Example sheet 2 hint sheet

Inference

Foundations of Data Science—DJW—2018/2019

This example sheet covers material up to Lecture 7 on 22 October. Questions 1(d), 3, 7(c), 8, and 9 are conceptually challenging. Questions 5(c) and 9 are mathematically involved.

Questions 1 and 2 are core Bayesian questions. Questions 4 and 6 are core frequentist questions. The rest are supplementary / mind-broadening.

Question 1. I sample $X_1, \ldots, X_n$ from Uniform$[0, \theta]$. The parameter θ is unknown, and I shall use $\Theta \sim \text{Pareto}(\theta_m, \alpha)$ as my prior, where $\theta_m > 0$ and $\alpha > 1$ are known:

$$\mathbb{P}(\Theta > \theta) = \begin{cases} (\theta/\theta_m)^{-\alpha} & \text{if } \theta \geq \theta_m \\ 1 & \text{if } \theta < \theta_m. \end{cases}$$

(a) What is the prior density of Θ?

(b) Find the posterior distribution for Θ.

(c) Find a 95% posterior confidence interval for Θ.

(d) Find a different 95% posterior confidence interval. Which is better? Why?

The keywords prior *and* posterior *tell you this is about Bayesian inference.*

For part (b), answer it like you answer every Bayesian question: write down the prior density $\Pr(\theta)$*, write down the data density* $\Pr(x_1, \ldots, x_n \mid \theta)$*, then use Bayes's rule i.e. multiply them and stick in a constant factor to get the posterior density* $\Pr(\theta \mid x_1, \ldots, x_n)$*. In this question, I recommend you keep track of bounds by using indicator functions,* $1_{\theta \geq \theta_m}$ *for the prior density and* $1_{0 \leq x_i \leq \theta}$ *for the data density. After you apply Bayes's rule, don't try to find the normalizing constant—just look at your posterior density function as a function of* θ*, recognize that you've seen it before, and write down the standard name and parameters.*

For part (c), you might perhaps have found $[0.025, 0.975]$ *quantile points of the posterior distribution. Why these, and not* $[0, 0.95]$ *or* $[0.05, 1]$*? Sketch the posterior density function, and sketch these confidence intervals, and this will suggest the answer to part (d).*

Question 2. I start with a prior belief that $\mu \sim \text{Normal}(\mu_0, \sigma_0^2)$. I then observe $x_1, \ldots, x_n$, which I take to be drawn from Normal(μ, ρ_0^2). Find my posterior distribution for μ, taking σ_0, μ_0, and ρ_0 as known. *Hint. The posterior distribution is also Normal, you just have to find the parameters.*

The keywords prior *and* posterior *tell you this is about Bayesian inference. The important thing is to write down the correct prior density, data density, and posterior density.*

When you have written the posterior density for μ*, remember: you want to simplify this expression* as a function of μ*, and you don't care about constant factors that don't involve* μ*. You should end up with* exp(*quadratic in* μ*), and you should try to simplify the quadratic.*

The answer is surprisingly simple and interpretable: the posterior mean is $A\mu_0 + (1 - A)\bar{x}$ *where* $\bar{x}$ *is the mean of the sample and* A *depends on the variances and on* n*. In other words, we shift our belief from* μ_0 *to something closer to the mean of the observed data.*

Question 3. I have a coin, which might be biased. I toss it n times and get x heads. To reflect my uncertainty about possible bias, my prior belief is that either the coin is unbiased (with prior probability $1 - \pi$); or it is biased (with prior probability π) in which case the probability of heads is $\Theta \sim \text{Beta}(\delta, \delta)$ with $\delta = 1$. The probability of seeing x heads is thus

$$\Pr(x \mid m, \theta) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{if } m = \text{biased} \\ \binom{n}{x}(1/2)^x(1-1/2)^{n-x} & \text{if } m = \text{unbiased} \end{cases}$$

Useful fact: the Beta(α, β) distribution has density

$$f(x) = \kappa x^{\alpha-1}(1-x)^{\beta-1}$$

where κ is

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

where m indicates which of the two possibilities is true, and my prior is

$$\Pr(m, \theta) \propto \pi^{1[m=\text{biased}]}(1-\pi)^{1[m=\text{unbiased}]}\theta^{\delta-1}(1-\theta)^{\delta-1}$$

$1[A]$ is another way to write the indicator function 1_A

(a) Find the posterior distribution of (M, Θ) given the data.
(b) Find $\mathbb{P}(M = \text{unbiased} \mid x)$, i.e. the posterior probability that the coin is unbiased.
(c) What is the posterior predictive probability that the next coin toss will be heads?

The discussion in section 3.3.3 may be helpful.

