Bayesian Statistics[1]

Prof. Hans-Rudolf Künsch, SS 2014
19.08.2014, 14.00-14.20

Summary The exam took place in the Professor's office; I was given paper to write on and a pen. I sat between Professor and Assistant and the Assistant was taking notes for the protocol. The atmosphere was relaxed and the Professor was calm: he gave me feedback every time I did or wrote correct statements. He came to pick me from the waiting room, let me kindly in the office and let me leave my bag and take a seat.

Description of the contents: Bayes Formula (Ch.1), Choice of Prior given Normal likelihoods with known and unknown variance (Ch.1,2), Bayesian Regression and Model Selection (Ch.3)[2]

Execution of the Exam (())

Prof: Bayes formula plays a central role in Bayesian Statistics. Can you write it down?

Me: Assume we are in discrete case:

$$\Omega = \cup_{i\in I} A_i, \quad I \text{ finite or countable set}$$
$$A_i \cap A_j = \emptyset$$
$$B \in \Omega, \ \mathbf{P}(B) > 0$$

Then Bayes formula is:

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(B|A_i)\mathbf{P}(A_i)}{\sum_k \mathbf{P}(B|A_k)\mathbf{P}(A_k)}$$

In the continuous case, assume we have two random variables X, Y and the joint and marginal densities:

$$\mathbf{P}(X = x, Y = y) = f_{X,Y}(x,y), \quad f(x) = \int_{\mathbb{R}} f(x,y)dy, \quad f(y) = \int_{\mathbb{R}} f(x,y)dx \tag{1}$$

Then the conditional probability of $X|Y$ is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f(y)} \Rightarrow f_{X,Y}(x,y) = f_{X|Y}(x|y)f(y) \tag{2}$$

[1] MSc Statistics, Specialisation Areas and Electives, Statistical and Mathematical Courses
[2] The chapter number refer to the script .

It follows that the joint probability is the product of the marginal times the conditional.
We exploit this result and obtain Bayes formula as follows:

$$f(y|x) = \frac{f(x,y)}{f(x)} \overset{(2)}{=} \frac{f_{X|Y}(x|y)f(y)}{f(x)} \overset{(1)+(2)}{=} \frac{f_{X|Y}(x|y)f(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y)f(y)dy}$$

Prof: Ok, it is correct. Do these formulas look like this in all cases?

Me: For the discrete case the formula holds if $\mathbf{P}(A_i) > 0$, otherwise $\mathbf{P}(A_i|B) \equiv 0$.

Prof: And in the continuous case? Look at the denominator: what is the probability of X being equal to ... (Professor said a number at random) ?

Me: That probability is of course zero because it is an event of Lebesgue measure zero.

Prof: Exactly, so note that we are conditioning on an event of probability zero but the formula holds provided the denominator is neither zero nor infinity. Suppose now we have:

$$X_1, \dots, X_n \sim N(\mu, 1)$$

and μ is unknown. How would you choose the prior for μ ?

Me: For easy calculation I would choose a normal prior for μ, as it would give rise to a conjugate model and hence a normal posterior. To obtain the posterior I multiply prior times likelihood:

$$\mu \sim N(\theta, \tau^2)$$
$$\pi(\mu|X_1, \dots, X_n) \propto \exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - \theta)^2\}$$

Prof: Don't do all the computations, just explain what the next steps are.

Me: Ok. What we have to do next is collecting together all terms containing μ and μ^2 and constructing the square of a binomial of the form $\alpha/2(\mu - \dots)^2$, where α is the inverse of the posterior variance and $\dots$ is the posterior mean.

Prof: It is correct. Suppose now that also the variance is unknown, i.e.

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

How would you now choose the prior for the unknown parameters?

Me: In this case, again to obtain a conjugate model, I would choose, for $\mu|\sigma^2$ a normal prior $N(\mu_0, \frac{\sigma^2}{n_0})$ and for σ^{-2} I would choose a Gamma prior $\Gamma(\gamma, \lambda)$ so that the product of these two

distributions gives the joint prior of the unknown parameters and we obtain as posteriors the same prior distributions but with updated parameters. (I wrote down only the posterior for $\mu|\sigma^2$)

Prof: We have seen during the course Bayesian Regression. Let's now discuss about Regression and Model Selection.

