Homework 4 Theory

Question 1

Assume X is a discrete random variable that takes values in $\{1,2,3\}$, with probability defined by

$$\begin{aligned} \Pr(X=1) &= \theta_1 \\ \Pr(X=2) &= 2\theta_1 \\ \Pr(X=3) &= \theta_2 \end{aligned}$$

where $\theta = [\theta_1, \theta_2]$ is an unknown parameter to be estimated.

Now assume we observe a sequence $D := \left\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right\}$ that is *independent and identically distributed (i.i.d.)* from the distribution. We assume the number of observations of the values: $1,2,3$ in D are s_1, s_2, s_3, respectively.

(a) To ensure that $\Pr(X=i)$ is a valid probability mass function, what constraint should we put on $\theta = [\theta_1, \theta_2]$? Write your answers quantitatively as expressions that include θ_1 and θ_2.

As these proabilities make up a complete pmf, we know that each probability is between 0 and 1 and that the sum of them is equal to 1.

$$\theta_1 + 2\theta_1 + \theta_2 = 1.$$

Which can be simplified to:

$$3\theta_1 + \theta_2 = 1.$$

As mentioned above the constraints are that these probabilities are the only probabilities that make up the pmf and that they are non-negative and their sum is exactly equal to 1.

(b) Write down the joint probability of the data sequence

$$\Pr(D \mid \theta) = \Pr\left(\left\{x^{(1)}, \ldots, x^{(n)}\right\} \mid \theta\right)$$

and the $\log$ probability $\log \Pr(D \mid \theta)$.

Given the probabilities above as well as the counts s_1, s_2, s_3 respectively, the joint probability is below:

$$\Pr(D \mid \theta) = \prod_{i=1}^{n} \Pr(X = x^{(i)} \mid \theta).$$

Since we know the count of observations for each respective probability, we can expand it as follows:

$$\Pr(D \mid \theta) = \theta_1^{s_1} \cdot (2\theta_1)^{s_2} \cdot \theta_2^{s_3}.$$

To find the log probability, we can first combine terms to separate off the θs and then apply the log of each side:

$$\log \Pr(D \mid \theta) = \log\left(\theta_1^{s_1+s_2} \cdot 2^{s_2} \cdot \theta_2^{s_3}\right).$$

Which simplifies to:

$$\log \Pr(D \mid \theta) = (s_1 + s_2) \log \theta_1 + s_2 \log 2 + s_3 \log \theta_2.$$

(c) Calculate the maximum likelihood estimation $\hat{\theta}$ of θ based on the sequence D.

First in order to reduce the number of varibales we will simplify the expression by using θ_2 in terms of θ_1 by using the fact that the probabilities sum to 1:

$$\theta_2 = 1 - 3\theta_1.$$

Substitute this into our previous expression in part b:

$$\log \Pr(D \mid \theta) = (s_1 + s_2) \log \theta_1 + s_2 \log 2 + s_3 \log(1 - 3\theta_1).$$

Now we will maximize this function with respect to θ_1 by taking the derivative and setting it equal to 0:

$$\frac{d}{d\theta_1}[(s_1 + s_2) \log \theta_1 + s_2 \log 2 + s_3 \log(1 - 3\theta_1)] = 0.$$

Which becomes:

$$\frac{s_1 + s_2}{\theta_1} - \frac{3s_3}{1 - 3\theta_1} = 0.$$

Solving for θ_1 we get:

$$\theta_1 = \frac{s_1 + s_2}{3(s_1 + s_2 + s_3)}.$$

Now we solve for θ_2 by using the substitution expression we used earlier to get:

$$\theta_2 = 1 - 3\left(\frac{s_1 + s_2}{3(s_1 + s_2 + s_3)}\right),$$

Which becomes:

$$\theta_2 = \frac{s_3}{s_1 + s_2 + s_3}.$$

Thus, the MLE is equal to θ_1 and θ_2 that are solved for above.

Question 2

Let $\{x^{(1)}, \ldots, x^{(n)}\}$ be an *i.i.d.* sample from an exponential distribution, whose the density function is defined as

$$f(x \mid \beta) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad \text{for} \quad 0 \leq x < \infty$$

Please find the maximum likelihood estimator (MLE) of the parameter β. Show your work.

To solve for this we will do the exact same process that we did above for θ but instead solve for the MLE of β.

First we define the likelihood function from the density function given above:

$$L(\beta \mid x^{(1)}, \ldots, x^{(n)}) = \prod_{i=1}^{n} \left(\frac{1}{\beta} \exp\left(-\frac{x^{(i)}}{\beta} \right) \right).$$

Now we apply the log to the function in order to find the log likliehood function:

$$\ell(\beta \mid x^{(1)}, \ldots, x^{(n)}) = \log\left(\left(\frac{1}{\beta} \right)^{n} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x^{(i)} \right) \right).$$

Which simplifies to:

$$\ell(\beta \mid x^{(1)}, \ldots, x^{(n)}) = -n \log \beta - \frac{1}{\beta} \sum_{i=1}^{n} x^{(i)}.$$

Now we find the derivative, by using the fact that if $f(x) = log(x)$ then $f'(x) = \frac{1}{x}$, and set it equal to 0:

$$\frac{\partial \ell(\beta \mid x^{(1)}, \ldots, x^{(n)})}{\partial \beta} = -n\frac{1}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x^{(i)} = 0$$

We will simplify the expression by multiply by β^2 and then solving for β:

$$-n\beta + \sum_{i=1}^{n} x^{(i)} = 0$$

$$\beta = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} = \bar{x}$$

So the MLE of β is simplified to equal the sample mean of the dataset, which makes sense given the density function provided.

