MATH 341 / 650.3 Spring 2020 Homework #2

Frank Palma Gomez

Friday 21st February, 2020

Problem 1

These are questions about McGrayne's book, chapters 4-7.

(a) [easy] Describe four things Bayesian modeling was applied to during WWII and identify the people who developed each application.

- Cracking Enigma Code, Alan Turing
- Bayesian Artillery, Andrei Kolmogorov
- Locating U Boats, Bernard Osgood Koopman
- Cryptography, Claude Shannon

(b) [harder] What do you think was the main reason Bayesian Statistics fell out of favor at the end of WWII?

There was a lack of recognition to the people who applied the methods. Publications were not published until post war and the public didn't get the chance to see the problems that Bayesian Statistics was solving.

(c) [harder] Why weren't the leaders of Statistics world in the 1950's able to answer the think-tank's question about the $\mathbb{P}$(war in the next 5 years)?

Because they were under interpretation that probability applies to a long sequence of repeatable events and that having was in the next 5 years was just a unique situation. Multiple academic papers were published years after the war ended and by that time the people using the Bayesian approach had passed away.

(d) [easy] Who was responsible for reviving the interest in Bayesian Statistics post-WWII and why?

Author Bailey became the first to challenge the anti-Bayesian status quo. Author Baily was using Bayesian Statistics to determine the new years insurances premium rates.

(e) [difficult] In 1955, there were no midair collisions of two planes. How was the actuary able to estimate that the number would be above zero?

(f) [easy] The main attack on Bayesian Statistics has always been subjectivity. Answer the following question how Savage would have answered it: "If prior opinions can differ from one researcher to the next, what happens to scientific objectivity in data analysis?" Do you believe Savage's idea is the way science works in the real world?

Yes because there exists a correlation between the amount of data and the agreence of opinions. As the amount of data increases, the subjectivity between person to person differs by the slightest. In the other hand, without any concrete evidence one can make an assumption based on its beliefs if the data is not sufficient enough.

(g) [difficult] [MA] On page 104, Sharon writes, "Bayesians would also be able to concentrate on what happened, not on what *could* have happened according to Neyman Pearson's samping plan". (Note that the "Neyman Pearson's samping plan" is synonymous with Frequentist Statistics). Explain (1) how Bayesians concentrate on "what happened" and (2) how Frequentists concentrate on what "*could* have happened" in the context on page 104.

(h) [easy] Who were the two tireless champions of Bayesian Statistics throughout the 50's, 60's and 70's and where geographically were they located during the majority of their career?

- Jimmie Savage, The University of Chicago
- Dennis Lindley, Cambridge University

Problem 2

We will now be looking at the beta-prior, binomial-likelihood Bayesian model and introduce credible regions as well.

(a) [easy] Using the principle of indifference, what should the prior on θ (the parameter for the Bernoulli model) be?

$$\mathbb{P}(\theta) = U(0,1) = \text{Beta}(1,\ 1)$$

(b) [harder] [MA] Can any discrete distribution satisfy the principle of indifference? Prove or disprove.

(c) [easy] Let's say $n = 6$ and your data is $0, 1, 1, 1, 1, 1$. What is the likelihood of this event?

Applying the principle of indifference, if each event is equaly likely and $n = 6$ then the likelihood is :

$$(\frac{1}{2})^6 = \frac{1}{64}$$

(d) [easy] Does it matter the order as to which the data came in? Yes/no.

No

(e) [harder] Show that the unconditional joint probability (the denominator in Bayes rule) is a beta function and specify its two arguments.

Recall the beta function,

$$B(\alpha, \beta) = \int_{-\infty}^{\infty} t^{\alpha-1}(1-t)^{\beta-1}\, \mathrm{d}t$$

Therefore,

$$\mathbb{P}(X) = \int_0^1 \mathbb{P}(X|\theta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = B(\alpha, \beta)$$

(f) [harder] Calculate this beta function explicitly.