The keywords prior *and* posterior *tell you this is a question about Bayesian inference. The setup is long and convoluted, but part (a) is exactly like every other Bayesian question: write down the prior* $\Pr(m, \theta)$*, write down the density* $\Pr(x \mid m, \theta)$*, and multiply them together (with a constant factor) to get the posterior* $\Pr(m, \theta \mid x)$*. It's up to you whether to write the cases out longhand (as I did for* $\Pr(x \mid m, \theta)$*) or to use the indicator function notation (as I did for* $\Pr(m, \theta)$*). I use the longhand notation myself, except for situations where it gets too cumbersome.*

Part (b) is about nuisance parameters *as in page 35 of lecture notes. You've found* $\Pr(m, \theta \mid x)$*, and you want* $\Pr(m \mid x)$*, so integrate out* θ *to get the marginal distribution of* m*.*

Part (c) is about posterior predictive probabilities*, and it is the same style of calculation as on page 35 of lecture notes, using the law of total probability. Integrate out the parameters* (m, θ) *— or, to be precise, sum over* m *since it's discrete, and integrate over* θ *since it's continuous.*

Question 4. We are given a dataset $x_1, \ldots, x_n$ which we believe is drawn from Normal(μ, σ^2) where the parameters μ and σ^2 are unknown.

(a) Find the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$.
(b) Given $\delta_1 > 0$ and $\delta_2 > 0$, give pseudocode to compute

$$\mathbb{P}\Big(\sigma \in [\hat{\sigma} - \delta_1, \hat{\sigma} + \delta_2]\Big)$$

using parametric resampling, and also using non-parametric resampling.
(c) Give pseudocode to compute a 95% confidence interval for σ.

The keyword resampling *tells you this is a question about frequentist inference.*

For part (b), the first thing to ask yourself is: what is the random variable in this probability expression? It's not σ *since (according to a frequentist) unknown parameters should be treated as fixed unknown quantities, not as random variables. It's not* δ_1 *or* δ_2*, since they're given. It must be* $\hat{\sigma}$*. But in what sense is* $\hat{\sigma}$ *random? The only way to make sense of this probability expression is if we view* $\hat{\sigma}$ *as a function of a random dataset,*

$$\mathbb{P}\Big(\sigma \in [\hat{\sigma}(X) - \delta_1, \hat{\sigma}(X) + \delta_2]\Big)$$

where X *refers to a random sample of* n *values, drawn from the distribution specified in the question. This probability expression involves* σ *(which we don't know), and it involves* X *(whose distribution depends on* μ *and* σ*, which we don't know), so we can't compute it directly. You should instead use the bootstrap resampling procedure, page 39 of lecture notes.*

There are two ways to interpret (c).

- *It could mean "We've found a maximum likelihood estimator* $\hat{\sigma}(x)$*, that returns an estimate of* σ *given the data* x*; now find the spread of values we might expect to see if we re-ran the experiment and computed* $\hat{\sigma}$ *again" and you can answer this by resampling the dataset and looking at the* $\hat{\sigma}(X^*)$ *you get.*
- *Or it could mean "In part (b) we computed the confidence level of an interval for nature's true unknown* σ*; now tune the parameters* δ_1 *and* δ_2 *so as to give the answer 95%".*

I think the second interpretation is more natural here, since the question said "a 95% confidence interval for σ*" rather than "a 95% confidence interval for* $\hat{\sigma}$*". There is an obvious brute force answer. There is also a cunning algorithmic answer, for which you need to 'unwrap' the code you wrote for part (b) and ask "what value of the inputs would give me the output I want?" (This is a question you find yourself asking whenever you're debugging—what on earth could possibly have led to the answer it's showing me?)*

Question 5. I have a coin which might be biased. I toss it n times and get X heads where $X \sim \text{Binom}(n, \theta)$ and θ is unknown.

(a) Show that the maximum likelihood estimator for θ given X is $\hat{\theta} = X/n$.

(b) Find functions lo and hi such that $\mathbb{P}\Big(\hat{\theta} \geq \mathsf{lo}(\theta) \text{ and } \hat{\theta} \leq \mathsf{hi}(\theta)\Big) \approx 0.95$.

(c) Rearrange your answer to (b) to give an approximate 95% confidence interval for θ in terms of $\hat{\theta}$. *A pseudocode answer is easier than an algebraic answer.*

The hint in part (b) is the $\approx$ sign: we're looking for an approximate *95% confidence interval, so look at the rule of thumb for approximate confidence intervals on page 19 of lecture notes.*

For part (c), I really do want you to just rearrange your answer. Don't do any clever data science or probability, just rewrite the expression as $\theta \in [\ldots, \ldots]$.

The point of this question is that it gives you a confidence interval, without having to go via the resampling / bootstrap route. It's a useful trick for fast code.

