-validation setting.

- The log predictive density of y_i conditional on all other data $\mathbf{y}_{-i}$, $\log(E_{\theta|\mathbf{y}_{-i}}(p(y_i|\theta)))$ can be estimated using

$$\log\left(\frac{\sum_{s=1}^{S} p(y_i|\theta^{is})}{S}\right), \quad \text{where } \theta^{is} \sim p(\theta|\mathbf{y}_{-i}).$$

- This leads us to define lppd_{loo-cv} as $\sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} p(y_i|\theta^{is})}{S}\right)$.
  - In addition, we can include bias corrections, $b = \text{lppd} - \overline{\text{lppd}}_{-i}$, where $\overline{\text{lppd}}_{-i} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E_{\theta|\mathbf{y}_{-i}}(\log(p(y_j|\theta)))$ such that $\text{lppd}_{cv} = \text{lppd}_{loo-cv} + b$ and estimate the number of parameters p by calculating the difference $\text{lppd} - \text{lppd}_{cv}$.

Example:

Now, let's look at these measures in the context of a linear regression.

- Note the code for this example will be put up as a separate R script after the lecture.

Model comparison so far

So far, we focused on measures of predictive accuracy (AIC, DIC, WAIC) that were based on information criterion.

While in the previous example, we did not use these measures of predictive accuracy for the purpose of model comparison, we did look at how model comparison can be justified based on information criterion. For AIC, DIC, WAIC, this implied that the model with the lowest value is preferred.

Now we will consider approaches where the model comparison aspect is more explicit.

Bayes Factors

- Consider the situation where we have a sequence of candidate models $M_1, \ldots M_m$, each with a unique set of parameters $\theta_{M_i}, i = 1, \ldots, m$
- In addition to assigning prior distributions for each set of parameters, θ_{M_i}, we can also assign prior probabilities for the models M_i.
  - Which must mean that for each model, we can calculate the posterior distribution of the model, that is $p(M_i|\mathbf{y})$.
- So model comparison is one of comparing posterior distributions, for instance

$$\frac{p(M_2|\mathbf{y})}{p(M_1|\mathbf{y})} = \frac{p(M_2, \mathbf{y})/p(\mathbf{y})}{p(M_1, \mathbf{y})/p(\mathbf{y})} = \frac{p(M_2, \mathbf{y})}{p(M_1, \mathbf{y})} = \frac{p(M_2)}{p(M_1)} \times \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)}$$

Bayes Factors

$$\frac{p(M_2|\mathbf{y})}{p(M_1|\mathbf{y})} = \frac{p(M_2,\mathbf{y})/p(\mathbf{y})}{p(M_1,\mathbf{y})/p(\mathbf{y})} = \frac{p(M_2,\mathbf{y})}{p(M_1,\mathbf{y})} = \frac{p(M_2)}{p(M_1)} \times \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)} \tag{1}$$

- In (1), the term $p(M_2)/p(M_1)$ is the prior ratio for the models being compared.
- The term $p(\mathbf{y}|M_2)/p(\mathbf{y}|M_1)$, is known as the Bayes factor, $BF(M2 : M1)$. But what does this correspond to?
  - $p(\mathbf{y}|M_i)$ is the prior predictive distribution of the observed data y assuming model M_i. Hence the Bayes factor is equivalent to,

$$BF(M2 : M1) = \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)} = \frac{\int p(\mathbf{y}|\theta_2, M_2)p(\theta_2|M_2)d\theta_2}{\int p(\mathbf{y}|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}$$

- Now we will consider some examples.

Bayes Factors example 1:

Bayes factors are closely related to Likelihood ratio tests used in hypothesis testing.

- Consider the hypothesis $H_0 : p = p_0$ vs. $H_A : p \sim U(0,1)$. In this setting $p = p_0$ is M_2 and $p \sim U(0,1)$ is M_1. The likelihood is $\text{Bin}(n, p)$.
- For M_2, the prior predictive distribution is just the likelihood,

$$p(y|M_2) = p(y|p_0, M_2) = \binom{n}{y} p_0^y (1-p_0)^{n-y},$$

while for M_1, the prior predictive distribution is an integrated likelihood,

$$p(y|M_1) = \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \times 1 dp = \frac{\binom{n}{y}\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} = \frac{1}{n+1},$$

- and the Bayes Factor is $\frac{p(y|M_2)}{p(y|M_1)} = (n+1)\binom{n}{y} p_0^y (1-p_0)^{n-y}$.
- Which leaves the question, would the Bayes Factor always give sensible results? To determine this, we will now turn to R.

Comments and second example:

- In the previous example, we used a flat prior for p. In earlier lectures, we discussed how flat priors are a popular choice of 'uninformative' prior.
- However, if the range of the parameter θ is unbounded, then a flat prior would be improper. While we have shown improper priors can lead to proper posterior densities, for the Bayes Factor to work it is the prior predictive distribution that needs to be proper.
- To understand this further, consider the case where $y_i \sim \mathcal{N}(\mu, \sigma^2)$ with μ unknown and σ^2 known.

Model expansion

- In our discussion of model comparison, we have always worked under the assumption that a discrete number of potential models exist.
- However, we have also discussed that the methods of model comparison we have considered all have potential flaws.
- Moreover our discussion of model comparison ignores the flexibility gained by being able to specify prior distributions. Often, the candidate models can be viewed as special cases of a more general model.

Example: One factor ANOVA

- In previous subjects, you have likely encountered analysis of variance. In the one factor case with K groups, the set of hypotheses tested are

$$H_0 : \mu_1, \ldots, \mu_K = \mu \quad \text{vs.} \quad H_A : \text{At least one population mean is different}$$

  with variation within each group being constant.
- Now say you were asked to analyse this problem in a Bayesian framework. How would you specify the model. One example could be
  - $y_{ij}|\mu_j, \sigma^2 \sim \mathcal{N}(\mu_j, \sigma^2)$.
  - $\mu_j \sim \mathcal{N}(\mu, \sigma^2_\mu)$
  - $\mu \sim 1$
  - $(\sigma^2_\mu)^{-1} \sim \text{Ga}(\alpha_\mu, \beta_\mu)$
  - $(\sigma^2)^{-1} \sim \text{Ga}(\alpha, \beta)$.

Example: One factor ANOVA

- Looking at the specifications,
  - $y_{ij}|\mu_j, \sigma^2 \sim \mathcal{N}(\mu_j, \sigma^2)$.
  - $\mu_j \sim \mathcal{N}(\mu, \sigma_\mu^2)$
  - $\mu \sim 1$
  - $(\sigma_\mu^2)^{-1} \sim \text{Ga}(\alpha_\mu, \beta_\mu)$
  - $(\sigma^2)^{-1} \sim \text{Ga}(\alpha, \beta)$,

  what model(s) are we considering?
- The null hypothesis would correspond to the case where $\sigma_\mu^2 = 0$.
- The ANOVA alternative hypothesis, would correspond to the case where $\sigma_\mu^2 > 0$.
- The hypothesis that each group is different would correspond to the case where $\sigma_\mu^2 \to \infty$.