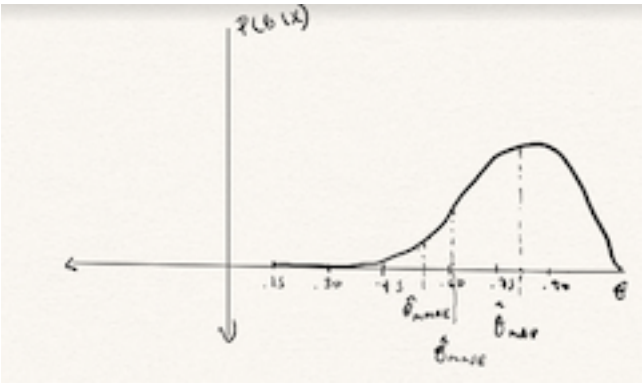
$$\int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} d\theta = \int_0^1 \theta^5 - \theta^6 d\theta = \frac{1}{42}$$

(g) [harder] Put your answers together to find the posterior probability of θ given this dataset. Do not use the beta function in your answer. Plot this posterior density function as best as you can.

Assuming the principle of indifference where the prior is $\mathbb{P}(\theta) = \text{Beta}(1,\, 1)$

$$\text{Beta}\left(\sum x_i + \alpha,\, n - \sum x_i + \beta\right) = \text{Beta}(5+1,\, 6-5+1) = \text{Beta}(6,\, 2)$$

(h) [easy] Sketch / plot / illustrate this posterior density function as best as you can by hand. Indicate where the $\hat{\theta}_{\text{MAP}}$, $\hat{\theta}_{\text{MMSE}}$ and $\hat{\theta}_{\text{MMAE}}$ estimates for θ are on the illustration using vertical lines as best as you can. No need to calculate them now explicitly. Just eyeball it on your drawing.

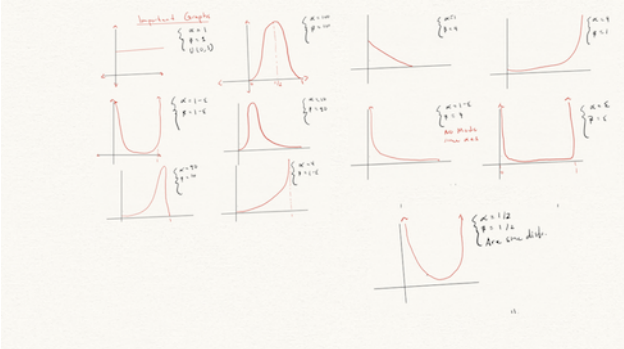


(i) [harder] Show that the posterior is a beta distribution and specify its parameters.

$$\text{Beta}\left(\sum x_i + \alpha,\, n - \sum x_i + \beta\right) = \text{Beta}\left(5+1,\, 6-5+1\right) = \text{Beta}\left(6,\, 2\right)$$

(j) [difficult] [MA] Prove that the posterior expectation is the mean squared error optimal estimator given your prior and the data.

(k) [easy] Now imagine you are not indifferent and you have some idea about what θ could be a priori and that subjective feeling can be specified as a beta distribution. (1) Draw the basic shapes that the beta distribution can take on, (2) give an example of α and β values that would produce these shapes and (3) write a sentence about what each one means for your prior belief. These shapes are in the notes.



(l) [harder] Imagine n data points of which you don't know the realization values. Using your prior of $\theta \sim \text{Beta}\left(\alpha,\, \beta\right)$, show that $\theta \mid X \sim \text{Beta}\left(\alpha + x,\, \beta + (n - x)\right)$. Note that $x := \sum_{i=1}^{n} x_i$ which is the total number of successes and thereby $n - x$ is the total number of failures.

If $\theta \sim \text{Beta}\left(\alpha,\, \beta\right)$ then,

$$\mathbb{P}\left(\theta|X\right) \propto \mathbb{P}\left(X|\theta\right)\mathbb{P}\left(\theta\right)$$

$$= \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)\theta^{\alpha-1}(1-\theta)^{\beta-1}}$$

$$= \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

$$\mathbb{P}(\theta|X) = \text{Beta}(\alpha + x,\ \beta + n - x)$$

(m) [easy] What does it mean that the beta distribution is the "conjugate prior" for the binomial likelihood?