Question 6. A common task in data processing is counting the number of unique items in a collection. When the collection is too large to hold in memory, we may wish to use fast approximation methods, such as the following: Given a collection of items $A_1, A_2, \ldots$, compute the hash of each item $X_1 = h(A_1)$, $X_2 = h(A_2)$, $\ldots$, then compute

$$T = \max_{1 \leq i \leq n} X_i.$$

If the hash function is well designed, then each X_i can be treated as uniformly distributed in $[0, 1]$, and unequal items will yield independent X_i.

(a) Show that $\mathbb{P}(T \leq t) = t^m$, where m is the number of unique items in the collection. Find the density function for T.

(b) Find the maximum likelihood estimator for m.

(c) Explain how to use the resampling method to find a confidence interval for m.

The keyword resampling *tells you this is a question about frequentist inference.*

There are two challenges in part (c). First challenge: what is it actually asking for? As in question 4, is it asking for a confidence interval for the maximum likelihood estimator $\hat{m}$, i.e. for numbers lo and hi such that

$$\mathbb{P}(\hat{m}(X) \in [\mathsf{lo}, \mathsf{hi}]) \approx 95\%?$$

Or is it asking for a confidence interval for the true unknown parameter m, i.e. for an output procedure i.e. an interval

$$\mathbb{P}(m \in [\mathsf{lo}(X), \mathsf{hi}(X)]) \approx 95\%?$$

(In this case, lo and hi have to be random variables, i.e. functions of the data X, since there needs to be something random for this probability expression to make sense.) I meant the latter.

Next, what functions $\mathsf{lo}(X)$ and $\mathsf{hi}(X)$ should you use? It's up to you to invent whatever interval you like. Have a look at the example page 37 for a suggestion.

From here on, it's a straightforward application of bootstrap resampling... except for the very subtle question of how to resample. Here's one way to think about resampling. The dataset consists of n values in $[0, 1]$, coming from m distinct Uniform$[0, 1]$ *random variables plus $n - m$ repeats. The unknown parameter in this statement is m. What is the parametric-resampling way to deal with unknown parameters?*

Question 7. I have built a text sentiment analyzer, and I hope to prove it is better than the state of the art analyzer. I ran them both on a validation set of documents, and obtained a collection of values $x_i \in \{-, 0, +\}$, $1 \leq i \leq n$, where $+$ means that mine did better, $-$ means that mine did worse, and 0 means that both did just as well.

(a) For the model $\Pr(-) = \Pr(+) = q/2$, $\Pr(0) = 1 - q$, find the maximum likelihood estimate for q.

(b) Let n_0 be the number of cases where $x_i = 0$, and similarly n_- and n_+. Consider the test statistic $t = n_+ + n_0/2$. Explain how to use resampling to find the distribution of t under the hypothesis that both analyzers are equally good. Give pseudocode for a hypothesis test.

(c) Let the alternative hypothesis be that my analyzer is better. Find a test statistic for comparing the two hypotheses, based on likelihood ratio. *Likelihood ratio is defined on page 44 of the notes.*

The keyword hypothesis test *tells you that this question is about pages 43–44 from lecture notes. The actual problem is taken from Machine Learning and Real World Data, where you were asked to conduct the so-called sign test, which uses exactly the test statistic from part (b).*

The general mechanism of a hypothesis test is always exactly the same, laid out on page 43 of lecture notes. There are three places that call for creativity:

- *What test statistic should I use? (In general this is entirely up to you. In this case, the question tells you.)*
- *How do I resample the data? In other words, assuming that H_0 is true, what would I expect to see if I re-ran the experiment? Part (a) is a hint, about how you might do parametric resampling.*
- *Should I do a one-sided or two-sided test? In other words, if H_0 isn't true, then what would I expect the test statistic to return—will it always be large positive, or can it be both large positive and large negative?*

Part (c) is about likelihood ratio tests, which are described in lecture notes but which weren't covered in lectures (and which are non-examinable). It's a good way to invent a test statistic, if you have no other ideas. Let $\Pr(-) = p_1$, $\Pr(+) = p_2$, $\Pr(0) = 1 - p_1 - p_2$, and let $\text{lik}(p_1, p_2 \mid data)$ be the likelihood function. The null hypothesis is that $p_1 = p_2$ and the alternative hypothesis is $p_1 < p_2$, so the likelihood ratio is

$$\frac{\max_q \text{lik}(q/2, q/2 \mid data)}{\max_{p_1,p_2\,:\,p_1<p_2} \text{lik}(p_1, p_2 \mid data)}.$$

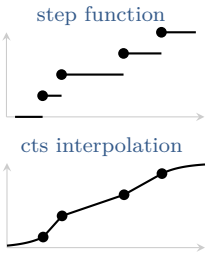
The calculus is a bit fiddly, but the eventual answer is intuitively sensible.

