Statistical Methods - Model evaluation, comparison and selection

Lecture 10

Mattias Villani

Sveriges Riksbank and Stockholm University

May 24, 2010

Overview Lecture 10

- Bayesian model inference
- Cross-validation

Bayesian 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.

Bayesian model comparison, cont.

- 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

$$p_i(y) = \int p_i(y|\theta_i) p_i(\theta_i) d\theta_i.$$

- 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 example

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

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

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

$$p(x_1, ..., x_n | M_0) = \theta_0^y (1 - \theta_0)^{n-y},$$

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

where $y = \sum_{i=1}^n x_i$, and $B(\cdot, \cdot)$ is the Beta function, and we have used that the Beta density integrates to one.

- Posterior model probabilities

$$Pr(M_i | x_1, ..., x_n) \propto p(x_1, ..., x_n | M_i) Pr(M_i), \ \ i = 0, 1.$$

- The Bayes factor

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

Bayesian hypothesis testing, cont.

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

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

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

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

  Then it can be shown that:

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

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

Bayesian 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.
- Example: What determines public expenditures?

Marginal likelihood and out-of-sample 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 $p(\theta|y_1, ..., y_{t-1})$ is the posterior distribution of θ using data up to time t. We have assume that y_t is independent of $y_1, ..., y_{t-1}$ conditional on θ.

- Note: the prediction of y_1 is based on the prior of θ, whereas the prediction of y_T uses almost all the data to infer θ and is therefore very little influenced by the prior.

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 $p_i(\gamma|y)$ is the marginal posterior of γ conditional on model i.

- A particular case is when $\gamma = (y_{T+1}, ..., y_{T+h})$' which means that we can get a predictive distribution that includes three sources of uncertainty:
  - Future errors/disturbances (e.g. the ε's in a regression)
  - Parameter uncertainty (the predictive distribution $p(y_{future}|y_{known})$ has the parameters integrated out by their posteriors)
  - Model uncertainty (by model averaging)

Computing the marginal likelihood

- Usually difficult to evaluate the integral

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

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

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

  Not stable if the posterior is very different from the prior (usual case).

- Bayes theorem may be rearranged to give the identity:

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

  The problem is that $p(\theta|y)$ is rarely known (including normalization constant), except in very simple models. Chib (1995, JASA) proposes a smart scheme based on Gibbs sampling draws to overcome this problem.

Computing the marginal likelihood, cont.

- The Laplace approximation:

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

  where $I(\hat{\theta})$ is the Fisher information evaluated at the posterior mode $\hat{\theta}$, k is the number of unrestricted parameters of the model.
- Cruder version of the Laplace: The SBC (BIC) approximation

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

- 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 i is an estimate of $\Pr(M_i|y)$. Reversible Jump MCMC (RJMCMC).

Marginal likelihoods - 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.
- As $N \to \infty$, $Pr(M_j|y) \to 1$ for the model that is closest to the true data generating process (in the Kullback-Leibler sense). This is sometimes a good property, but it also means that we will eventually (i.e. when $N \to \infty$) end up chosing one, and only one, model in the model set, even when no model is true. Usually better to combine many of the models, but the logic of the marginal likelihood does not allow it.

Cross-validation

- Cross-validation directly estimates the expected prediction error

$$Err = \mathrm{E}[L(Y, \hat{f}(X))].$$

- K-fold crossvalidation splits the data in K equally size parts. Uses $K - 1$ of the partitions to estimate the model and then predicts the left-out partition.

Validation	Training	Training	Training	Training
Training	**Validation**	Training	Training	Training
Training	Training	**Validation**	Training	Training
Training	Training	Training	**Validation**	Training
Training	Training	Training	Training	**Validation**

Table: The data splits in 5-fold cross-validation.

- Cross-validation can estimate *Err* since the validation/test data is never used in the training.

Cross-validation, cont.

- The cross-validation estimate of the prediction error can be written

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i)),$$

  where $\kappa(i) : \{1, ..., n\} \to \{1, K\}$ is an indexing function that indicates which of the partitions each observation belongs to, and $\hat{f}^{-k}(x)$ denotes the fitted function using all the data in fitting except the observations in the kth partition.

- This is a very general method, any model can be used as long a the predictions $\hat{f}(x)$ can be computed.

How to choose K?

- Large K:
  - uses a lot of data in each training and therefore has **little bias** in estimating the true prediction error.
  - all the training data sets are very similar, so the estimate typically has **large variance.**
- Small K:
  - uses less data in each training sample and risks that the learning curve of estimator is still very steep $\rightarrow$ Bias.
  - training data sets are likely to more different $\rightarrow$ smaller variance.
- Drawback of cross-validation. Paralell computations helps a lot!

Leave-one-out and generalized cross-validation

- $K = N$ is usually referred to as leave-one-out crossvalidation. Requires N estimations. Computationally demanding.
- Computational short-cuts for leave-one-out for some linear smoothers and quadratic loss ($\hat{y} = Ly$):

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{f}(x_i)}{1 - L_{ii}}\right)^2$$

- For some models (or at least approximately) this simplifies further to

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{f}(x_i)}{1 - N^{-1}\mathrm{tr}(L)}\right)^2,$$

  which is usually (strangely) called **generalized cross-validation** (**GCV**).
- Note that both these versions only uses the fit from the whole sample and the smoother matrix.