It means that if we started with a beta prior, we get a distribution as a posterior.

(n) [difficult] Show that if $Y \sim \text{Beta}(\alpha,\ \beta)$ then $\mathbb{V}\text{ar}[Y] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Recall,

$$\mathbb{V}\text{ar}[Y] = E(x^2) - [(E(x)]^2 = \int_0^0 x^2 f_x(x) d\boldsymbol{x} - \int_0^1 x f_x(x) d\boldsymbol{x}$$

$$\int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\text{Beta}(\alpha,\ \beta)} d\boldsymbol{x} - \frac{\alpha^2}{(\alpha+\beta)^2}$$

$$= \frac{\text{Beta}(2+\alpha,\ \beta)}{\text{Beta}(\alpha,\ \beta)} - \frac{\alpha^2}{(\alpha+\beta)^2}$$

$$= \frac{\Gamma(2+\alpha)\Gamma(\beta)}{\Gamma(2+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2}$$

$$\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2}$$

$$= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\mathbb{V}\text{ar}[Y] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

(o) [E.C.] Prove that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

$$\boldsymbol{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} = \frac{(x-1)!(y-1)!}{(x+y-1)!}$$

$$\Gamma(s) = \int_0^\infty x^{s-1}\exp(-x)\, d\boldsymbol{x}$$

$$\Gamma(m)\Gamma(n) = \int_0^\infty x^{m-1}\exp(-x)\, d\boldsymbol{x} \int_0^\infty y^{n-1}\exp(-y)\, d\boldsymbol{y}$$

$$= \int_0^\infty \int_0^\infty x^{m-1}y^{n-1}\exp(-(x+y))\, d\boldsymbol{x}d\boldsymbol{y}$$

$$x = vt, y = v(1-t)$$

$$= \int_0^\infty (vt)^{m-1}(v(1-t))^{n-1}\exp(-(vt+v(1-t))\, d\boldsymbol{x}d\boldsymbol{y}$$

$$\Gamma(m)\Gamma(n) = \int_0^\infty t^{m-1}(1-t)^{n-1}dt \int_0^\infty v^{m+n-1}\exp(-v)\, dv$$

$$\Gamma(m)\Gamma(n) = B(m, n)\Gamma(m+n)$$

(p) [harder] The posterior is $\theta \mid X \sim \text{Beta}(\alpha + x, \beta + (n-x))$. Some say the values of α and β can be interpreted as follows: α is considered the prior number of successes and β is considered the prior number of failures. Why is this a good interpretation? Writing out the PDF of $\theta \mid X$ should help you see it.

If θ is considered the prior number of successes and β is considered the prior number of failures then anything that is added to each must be the same unit. This is a good interpretation because it is consistent with the way we add the known successes to θ and the know failure to β respectively

(q) [harder] If you employ the principle of indifference, how many successes and failures is that equivalent to seeing a priori?

One pseudo success and one pseudo failure.

$$U(0,1) = B(1,1)$$

(r) [easy] Why are large values of α and/or β considered to compose a "strong" prior?

α and β are pseudo successes and pseudo failure respectively. If θ and β are large then the stronger your belief is. At the end, if you have enough data, the data "swamps the prior" if your prior is weak.

(s) [harder] [MA] What is the weakest prior you can think of and why?

Using a uniform prior because it gives 50/50 odds to each event.

(t) [difficult] I think a priori that θ should be expected to be 0.8 with a standard error of 0.02. Solve for the values of α and β based on my a priori specification.

