Bayesian Learning - Lecture 11

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Likelihood is no good for model comparison

- Consider two models: M_1 and M_2. Let :$p_i(y|\theta_i)$ denote the data density under model M_i. If we knew the values of θ_1 and θ_2, then the likelihood ratio
$$\frac{p_1(y|\theta_1)}{p_2(y|\theta_2)},$$
could be used to compare the models.
- What if the model parameters are unknown? The estimated likelihood ratio:
$$\frac{p_1(y|\hat{\theta}_1)}{p_2(y|\hat{\theta}_2)}.$$
where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates.
- Estimated likelihood ratio is useless in itself as the larger model always has larger likelihood. Comparison with sampling distribution of the estimated likelihood ratio is one solution.

Enter Bayes

- Bayesian: use your priors $p_1(\theta_1)$ och $p_2(\theta_2)$ and compute the **marginal likelihood**, or **prior predictive density**, for each model

$$p_k(y) = \int p_k(y|\theta_k) p_k(\theta_k) d\theta_k.$$

- The **Bayes factor** can be used to compare to models

$$B_{12}(y) = \frac{p_1(y)}{p_2(y)}.$$

- The marginal likelihoods may be converted into posterior probabilities of the models (M_1, M_2):

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(M_1)}{p(M_2)} B_{12}(y),$$

where $B_{12}(y)$ is the Bayes factor in favor of M_1.

Posterior model odds ratio = Prior model odds ratio · Bayes factor

Bayesian hypothesis testing - Bernoulli

- **Hypothesis testing** is just a special case of model selection:

$$M_0 : x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta_0)$$

$$M_1 : x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta), \theta \sim Beta(\alpha, \beta)$$

$$p(x_1, ..., x_n | M_0) = \theta_0^s (1 - \theta_0)^f,$$

$$\begin{aligned} p(x_1, ..., x_n | M_1) &= \int_0^1 \theta^s (1-\theta)^f B(\alpha, \beta)^{-1} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= B(\alpha + s, \beta + f) / B(\alpha, \beta), \end{aligned}$$

  where $B(\cdot, \cdot)$ is the **Beta function**.
- Posterior model probabilities

$$Pr(M_k | x_1, ..., x_n) \propto p(x_1, ..., x_n | M_k) Pr(M_k), \text{ for } k = 0, 1.$$

- The Bayes factor

$$BF(M_0; M_1) = \frac{p(x_1, ..., x_n | H_0)}{p(x_1, ..., x_n | H_1)} = \frac{\theta_0^s (1 - \theta_0)^f B(\alpha, \beta)}{B(\alpha + s, \beta + f)}.$$

Bayesian hypothesis testing - Bernoulli example

- This is equivalent to the posterior under the following 'spike-and-slab' prior:

$$p(\theta) = \pi I_{\theta_0}(\theta) + (1 - \pi) Beta(\alpha, \beta)$$

- Note: data can now *support* a null hypothesis (not only reject it). This is all due to the introduction of a prior.

Bayesian hypothesis testing, cont.

- Bayes tests are consistent (not true for frequentist test)

$$p(H_k|\mathbf{x}) \to 1 \text{ as } n \to \infty \text{ if } H_k \text{ is true.}$$

- The priors must be proper. Example: Let $x_1, ..., x_n$ be an independent sample from $N(\theta, 1)$.

$$\begin{aligned} H_0 &: \quad \theta = \theta_0 \\ H_1 &: \quad \theta \neq \theta_0, \ \text{ with prior } N(\mu_0, \tau_0^2) \text{ if } H_1 \text{ holds.} \end{aligned}$$

  Then it can be shown that:

$$p(H_0|\mathbf{x}) \to 1 \text{ as } \tau_0^2 \to \infty,$$

  regardless of which hypothesis is the true one.
- This result is entirely in the logic of Bayesian testing!

Example - Variable selection

- Linear regression:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

  Which variables have non-zero coefficient? Example of hypotheses:

$$\begin{aligned} H_0 &: \beta_0 = \beta_1 = ...\beta_p = 0 \\ H_1 &: \beta_1 = 0 \\ H_2 &: \beta_1 = \beta_2 = 0 \end{aligned}$$

  we could consider all possible subsets of β coefficients to be zero. Easy! Just compute the marginal likelihood of each hypothesis.
- MCMC sampling algorithms for variable selection. Introduce variable indicators:

$$I_j = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0 \end{cases}$$

- Sample from the joint posterior $p(\beta_0, \beta_1, ..., \beta_p, I_1, I_2, ..., I_p | y, x)$ using Gibbs sampling (linear Gaussian regression) or Metropolis-Hastings (everything else).