Me: Given the regression model

$$y = \alpha\mathbf{1} + X_\gamma\beta_\gamma + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where X_γ denotes the columns of the original data matrix X that have been selected (γ is a vector of $\{0,1\}^p$, p is the number of columns of X), the unknown parameters we have to put a prior on are: $\gamma, \beta_\gamma, \alpha, \sigma^2$. We fix γ and assume α to be independent a priori from β_γ and σ^2 and we take univariate Jeffreys priors for α, σ^2 (flat prior for α,inverse Gamma for σ^2). Hence we have that the joint prior is

$$\pi(\alpha, \beta_\gamma, \sigma^2) \propto \pi(\beta_\gamma|\sigma^2)\sigma^{-2}$$

Usual choice for the prior of β_γ is the g-prior of Zellner:

$$\pi(\beta_\gamma|\sigma^2) \sim \mathcal{N}(0, g\sigma^2(X_\gamma^T X_\gamma)^{-1})$$

where the prior gets uninformative as g goes to infinity.
Using these priors, after multiplication with the likelihood, we obtain the following posteriors for α, β_γ:

$$\beta_\gamma|y, \sigma^2 \sim \mathcal{N}(\frac{g}{g+1}\hat{\beta}_\gamma, \frac{g\sigma^2}{g+1}(X_\gamma^T X_\gamma)^{-1})$$
$$\alpha|y, \sigma^2 \sim \mathcal{N}(\hat{\alpha}, \frac{\sigma^2}{n})$$

where $\hat{\beta}_\gamma$ is the usual least squares estimator (and also MLE estimator for β_γ): $\hat{\beta}_\gamma = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y$, $\hat{\alpha}$ is the MLE estimator of α : $\hat{\alpha} = \bar{y}$, and n is the sample size.

Prof: Ok, posteriors are correct, you do not need to write down all the formulas. Let's discuss about the posterior mean of β_γ.

Me: The posterior mean is a convex combination of prior mean (1*0) and MLE ($g^*\hat{\beta}_\gamma$) for β. We see there is a shrinkage towards zero.

Prof: Good. Concerning g: how should we choose it? What happens if we choose $g \to \infty$?

Me: If we choose $g \to \infty$ we are approaching a non-informative prior for β_γ but we find a paradox in model selection. Namely, if we consider Bayes factor for doing model selection and compare a non-empty model γ (i.e. a model with some predictors) versus an empty model, we find that $B(\gamma, 0) \to 0$, which is known as Bartlett's paradox, that is in the limit we will always choose the empty model.

Prof: Which makes no sense. And if we keep g fixed?

Me: If we take g fixed, we could have again problems in model selection. In fact, in the limit of a perfect model ($R^2_\gamma \to 1$), the Bayes factor of our model γ vs. an empty model does not go to infinity but to a constant, differently from what we would expect. This case is known as Information paradox. Hence there is no good fixed choice of g a priori.

Prof: How could we then treat g?

Me: We can proceed in two ways: Empirical Bayes approach or putting a hyperprior on g. In the empirical Bayes we do as follows:

$$\hat{g} = \arg\max_g f(y|g,\gamma)$$

where $f(y|g,\gamma)$ is the likelihood, while in the fully Bayes approach we choose a prior as to make calculations easier to obtain the marginal likelihood:

$$f(y|\gamma) = \int f(y|\gamma,g)\pi(g)dg \propto \int (1+g)^{1/2}\pi(g)dg \to \infty \implies \pi(g) \propto (1+g)^{-\frac{a}{2}}, \qquad a \in (2,3]$$

Prof: And why does the interval of a have this form?

Me: So that we can avoid the Information paradox.

Prof: And also because for $a > 2$ the prior is integrable.
Ok, time is over. Thank you and have a good day.

Final Remarks The Professor has been very kind and calm for the entire exam. In some cases I was writing down formulas I knew by heart and he told me that I did not need writing them all: what really matters for him is understanding the procedures and the derivations and no that we learn all the formulas by heart.

Expected mark: 5.75
Received mark: 6