Let π_0 be expectation and σ^2 the stardard error, respectively.

$$\alpha = \frac{\pi_0(\pi_0(1-\pi_0) - \sigma^2)}{\sigma^2} = \frac{0.8(0.8(1-0.8) - 0.02^2)}{0.02^2} = 319.2$$

$$\beta = \frac{\pi_0(1-\pi_0) - \sigma^2}{\sigma^2} - \alpha = \frac{(0.8)(1-0.8) - (0.02)^2}{(0.02)^2} - 319.2 = 79.8$$

(u) [easy] Assume the dataset in (b) where $n = 6$. Assume $\theta \sim \text{Beta}\,(\alpha = 2, \beta = 2)$ a priori. Find the $\hat{\theta}_{\text{MAP}}$, $\hat{\theta}_{\text{MMSE}}$ and $\hat{\theta}_{\text{MMAE}}$ estimates for θ. For the $\hat{\theta}_{\text{MMAE}}$ estimate, you'll need to obtain a quantile of the beta distribution. Use `R` on your computer or online using rextester. The `qbeta` function in `R` finds arbitrary beta quantiles. Its first argument is the quantile desired e.g. 2.5%, the next is α and the third is β. So to find the 97.5%ile of a $\text{Beta}\,(\alpha = 2, \beta = 2)$ for example you type `qbeta(.975, 2, 2)` into the `R` console.

$$\hat{\theta}_{\text{MAP}} = Mode[\theta|X] = \frac{\sum x_i + \alpha}{n - \sum x_i + \beta} = 0.75$$

$$\hat{\theta}_{\text{MMSE}} = E[\theta|X] = \frac{\sum x_i}{n + \theta + \beta} = 0.70$$

$$\hat{\theta}_{\text{MMAE}} = \texttt{qbeta}(0.5, \sum x_i + \alpha, n - \sum x_i + \beta) = \texttt{qbeta}(0.5, 7, 3) = 0.714$$

(v) [harder] Why are all three of these estimates the same?

Due to the small size of the dataset

(w) [easy] Write out an expression for the 95% credible region for θ. Then write out the answer using the `qbeta` function from the `R` language.

$$CR_{\theta,95\%} = [Quantile(2.5\%, 7, 3), Quantile(97.5\%, 7, 3)]$$

$$CR_{\theta,95\%} = [\texttt{qbeta}(.025, 7, 3), \texttt{qbeta}(.975, 7, 3)] = [0.609, 0.925]$$

(x) [easy] Compute a 95% frequentist CI for θ.

$$CI_{\theta,95\%} = [0.833 + (1.960)\frac{0.4082}{\sqrt{6}}, 0.833 + (1.960)\frac{0.4082}{\sqrt{6}}] = [0.506, 1.159]$$

(y) [difficult] Let $\mu : \mathbb{R} \to \mathbb{R}^+$ be the Lebesgue measure which measures the length of a subset of $\mathbb{R}$. Why is $\mu(\text{CR}) < \mu(\text{CI})$? That is, why is the Bayesian Confidence Interval tighter than the Frequentist Confidence Interval? Use your previous answers.

Bayesian confidence intervals use the data to construct the interval whereas an Frequentist confidence interval does not take it into consideration and it assumes an asymptotic normality.

(z) [easy] Explain the disadvantages of the highest density region method for computing credible regions.

Highest Density Regions are computationally intense. Additionally, they are non-contiguous as they supply multiple possible locations for θ.

(aa) [harder] Design a prior where you believe $\mathbb{E}[\theta] = 0.5$ and you feel as if your belief represents information contained in five coin flips.

$$Beta(2.5, 2.5)$$

(bb) [harder] Calculate a 95% a priori credible region for θ. Use `R` on your computer (or rdrr.io online) and its `qbeta` function.

$$CR_{\theta,95\%} = [\texttt{qbeta}(0.025, 2.5, 2.5), \texttt{qbeta}(0.975, 2.5, 2.5)] = [0.123, 0.877]$$

(cc) [easy] You flip the same coin 100 times and you observe 39 heads. Calculate a 95% a posteriori credible region for θ. Round to the nearest 3 decimal points.

If we use the principle of indifference,

$$\mathbb{P}(\theta|X) = Beta(1 + 39, 100 - 39 + 1) = Beta(40, 60)$$

$$CR_{\theta,95\%} = [\texttt{qbeta}(0.025, 40, 60), \texttt{qbeta}(0.975, 40, 60)] = [0.307, 0.497]$$