MSA101/MVE187 2018 Lecture 14

Petter Mostad

Chalmers University

October 18, 2018

Outline

- Model choice using context knowledge (recommended).
- Some other model choice ideas (less recommended).
- Model choice without using context knowledge: Aiming for simplicity.

Building Bayesian models – a recommended way to work

- Any data analysis should start with some data exploration!
- If possible, take as starting point your *understanding* of the process that has generated the data.
- Set up a model (a Bayesian Network?) with arrows pointing in the direction of *causality*.
- It should contain: Variables y you have *observed* (data), variables y_{NEW} you want to *predict*, and additional variables which may be divided into *parameters* θ and *augmented data* (or "extra variables") z, where we often will have $y_{NEW} \coprod z \mid \theta$.
- My advice: Better to build a good model with extra variables z than an ad-hoc model directly connecting y and y_{NEW} via θ.
- Generally, use "uninformative priors".
- When there is concrete documentable additional information about, say, θ, one may use it to create an informative prior.

Uninformative priors

- We want an "uninformative prior" on a parameter θ to represent "no knowledge". Unfortutately, it is not mathematically clear how this should be best defined.
- We have often used "flat" priors; however, a flat prior may not stay flat if θ is re-parametrized.
- If μ is a "location" parameter, you might use $\pi(\mu) \propto 1$; if λ is a "scale" parameter, a good alternative may be $\pi(\lambda) \propto 1/\lambda$.
- A number of theories have been developed. Some aim for maximized entropy. However, we will not go into these theories.

Model choice

When building a Bayesian model, how does one choose between different options? Some tools:

- Use the prior predictive and compare with contextual knowledge.
- Use the posterior predictive and compare with contextual knowledge.
- Aim for “simplicity”, informally or formally using “information criteria”.
- As far as possible, investigate robustness with respect to changes in the model (in particular changes in prior distributions).

Comparing the prior predictive with context knowledge

- The prior model should represent "prior knowedge": A way to check that it does this correctly is to simulate new data from the prior predictive and check if they look like what you expect a priori.
- Examples
  - Simulate from the prior of a stochastic model for tree growth.
  - Simulate from the prior of a stochastic model for geological faults.
  - Simulate from the prior of a stochastic model for image noise.
- Example: If one believes some unobserved quantities should follow some distribution, one may compute or simulate their quantiles in this distribution: They should then be uniformly distributed. (Example: Prior predictive p-values).
- This is closely connected to cross validation: From data $x_1, \ldots, x_n$, use all but x_i to fit the model and use the fitted model to predict x_i.

Comparing the posterior predictive with contet knowledge

- The prior will indicate that some "features" of the model can be "informed" by the data, while other "features" are fixed. Are there "features" that are fixed that need to be informed by the data? This can be investigated by comparing simulations from the posterior predictive with the actual data. Are there systematic differences?
- Very simple example:
  - Data, 4.33, 4.32, 4.35, 4.30.
  - Model: $y_i \sim \text{Normal}(\mu, \sigma^2)$.
  - If the prior is $\mu \sim \text{Normal}(0, 100)$, $\sigma^2 = 1$, simulations from the posterior predictive will have too much spread in the data.
  - If the prior is $\mu = 0$, $\pi(\sigma^2) \propto 1/\sigma^2$, simulations from the posterior predictive will hav both wrong mean and wrong spread.
- Posterior predictive p-values
- Heart transplant example in chapter 7 of Albert.

"Model choice" using weighed models and Bayes Factors

- Consider the theory for mixtures previously presented: Instead of mixtures of conjugate distributions, one can use mixtures of *any* set of priors $g_i(\theta)$. The prior predictives $f_i(x)$ and the posteriors $g_i'(\theta)$ exist, even if they may be difficult to compute.
- We get that the posterior is a mixture of the corresponding posteriors, with weights updated using the prior predictive values $f_i(x)$ for the data.
- If we have only $k = 2$ priors, with weights α_1 and $\alpha_2 = 1 - \alpha_1$, and if we denote the posterior weights α_1' and $\alpha_2' = 1 - \alpha_1'$, we get

  $$\frac{\alpha_1'}{1-\alpha_1'} = \frac{f_1(x)}{f_2(x)} \cdot \frac{\alpha_1}{1-\alpha_1}$$

  i.e., the posterior odds $\alpha_1'/(1-\alpha_1')$ is equal to the likelihood ratio $f_1(x)/f_2(x)$ times the prior odds $\alpha_1/(1-\alpha_1)$.
- $f_1(x)/f_2(x)$ is called the *Bayes factor*.

Difficulties using Bayes Factors for model choice

- Instead of determining prior weights for the models, one may compare the model likelihoods: If one is "sufficiently big", one may decide to go with only this model. (A practical alternative to using Hypothesis Testing for model selection).
- Alternatively, one may go on with a weighted mean of the models, but then actual prior weights must be determined. May be particularly difficult to do when the models are structurally different.
- Improper priors may cause difficulties in the setup above.
- Improper priors should not be replaced with "vague" priors for model comparison purposes!
- Main problem: You have to first come up with the list of "possible" models, before you can do model selection using Bayes factors!

Informal model checking: Hypothesis testing

- A practical problem with model comparison via Bayes factors is that both (or all) models need to be completely specified.
- Hypothesis testing lets you compare a model with an alternative that deviates from it in the direction measured by the test statistic, but may otherwise be unspecified.
- Thus, hypothesis testing can be used in Bayesian statistics as a way to indicate alternative models.
- Over-interpretation of p-values must be avoided.

Robustness

- Another approach to the choice of models and priors: Check if switching between different choices matters for the final result.
- NOTE: For any posterior, there exists a prior that will give this posterior (assuming nonzero densities).
- Revised question: Do *reasonable* changes in the prior affect the result much?
- If not, the prior is called *robust* for this likelihood.

The problem of “overfitting” often considered in connection to model choice

- A large portion of frequentist statistics deals with the problem of “overfitting”: The fitted model fits the old data so well that it does not predict new data well.
- In Bayesian statistics, replacing maximum likelihood parameter estimation with the use of posterior distributions for parameters will often alleviate such problems.
- However, a Bayesian version of “overfitting” can occur if “overusing” the data to construct the model (i.e., either the likelihood or the prior).

Comparing models based on their complexity

- For various reasons, one might want to build a stochastic model for some data *without* using (much) context knowledge.
- Example: Neural network models.
- The idea then is to weigh the complexity of the model against the likelihood of the data under the model.
- Use of *information criteria* that *penalize* the complexity of a model:
-
  - AIC, Akaike Information Criterion.
  - BIC Bayesian Information Criterion.
  - DIC Deviance Information Criterion.
  - ...

Model selection: Learning graphical network models from data

- Given a set of observations of a set of variables, one may assume this is a (random) sample from a joint distribution, and one may try to learn a reasonable set of conditional independencies from the data.
- More concretely, an algorithm produces one (or several) graphical models from the data.
- In principle, the same issues as for Bayesian model selection, or any other model selection, apply.
- Problem: There are a huge number of possible graphs for a moderately long list of variables.
- Important field of research.