Example sheet 2

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.

Question 1. I sample $X_1, \dots, X_n$ 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?

Question 2. I start with a prior belief that $\mu \sim \text{Normal}(\mu_0, \sigma_0)^2$. I then observe $x_1, \dots, x_n$, 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.*

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

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

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

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

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

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

where κ is

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

$1[A]$ 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.

Question 4. We are given a dataset $x_1, \dots, x_n$ 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.

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.*

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, ..., compute the hash of each item $X_1 = h(A_1)$, $X_2 = h(A_2)$, ..., then compute

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

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 $\mathbb{P}(T \le t) = t^m$, 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.

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 $x_i \in \{-, 0, +\}$, $1 \le i \le n$, 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.*

Question 8. Suppose we have a dataset $x_1, \dots, x_n$ 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,

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

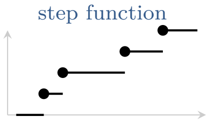
where X is a new value and Pr denotes the probability density for the fitted distribution. (Lower perplexity is better. If we wanted to choose between fitting a Normal distribution and fitting an Exponential distribution, for example, we'd choose whichever has lowest perplexity.) But we don't know the 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

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

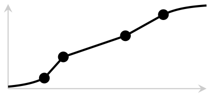
where $\Pr_{[-i]}$ denotes the distribution fitted to the dataset with x_i omitted. *This is called leave-one-out cross validation, and is discussed in section 3.3.2 of lecture notes.*

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?

step function



cts interpolation



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.

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.*