Model averaging

- Let γ be a quanitity with an interpretation which stays the same across the two models (for example a future value of the data $\tilde{y}$). The marginal posterior distribution of γ reads

$$p(\gamma|y) = p(M_1|y)p_1(\gamma|y) + p(M_2|y)p_2(\gamma|y),$$

  where $p_k(\gamma|y)$ is the marginal posterior of γ conditional on model k.
- Prediction: $\gamma = (y_{T+1}, ..., y_{T+h})'$.
- Predictive distribution includes three sources of uncertainty:
  - Future errors/disturbances (e.g. the ε's in a regression)
  - Parameter uncertainty (the predictive distribution $p(\tilde{y}|y)$ has the parameters integrated out by their posteriors)
  - Model uncertainty (by model averaging)

Marginal likelihood as measure of out-of-sample predictive performance

- The marginal likelihood of a sample $y_1, ..., y_T$ can be expressed as

$$p(y_1, ..., y_n) = p(y_1)p(y_2|y_1) \cdots p(y_n|y_1, y_2, ..., y_{n-1})$$

$$p(y_t|y_1, ..., y_{t-1}) = \int p(y_t|\theta)p(\theta|y_1, ..., y_{t-1})d\theta$$

  where we assume that y_t is independent of $y_1, ..., y_{t-1}$ conditional on θ.
- The prediction of y_1 is based on the prior of θ, and is therefore sensitive to the prior.
- The prediction of y_T uses almost all the data to infer θ. Very little influenced by the prior when T is not small.
- Log Predictive Score.
- Cross-validation.

Marginal likelihood in conjugate models

- Computing the marginal likelihood requires integration w.r.t. θ.
- Short cut for conjugate models by rearragement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

- Bernoulli model example

$$\begin{aligned} p(\theta) &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ p(y|\theta) &= \theta^{s}(1-\theta)^{f} \\ p(\theta|y) &= \frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1} \end{aligned}$$

- Marginal likelihood

$$p(y) = \frac{\theta^{s}(1-\theta)^{f}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}} = \frac{B(\alpha+s,\beta+f)}{B(\alpha,\beta)}$$

Computing the marginal likelihood

- Usually difficult to evaluate the integral

$$p(y) = \int p(y|\theta)p(\theta)d\theta = E_{p(\theta)}[p(y|\theta)].$$

- A (naive) first try is to draw from the prior $\theta^{(1)}, ..., \theta^{(N)}$ and estimating the marginal likelihood by the average likelihood

$$\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|\theta^{(i)}).$$

  Unstable if the posterior is very different from the prior.
- Importance sampling. Let $\theta^{(1)}, ..., \theta^{(N)}$ be iid draws from some density $g(\theta)$.

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta = \int \frac{p(y|\theta)p(\theta)}{g(\theta)}g(\theta)d\theta \\ &= \mathrm{E}_g\left[\frac{p(y|\theta)p(\theta)}{g(\theta)}\right] \approx N^{-1}\sum_{i=1}^{N}\frac{p(y|\theta^{(i)})p(\theta^{(i)})}{g(\theta^{(i)})}. \end{aligned}$$

Computing the marginal likelihood, cont.

- Rearrangement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

- Problem: we must know the posterior, **including** the normalization constant. The $\propto$ trick does not work here.
- But we only need to know $p(\theta|y)$ in a single point θ_0.
- Kernel density estimator may be used to approximate $p(\theta_0|y)$. Unstable. Chib (1995, JASA) and Chib-Jeliazkov (2001, JASA) provide better solutions.
- Reversible Jump MCMC (RJMCMC) for model inference.
  - MCMC methods can be extended to not only move in the parameter space for a given model, but also jumping between models.
  - The proportion of iterations spent in model k is an estimate of $\Pr(M_k|y)$.

Approximate marginal likelihoods

- Taylor approximation of the log posterior

$$\ln p(\mathbf{y}|\theta)p(\theta) \approx \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{1}{2}J_{\hat{\theta},\mathbf{y}}(\theta - \hat{\theta})^2,$$

$$p(\mathbf{y}|\theta)p(\theta) \approx p(\mathbf{y}|\hat{\theta})p(\hat{\theta}) \exp\left[-\frac{1}{2}J_{\hat{\theta},\mathbf{y}}(\theta - \hat{\theta})^2\right],$$

  which can be integrated analytically w.r.t. θ, using properties of the multivariate normal pdf.
- The Laplace approximation:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{1}{2}\ln\left|J_{\hat{\theta},\mathbf{y}}^{-1}\right| + \frac{p}{2}\ln(2\pi),$$

  where p is the number of unrestricted parameters in the model.
- Cruder version of the Laplace: The SBC (BIC) approximation

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{p}{2}\ln n.$$

Bayesian model inference - a critique

- Bayes factors (model probabilities) are very sharp inference objects. Handle with care.
- Minor differences in the prior can lead to large differences in the Bayes factor, especially in high-dimensional non-linear models.
- Continuous model expansion is usually a better alternative, when feasible.
- Improper priors cannot be used to compute Bayes factors. Several tricks have been developed to handle this, but they are non-Bayesian.
- Bayes factors are relative measures, all models under consideration may be bad approximations to the data.
- Bayes model probabilities essentially assume the true data generating process (DGP) is among the compared models. Box: All models are false, but some are useful.
- When none of the compared models are true: $Pr(M_i|y) \to 1$ for the model which is closest to the DGP in the Kullback-Leibler sense. Putting all eggs in one basket is not always a good idea.