MSA101/MVE187 2022 Lecture 15a
Applying Bayesian statistics

Petter Mostad

Chalmers University

October 17, 2022

Overview

- Applied Bayesian modelling
- Model selection
- Connections between Bayesian Learning and Machine Learning
- An example of a paper using Bayesian modelling (separate overheads).

Applied Bayesian modelling

1. Decide on a set of variables whose relationship models the core of your situation. The should include variables representing data, and variables for things you want to predict.
2. Formulate a joint model.
3. Is the model appropriate?
   - Check your model!
   - Compare different models!
4. Now you want to compute or approximate your prediction in the model conditional on data. Purely a computational problem!

How to construct models

General advice:

- Use what you believe are cause and effect to guide your model specification: The *effect* of something is modelled as a stochastic variable conditional on the things that *caused* it.
- Write down a corresponding Bayesian network to get an overview!
- Examples...

How to check models

- Checking a single model, whether it is "reasonable":
  - Simulating posterior predictive values!
  - Simulating prior predictive values!
  - Simulate some variables that it is easy to have an opinion about!
- Examples ...

Bayesian model comparison

- Assume you are considering n different models connecting your data Y_d with your prediction Y_p.
- Let λ have possible values $1, \ldots, n$ and let $\pi(Y_p, Y_d \mid \lambda = i)$ indicate model i.
- If you specify a prior belief in each model, you can use a combined *weighted model*

$$\pi(Y_p, Y_d) = \sum_{i=1}^{n} \pi(\lambda = i)\pi(Y_p, Y_d \mid \lambda = i)$$

  with weights $w_i = \pi(\lambda = i)$.
- We get

$$\begin{aligned}
\pi(Y_p \mid Y_d) &= \frac{\pi(Y_p, Y_d)}{\pi(Y_d)} = \frac{\sum_{i=1}^{n} \pi(\lambda = i)\pi(Y_d \mid \lambda_i)\pi(Y_p \mid Y_d, \lambda_i)}{\sum_{j=1}^{n} \pi(Y_d \mid \lambda = j)} \\
&= \sum_{i=1}^{n} \left(\frac{\pi(\lambda = i)\pi(Y_d \mid \lambda = i)}{\sum_{j=1}^{n} \pi(\lambda = j)\pi(Y_d \mid \lambda = j)} \right) \pi(Y_p \mid Y_d, \lambda = i)
\end{aligned}$$

Bayesian model comparison

- The prediction $\pi(Y_p \mid Y_d)$ using the weighted model uses a weighting of the predictions $\pi(Y_p \mid Y_d, \lambda = i)$ from each individual model, where the weights are updated from $w_i = \pi(\lambda = i)$ to

$$w_i' = \frac{\pi(\lambda = i)\pi(Y_d \mid \lambda = i)}{\sum_{j=1}^{n} \pi(\lambda = j)\pi(Y_d \mid \lambda = j)}.$$

- The value $\pi(Y_d \mid \lambda = i)$ is the probability of observing the data Y_d given model i.
- Except the notation, formulas are exactly the same as when using *mixtures of conjugate priors* (see Lecture 3).
- If one posterior weight w_i' is close to 1, we may approximate by *discarding* all models but model i. *The procedure becomes a model selection procedure*.
- Note: When $n = 2$ we get that $w_2'/w_1' = w_2/w_1 \cdot \pi(Y_d \mid \lambda = 2)/\pi(Y_d \mid \lambda = 1)$.
- To use the formulas in practice, we need to be able to compute $\pi(Y_d \mid \lambda = i)$ for all models i.

Bayesian model comparison

- Note: The ideas above cannot be used (directly) to compare a model i with an *improper prior*: Then $\pi(Y_d \mid y = i)$ cannot be computed.
- Note: An improper prior should not be interpreted as a limit of a sequence of proper priors.
- Note: How to determine if models are good apriori? (How to determine prior weights w_i?)

