MSA101/MVE187 2020 Lecture 14.1
Model comparison and choice

Petter Mostad

Chalmers University

October 9, 2020

Comparing models

- In our "cookbook" of Lecture 13.1, the fourth step was described as:

  Compare the possible candidate models by looking at the probability of observing the actual data under each model. Decide on one or a combination of models.

  In this section of lecture 13 we describe how.
- The above is called Bayesian model choice. There are also many non-Bayesian ways to compare models, we discuss this briefly.

Bayesian model choice

- In Bayesian statistics, just as we can use a prior that is a mixture of other priors, we can use a model that is a mixture of other models:
- Let y denote the data, and let $M_1, \ldots, M_k$ denote k different stochastic models for this data.
- For example, if model i has parameters θ and additional variables x, we can write the joint density as $\pi(y, \theta, x \mid M_i)$.
- *If* the joint density is *proper* we can integrate out x and θ to get a marginal distribution for the data under model M_i: $\pi(y \mid M_i)$.
- If we have for the models *prior probabilities* $\pi(M_1), \pi(M_2), \ldots, \pi(M_k)$ that sum to 1, we can use a mixed model:

$$\pi(y) = \sum_{i=1}^{k} \pi(y \mid M_i)\pi(M_i)$$

- Computation can proceed as with any other Bayesian model.

Computations

- Example: For predicting new data y_{new} given the information from old data y, we use a mixture of updated models with updated weights:

$$\pi(y_{new} \mid y) = \sum_{i=1}^{k} \pi(y_{new} \mid M_i, y)\pi(M_i \mid y)$$

- The updated weights can be computed with Bayes formula:

$$\pi(M_i \mid y) = \frac{\pi(y \mid M_i)\pi(M_i)}{\sum_{j=1}^{k} \pi(y \mid M_j)\pi(M_j)}$$

- Note that we introduce no new theory, we simply introduce an extra variable M which has k different values, so that conditionally on this variable being i, the model has a particular form M_i.

Example: Comparing three simple models

- Toy example: Data $(x_1, y_1), \ldots, (x_{10}, y_{10})$ to be fitted with regression. (Visualize toy data)
- We consider three possible models:
  - Model 1

$$
\begin{aligned}
y_i \mid a, x_i &\sim \text{Normal}(ax_i, 0.5^2) \\
a &\sim \text{Normal}(0.3, 0.1^2)
\end{aligned}
$$

  - Model 2

$$
\begin{aligned}
y_i \mid a, b, x_i &\sim \text{Normal}(ax_i + bx_i^2, 0.5^2) \\
a &\sim \text{Normal}(0.3, 0.1^2) \\
b &\sim \text{Normal}(0.02, 0.01^2)
\end{aligned}
$$

  - Model 3

$$
\begin{aligned}
y_i \mid a, x_i &\sim \text{Normal}(\exp(ax_i), 0.5^2) \\
a &\sim \text{Normal}(0.1, 0.01^2)
\end{aligned}
$$

- Our prior probability for each model is $1/3$.

Computations for the example

- Let's say we want to find the posterior weights for the models.
- For each model, we need to compute the probability of the data under the model. For example, in the first model, we get

$$\pi(y_1, \ldots, y_{10} \mid M_1) = \int_{-\infty}^{\infty} \pi(y_1, \ldots, y_{10} \mid a, x_1, \ldots, x_{10}) \pi(a) \, da$$

$$= \int_{-\infty}^{\infty} \left[\prod_{i=1}^{10} \text{Normal}(y_i; ax_i, 0.5^2) \right] \text{Normal}(a; 0.3, 0.1^2) \, da$$

- For the second model we get

$$\begin{aligned}
& \pi(y_1, \ldots, y_{10} \mid M_2) \\
= & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \pi(y_1, \ldots, y_{10} \mid a, b, x_1, \ldots, x_{10}) \pi(a) \pi(b) \, da \, db \\
= & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\prod_{i=1}^{10} \text{Normal}(y_i; ax_i + bx_i^2, 0.5^2) \right] \cdot \\
& \text{Normal}(a; 0.3, 0.1^2) \, \text{Normal}(b; 0.02, 0.01^2) \, da \, db
\end{aligned}$$

Computations for the example

- Computations in R (using discretiaztion) give us

$$\pi(y_1, \ldots, y_{10} \mid M_i) = \begin{cases} 0.00038832 & \text{for } i = 1 \\ 0.00008302 & \text{for } i = 2 \\ 0.00055191 & \text{for } i = 3 \end{cases}$$

- Using Bayes formula and the uniform prior on the weigts gives us the posterior weights

$$\pi(M_1 \mid y_1, \ldots, y_{10}) = \begin{cases} 0.38 & \text{for } i = 1 \\ 0.08 & \text{for } i = 2 \\ 0.54 & \text{for } i = 3 \end{cases}$$

- If we want, we can make predictions for new observations y_{new} with a mix of the posterior models using these posterior weights.

