STAT 339: HOMEWORK 2B (LIKELIHOOD AND BAYESIAN INFERENCE)

UPDATED: DUE ON BLACKBOARD BY START OF CLASS ON MONDAY. MAR. 6

Instructions. Turn in your writeup and code to Blackboard as an archive file (e.g., `.zip`, `.tar`, `.gz`) by the start of class on Friday 3/3.

As always, you may use any language you like for the progamming components of this assignment — the tasks are stated in a language-neutral way. You may also use any typesetting software to prepare your writeup, but the final document should be a PDF. LaTeXis encouraged; a reproducible research format in which code is embedded into the document (e.g., knitr, RMarkdown, Jupyter or IPython Notebook) is even more encouraged.

1. **Bernoulli MLE.** (Adapted from FCML Ex. 2.9): Assume that a dataset of N binary values, $x_1, ..., x_n$, was sampled *i.i.d.* from a Bernoulli distribution with success parameter q.
   (a) Write out the likelihood function for q. Be clear about the domain!
   (b) Find a formula for the maximum likelihood estimator (MLE), $\hat{q}$ (Hint: the value of q that maximizes the log of the likelihood also maximizes the likelihood.)

2. **Univariate Normal (Gaussian) MLE.** (Adapted from FCML Ex. 2.8) Assume that a dataset of N real-valued observations was generated by a $\mathcal{N}(\mu, \sigma^2)$ distribution.
   (a) Write down the likelihood function for μ and σ^2 based on the full sample of all N observations.
   (b) Find the maximum likelihood estimates of the mean, μ, and variance, σ^2. (Hint 1: Remember that the product of exponentials is also the exponential

Date: Last Revised: February 27, 2017.

of a sum. Hint 2: The parameters that maximize the log likelihood also maximize the likelihood. Hint 3: You will need to differentiate the log likelihood separately with respect to μ and σ^2, and set both derivatives to zero simultaneously. You may need to find the MLE of one parameter first in terms of the other and then substitute.)

3. **MLE of Noise Variance in Linear Regression.** (FCML Ex. 2.11) Show that the MLE of the noise variance in the linear model, which is given in the textbook as

$$\hat{\sigma}^2 = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\hat{\mathbf{w}})$$

can also be written as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(t_n - \mathbf{x}_n^\mathsf{T}\mathbf{w})^2)$$

(Hint: start with the second expression and work backwards to get the first.)

4. **MLE for Linear Regression With Non-Constant Noise Variance.** Suppose we have a regression model

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

where $\varepsilon \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, and where $\boldsymbol{\Sigma}$ is a known diagonal matrix with $\boldsymbol{\Sigma}_{nn} = \sigma_n^2$. Show that the parameter vector $\mathbf{w}$ that maximizes the (log) likelihood also minimizes the weighted least squares loss

$$\mathcal{L}(\mathbf{w}; \mathbf{x}, \mathbf{t}) = \frac{1}{N}\sum_{n=1}^{N}\alpha_n(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2$$

for suitable choices of the α_n (and find the α_n).

5. **Bayesian inference for a proportion.** In order to determine how effective a magazine is at reaching its target audience, a market research company selects a random sample of n people from the target audience and interviews them. Let q represent the proportion of the target audience that has seen the latest issue and Y be the number in the interview group who has seen it.

   (a) Using a uniform prior on q, find the posterior distribution of q in terms of y and n.

(b) Find $\mathbb{E}[q|Y = y]$ in terms of y and n.

(c) Rewrite $\mathbb{E}[q|Y]$ as a weighted average of two terms: the MLE $\hat{q}$, and the prior mean, $\mathbb{E}[q]$. That is, find an expression for α so that $\mathbb{E}[q|Y = y] = \alpha\mathbb{E}[q] + (1 - \alpha)\hat{q}$.

(d) Show that the prior in 5a is a special case of a Beta distribution, and generalize the result in 5c for arbitrary choices of prior parameters.

(e) What does the expression for α suggest about the interpretation of the prior parameters (specifically, the way they affect posterior inferences about q)?

6. **Inferring a Detection Limit.** Suppose Y represents a non-negative measurement (i.e, $Y \geq 0$) that is that is detectable only up to a threshold θ. Suppose also that Y is uniformly distributed on its range, i.e., $Y \sim \mathsf{Unif}(0, \theta)$, but that the value of θ is unknown.

(a) Find a formula for the likelihood function for θ, given a single observation Y. Be careful to specify the domain of the function! Where is it largest? (You do not need calculus to answer this — draw the graph if you aren't sure.) Explain why, intuitively, it is largest there.