Example of Bayesian model selection

- The data consists of counts c_i, $i = 1, \ldots, n$, with $S = \sum_{i=1}^{n} c_i$.
- Model 1: $(i = 1, \ldots, n)$

$$\begin{aligned} \lambda &\sim \text{Gamma}(1,1) \\ c_i \mid \lambda &\sim \text{Poisson}(\lambda) \end{aligned}$$

- Model 2: $(i = 1, \ldots, n)$

$$\begin{aligned} p &\sim \text{Uniform}(0,1) \\ \lambda_0, \lambda_1 &\sim \text{Gamma}(1,1) \\ \pi(c_i \mid p, \lambda_0, \lambda_1) &= p \, \text{Poisson}(c_i; \lambda_1) + (1-p) \, \text{Poisson}(c_i; \lambda_0) \end{aligned}$$

- **Break to compute** $\log \pi(c \mid$ **Model 1**$)$**.**
- **Break to compute** $\log \pi(c \mid$ **Model 2**$)$**.**
- As $\pi(c \mid \text{Model 2}) / \pi(c \mid \text{Model 1})$ is very large, we see that the second model fits the data much better. Overwhelms any reasonable value for w_2/w_1!

Example: Continued

- Consider Model 3:

$$\pi(c_i) = \hat{p}\,\mathrm{Poisson}(c_i; \hat{\lambda}_1) + (1-\hat{p})\,\mathrm{Poisson}(c_i; \hat{\lambda}_0)$$

  where $(\hat{p}, \hat{\lambda}_0, \hat{\lambda}_1)$ is the mode of the `logpost` function.
- We get

$$\log \pi(c \mid \text{Model 3}) = \mathrm{logpost}(\hat{p}, \hat{\lambda}_0, \hat{\lambda}_1)$$

  where `logpost` is the function we programmed in R.
-

$$\pi(c \mid \text{Model 3})/\pi(c \mid \text{Model 2})$$

  becomes larger than 1. So should model 3 be preferred to model 2?
- NO: The **prior** probability for Model 3 is quite low, so w_3/w_2 should cancel out the factor above.
- Ignoring this leads to **overfitting**, a serious problem in non-Bayesian statistics.

Advice on statistical modelling

- Always start with data and a clear question.
- Always plot and explore your data, so you understand it as best you can.
- Understand the known science of what is going on as best as you can, to make a realistic model.
- In complicated models:
    1. Start with a Bayesian Network for variables needed to describe a model. Use causality as a guide!
    2. *Then* choose either fixed distributions, or distributions with uncertain parameters, to relate the variables.
- *Elicitation* for constructing informative priors. (Example: Use of `beta.select` in LearnBayes package).

Comparing Bayesian learning and machine learning (ML)

- Bayesian statistics and computation is an important part of ML technology.
- Bayesian inference of various types, e.g., Variational Bayes, has been used as a way to learn about weights in a neural network.
- However, the Bayesian paradigm, as used in this course, is generally not used in ML.

A possible way to connect ML with the Bayesian paradigm

- For concreteness, we look at the basic problem of classifying digits (0 - 9) from images, using the MNIST data set.
- Using the Bayesian paradigm, Y_{data} is the set of images and their classifications, and Y_{pred} is the classification of a new image. We want to define a joint distribution on these, and then use $\pi(Y_{pred} \mid Y_{data})$.
- Using ML, you may for example choose a neural network ending with a softmax layer used to give probabilities for the 10 classification outcomes. You also choose a particular stochastic algorithm for training of that network, to obtain a single neural network, which you then use for prediction.
- Is it possible to compare or connect the two approaches?

A possible way to connect ML with the Bayesian paradigm

- The neural network parameters should be identified with θ, the parameter of the Bayesian model.
- The likelihood defined by the data is the same in both approaches. We also have conditional independence of the observations, and of any new prediction, given the parameter θ.
- In Bayesian inference one would find a posterior for θ (i.e., a posterior on the set of networks) and average over it for predictions.
- In ML one uses (most often) a single network for predictions.
- To make a comparison, we assume the Bayesian approach is to sample a *single* $\hat{\theta}$ from the posterior.
- The Bayesian approach will sample $\hat{\theta}$ from a distribution whose logdensity is
$$\mathsf{Loglikelihood}(\theta) + \mathsf{Prior}(\theta) \tag{1}$$
where in ML `Loglikelihood` is the negative of the `Loss` and `Prior` is the negative of a regularization term.
- By comparison, ML will use a similar Equation 1 and a stochastic algorithm, but also test- and validation-data, to produce a NN $\hat{\theta}$.

A possible way to connect ML with the Bayesian paradigm

1. Given an NN, can we establish a clear correspondence

$$\text{Prior}(\theta) \text{ functions} \leftrightarrow \text{Stochastic ML algorithm producing } \hat{\theta}$$

2. Is such a correspondence of practical use when developing new algorithms / models?

- Note: Priors need to be more advanced than currently used regularization terms.
- Note: Simulation in the posterior is not straight-forward in the relevant high dimensions.