Special interpretation and nomenclature

- The value $\pi(y \mid M_i)$ is called the **evidence** for model M_i.
- Ratios like

  $$\frac{\pi(y \mid M_i)}{\pi(y \mid M_j)}$$

  are called **Bayes factors**.
- The setup above is called **model averaging**.
- If the weight of a model becomes "very small" after updating weights with data, we may choose to drop the model entirely, to simplify.
- If we drop all but one model, the method becomes a type of **model choice**!

Using Bayes factors without first establishing prior probabilities for models

- We saw above: The vector of posterior model weights is proportional to the vector of evidences times the vector of prior weights.
- As prior weights may be difficult to establish, one may instead first look at the relative values of the evidences, i.e., the Bayes factors.
- If the Bayes factor in favour of one model is "sufficiently large", one may directly decide to choose this model and to discard the others.

Problems with Bayesian model choice

There are two major problems with using the theory above

1. If M_i uses an *improper prior*, the joint model, before conditioning on data, will be improper, and $\pi(y \mid M_i)$ cannot be computed!
2. Even when M_i is proper, it may be computationally very difficult to compute the number $\pi(y \mid M_i)$.

Difficulty of computing $\pi(y \mid M_i)$

- Remember Bayes formula:

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}.$$

  We have mentioned many times that the denominator $\pi(y)$, which we now denote $\pi(y \mid M_i)$ and call the evidence, may be difficult to compute!

- Some approaches to compute or approximate $\pi(y)$:
  - In not-too-high dimensions, (numerical) integration of $\pi(y \mid \theta)\pi(\theta)$.
  - In not-too-high dimensions, you may approximately fit some density $f(\theta)$ using a sample from the posterior. Then the integration constant $\pi(y)$ is approximated by $\pi(y \mid \theta)\pi(\theta)/f(\theta)$ for any value θ.
  - Laplace approximation (see below).
  - Nested sampling (see below).
  - Computing the ELBO in a Variational Bayes approximation.
  - ...

The Laplace multivariate normal approximation

- Rewrite $\pi(y \mid \theta)\pi(\theta)$ as follows:

$$\pi(y \mid \theta)\pi(\theta) = \exp(-h(\theta))$$

  for some function h.

- Use a Laplace multivariate normal approximation (see Lecture 7):

$$\pi(y \mid \theta)\pi(\theta) \approx \exp(-h(\hat{\theta}))\exp\left(-\frac{1}{2}(\theta - \hat{\theta})^t H(\hat{\theta})(\theta - \hat{\theta})\right)$$

  where $\hat{\theta}$ is the θ maximizing the posterior and $H(\hat{\theta})$ is the Hessian matrix at $\hat{\theta}$.

- Integrating both sides over θ and using the formula for the multivariate normal density, we get

$$\pi(y) \approx \exp(-h(\hat{\theta}))|2\pi H(\hat{\theta})^{-1}|^{1/2}$$

Nested sampling (for orientation only)

A method to approximate the number $\pi(y) = \int \pi(y \mid \theta)\pi(y)\, d\theta$.

- The general idea of nested sampling:
  1. Sample N points from the prior and compute the likelihood in these points.
  2. Store and remove from list the point at which the likelihood is smallest.
  3. Add to list a point simulated from the prior, conditional on having at least as high likelihood as the removed point.
  4. Loop back to step 2 and repeat a number of times.
  5. From the stored points and their likelihoods, the integral of the posterior can be estimated.
- The conditoinal simulation of new points can be done in various ways (rejection sampling, MCMC,...)
- The method works well in situations with multiple peaks.

Non-Bayesian model selection (for orientation only)

- Given the drawbacks of Bayesian model choice using Bayes factors, there is of course a huge number of alternatives.
- Some focus on weighing how well the model fits the data against the complexity of the model.
- A large class of methods use *information criteria* that *penalize* the complexity of a model:
    - AIC Akaike Information Criterion.
    - BIC Bayesian Information Criterion.
    - DIC Deviance Information Criterion.
    - TIC Takeuchi Information Criterion.
    - FIC Focus Information Criterion.
    - ...