Question 3

(a) Assume that you want to investigate the proportion (θ) of defective items manufactured at a production line. You take a random sample of 30 items and found 5 of them were defective. Assume the prior of θ is a uniform distribution on $[0, 1]$. Please compute the posterior of θ. It is sufficient to write down the posterior density function upto a normalization constant that does not depend on θ.

We are given:

- $N = 30$
- $D = 5$
- Prior distribution: $\theta \sim \mathrm{Uniform}(0, 1)$

The likelihood function for the data given θ is a binomial distribution as we know that it is a binary choice between defective or not defective samples:

$$\Pr(D \mid \theta) = \binom{N}{D} \theta^{D} (1 - \theta)^{N-D}.$$

The prior we will denote as $P(\theta_0)$. By Bayes' theorem, we know that the posterior is proportional to the product of the likelihood and the prior as seen below:

$$\pi(\theta \mid D) \propto \Pr(D \mid \theta) P(\theta_0)$$

Since we know that $P(\theta_0)$ is a uniform distribution, we know it is equal to 1. Therefore we have:

$$\pi(\theta \mid D) \propto \binom{30}{5} \theta^5 (1-\theta)^{25} \cdot 1.$$

We can remove the constant coefficient $\binom{30}{5}$ since it does not depend on θ, as stated in the problem.

$$\pi(\theta \mid D) \propto \theta^5 (1-\theta)^{25}.$$

*(b) **Assume an observation $D := \{x^{(1)}, \ldots, x^{(n)}\}$ is *i.i.d.* drawn from a Gaussian distribution $\mathcal{N}(\mu, 1)$, with an unknown mean μ and a variance of 1 . Assume the prior distribution of μ is $\mathcal{N}(0, 1)$. Please derive the posterior distribution $p(\mu \mid D)$ of μ given data D.**

We will use Bayes' theorem to combine elements of the likelihood function and the prior.

From the lectures, using a Gaussian distribution $\mathcal{N}(\mu, 1)$, we can define the likelihood function as:

$$p(D \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2} \right).$$

Therefore:

$$p(D \mid \mu) = (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(x^{(i)} - \mu)^2}{2} \right).$$

From the given problem statement, we can define the prior with $\mu \sim \mathcal{N}(0, 1)$ as:

$$p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2} \right).$$

Substituting the likelihood and prior being proportional to the posterior, we get:

$$p(\mu \mid D) \propto (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(x^{(i)} - \mu)^2}{2} \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2} \right).$$

Simplifying:

$$\propto \sum_{i=1}^{n} \frac{(x^{(i)} - \mu)^2}{2} + \frac{\mu^2}{2}$$

$$\propto \frac{1}{2} \left(\sum_{i=1}^{n} (x^{(i)} - \mu)^2 + \mu^2 \right).$$

Expanding $\sum_{i=1}^{n}(x^{(i)}-\mu)^2$:

$$\propto \frac{1}{2}\left(\sum_{i=1}^{n}(x^{(i)})^2 - 2\mu\sum_{i=1}^{n}x^{(i)} + n\mu^2 + \mu^2\right)$$

$$\propto \frac{1}{2}\left(\sum_{i=1}^{n}(x^{(i)})^2 - 2\mu\sum_{i=1}^{n}x^{(i)} + (n+1)\mu^2\right).$$

We can recognize the potential here for completing the square of the quadratic form which we will do by taking the exponent part of μ. First we will substitute by stating $S = \sum_{i=1}^{n}(x^{i})$ (doing it this to make it simpler in Latex!):

$$(n+1)\mu^2 - 2\mu S = (n+1)\left(\mu^2 - \frac{2\mu S}{n+1}\right)$$

$$= (n+1)\left(\mu - \frac{S}{n+1}\right)^2 - \frac{S^2}{n+1}.$$

So it becomes:

$$-\frac{1}{2}\left((n+1)\left(\mu - \frac{S}{n+1}\right)^2 - \frac{S^2}{n+1} + S^2\right).$$

From this quadratic we can therefore derive the mean which is the value that makes the squared term zero, where $S = \sum_{i=1}^{n}(x^{i})$:

$$\mu = \frac{S}{n+1}.$$

While the variance is the inverse of the coefficient of the quadratic term of μ:

$$\sigma^2 = \frac{1}{n+1}.$$

Lastly we have:

$$p(\mu \mid D) = \mathcal{N}\left(\frac{\sum_{i=1}^{n}x^{(i)}}{n+1}, \frac{1}{n+1}\right)$$