Compare this question to what you did (or will do) in IA/IB Machine Learning and Real World Data. (i) What test statistic should you use? It's up to you to invent whatever you like. It's not handed down to you from above. The test statistic "score 0.5 for each case where the two sentiment analyzers agree" isn't a kludge, it's an arbitrary choice, and any arbitrary choice is acceptable. (ii) How do you actually do the test? The method used in MLRD is a kludge, and now that you know resampling you know how to do the test properly.

Question 8. Suppose we have a dataset $x_1, \ldots, x_n$ and we want to fit a distribution to it, so that we can generate new values. One way to measure the goodness of fit is the perplexity score,

$$\log \text{perplexity} = -\mathbb{E} \log \Pr(X)$$

where X is a new value and Pr denotes the probability density for the fitted distribution. (Lower perplexity is better. If we wanted to choose between fitting a Normal distribution and fitting an Exponential distribution, for example, we'd choose whichever has lowest perplexity.) But we don't know the distribution of X — that's why we're trying to fit a distribution in the first place — so instead we can approximate the perplexity by



$$\log \text{perplexity} \approx -\frac{1}{n} \sum_{i=1}^{n} \log \Pr_{[-i]}(x_i)$$

where $\Pr_{[-i]}$ denotes the distribution fitted to the dataset with x_i omitted. *This is called leave-one-out cross validation, and is discussed in section 3.3.2 of lecture notes.*

Given the dataset $[3.1, 4.2, 7.8, 10.0]$, which version of the empirical distribution is better: the step function or the continuous interpolation?

This question is very sophisticated. It's research level, not undergraduate level! Nonetheless, the answer is very short, and it doesn't actually need any maths or any coding.

What's the probability density function $\Pr(x)$ for the step function distribution? It's 0 for almost all x, since the step function is flat almost everywhere. Technically, $\log \Pr(x)$ is undefined, but it's more useful to take the limit as $\Pr(x) \to 0$ and say $\log \Pr(x) = -\infty$ for almost all x. This is enough to let you decide which of the two versions of the empirical distribution function is better.

I think the answer is fascinating. What features of the data come into play in the answer? Can you invent a dataset where the step function is better, and another dataset where the continuous version is better?

Question 9. An engineer friend tells you "Bayesianism is the Apple of inference. You just work out the posterior, and everything Just Works™, and you don't need to worry about irritating things like confounded variables." What do you think? Illustrate your answer with reference to a random sample drawn from $X \sim \text{Normal}(\mu + \nu, \sigma^2)$ where μ and ν are unknown parameters. What is the maximum a posteriori estimate of (μ, ν) when the sample is large?

The maximum a posteriori (MAP) estimate is the parameter value that maximimizes the posterior density.

The keyword confounded variables *tells you that this question relates to page 54 of notes (only covered in the lecture on 2018-10-29). The keyword* Bayesian *says it's related to Bayesianism, page 33. This question requires you to make a novel connection between two separate parts of notes.*

The question suggests a probability model, $X \sim \text{Normal}(\mu + \nu, \sigma^2)$, where μ and ν are unknown parameters. Start by analyzing this model in the Bayesian manner: invent a prior distribution for the unknown parameters, write out the density for a sample $(X_1, X_2, \ldots, X_n)$ drawn from the specified distribution, then find the posterior distribution. Question 2 is useful here.

Next, think about confounded variables as we discussed them on page 54 of lecture notes. When variables are confounded, we can't learn their values by maximum likelihood estimation. But then—how come the Bayesian says "I know the posterior distribution of (μ, ν)" but the machine learner (maximum likelihood estimation) says "I can't estimate (μ, ν)"?

The question tells you to compute the Bayesian maximum a posteriori *estimates, and explains what this means. The answer is messy, but if you take the limit as $n \to \infty$ it simplifies into something that gives insight into the difference between Bayesian and MLE.*

Question 10. You have two coins from the same mint. You believe that the coins might be biased, and that they are likely to have similar bias, but you don't know what that bias might be. Invent a Bayesian prior distribution for (θ_1, θ_2) that expresses this belief, where θ_1 and θ_2 are the two bias parameters. Your distribution should have the property that any $(\theta_1, \theta_2) \in [0, 1]^2$ is possible, but that small $|\theta_1 - \theta_2|$ are more likely. *A good way to visualize your distribution is to generate samples and show a scatterplot, using low opacity for the points.*

There are many ways to answer this. Try to answer it using the trick from page 56 of lecture notes, the softmax/logit function, which maps $\mathbb{R} \to [0, 1]$. How could you program a random number generator to produce a pair $(X_1, X_2) \in \mathbb{R}^2$, such that any point in $\mathbb{R}^2$ is possible but $|X_1 - X_2| \approx 0$ is more likely?