(b) Suppose we want to use Bayesian inference to estimate the upper bound, θ. The prior range of θ is $[0, \infty)$. Given an observation, $Y = y$, what is the posterior range for θ?

(c) Suppose θ has a Gamma prior: $\theta \sim \mathsf{Gamma}(a, b)$, with prior density

$$f(\theta) = \begin{cases} \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} & \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the non-constant component of the posterior PDF. Is the posterior also a Gamma distribution (for some values of a and b)? How do you know?

7. **A waiting time model.** The *exponential distribution* with *rate parameter* λ and range $[0, \infty)$ is often used to model the amount of time that passes between two events. Its density is given by

$$f(y|\lambda) = \lambda e^{-\lambda y}$$

(a) This density is a special case of a $\mathsf{Gamma}(a, b)$ density. Find the values of a and b that make the densities equivalent.

(b) Show that the $\mathsf{Gamma}(a, b)$ family is also a conjugate prior for the rate parameter. That is, if the prior density on λ has the form

$$f(\lambda) = k_{a,b} \cdot \lambda^{a-1} e^{-b\lambda}$$

where $a, b > 0$ are parameters and $k_{a,b} > 0$ is a normalizing constant that depends on a and b but not on λ, then the posterior density $f(\lambda|y)$ has the same form, for different values of a, b and k.

(c) Show that if $Y_1, \ldots, Y_n$ are i.i.d. exponential with rate λ, then using a $\mathsf{Gamma}(a, b)$ prior for λ results in a Gamma posterior density, $f(\lambda|y_1, \ldots, y_n)$, and find its parameters. (First write down the joint density for $Y_1, \ldots, Y_n$ / likelihood function for θ.)

8. **Simulating the Posterior for a Poisson parameter.** Consider the fish taco model you worked with on the last homework. We assume that the number of customers that buy fish tacos in a given hour can be modeled by a Poisson distribution with parameter λ, which has a distribution given by

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

(where $k!$ is k-factorial: $k! := k \times (k-1) \times \cdots \times 2 \times 1$) We now want to use data to infer λ.

(a) Suppose we have a dataset with customer counts for N different hours, represented by $\{Y_1, \ldots, Y_N\}$. Assuming the observations are independent and that λ is constant, find the likelihood function for λ.

(b) Find the MLE, $\hat{\lambda}$ (Hint: As always, the λ that maximizes the log likelihood also maximizes the likelihood).

(c) An alternative way to model arrival numbers is to consider the total number of customers in the entire N hour period as a single data point, generated from a Poisson distribution with mean N times as large as the single-hour distribution (i.e., $N\lambda$). Show that the likelihood function for λ in this scenario is a constant multiple of the likelihood function in the original scenario (and hence both MLE and Bayesian inferences about λ (for any prior) will be identical in both cases). Comment on what this means about the information contained in the order of the observations as it concerns inferences about λ in this model.

(d) The conjugate prior for the Poisson parameter is a Gamma distribution. Suppose the prior on λ is $\mathsf{Gamma}(a, b)$; that is

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

Find the posterior parameters in terms of the prior parameters and the data values, $\{y_1, \ldots, y_N\}$.

(e) In this case, since we are using a conjugate prior, we can find the posterior distribution analytically. But this will not be true for many other models. Hence we will often resort to approximation methods to do computations with the posterior distribution. An extremely naive (and inefficient) method is to take many samples from the joint distribution of the parameters and data (that is, generate a sequence of pairs $\{(\lambda_t^{(sim)}, y_t^{(sim)})\}$, $t = 1, \ldots, T$), and "condition" by retaining only those samples that yield data values identical to those observed (that is, where $y_t^{(sim)} = y_N$, where y_N is the real data). The posterior distribution is then approximated by the distribution of the parameter values used for the subset of the simulated data that is retained (that is, the set of $\lambda_t^{(sim)}$ such that $y_t^{(sim)} = y_N$). Implement this method using the formulation in 8c in which the data consists of a single count assumed to be drawn from a $\mathsf{Poisson}(N\lambda)$ distribution. Your function should take as input the prior parameters, a and b, the data value y_N, and a "stopping count", T_{kept}, which governs how many "retained" $(\lambda_t^{(sim)}, y_t^{(sim)})$ pairs must be generated before the algorithm stops. Run your algorithm for a few different combinations of y_N, a and b values, plotting both the theoretical posterior density and a histogram of the simulated posterior samples. Compare the theoretical and simulated means. Are they close?