CS/DSC/AI 391L: Machine Learning

Homework 4 - Theory

Lecture: Prof. Qiang Liu

Problem 1

Consider a discrete random variable X with the following probability mass function:

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

Part a)

Question: What constraints must be placed on θ_1 and θ_2 to ensure that $\Pr(X=i)$ is a valid probability mass function?

Work: To ensure that $\Pr(X=i)$ is a valid probability mass function (PMF), the probabilities must be non-negative and sum to 1.

1. Non-negativity Constraints:

$$\begin{aligned}\Pr(X=1) &= \theta_1 \geq 0 \\ \Pr(X=2) &= 2\theta_1 \geq 0 \\ \Pr(X=3) &= \theta_2 \geq 0\end{aligned}$$

Since θ_1 and θ_2 are probabilities or multiples of probabilities, they must be non-negative:

$$\theta_1 \geq 0, \quad \theta_2 \geq 0$$

2. Sum-to-One Constraint:

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

Answer: The constraints are:

$$\theta_1 \geq 0, \quad \theta_2 \geq 0, \quad \text{and} \quad 3\theta_1 + \theta_2 = 1$$

Part b)

Question: Suppose we observe a data sequence $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$. Let s_1, s_2, and s_3 denote the number of times we observe 1, 2, and 3 in the sequence, respectively. Write down the joint probability of the data $\Pr(D \mid \theta)$ and its logarithm.

Work: The joint probability of the data is the product of individual probabilities for each observation:

$$\begin{aligned}\Pr(D \mid \theta) &= \prod_{i=1}^{n} \Pr(x^{(i)} \mid \theta) \\ &= (\theta_1)^{s_1} \cdot (2\theta_1)^{s_2} \cdot (\theta_2)^{s_3} \\ &= 2^{s_2} \theta_1^{s_1+s_2} \theta_2^{s_3}\end{aligned}$$

Taking the logarithm of the joint probability:

$$\begin{aligned}\log \Pr(D \mid \theta) &= \log(2^{s_2}\theta_1^{s_1+s_2}\theta_2^{s_3}) \\ &= s_2 \log 2 + (s_1+s_2)\log\theta_1 + s_3\log\theta_2\end{aligned}$$

Answer: Joint Probability:

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

Log Probability:

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

Part c)

Question: Find the maximum likelihood estimates for θ_1 and θ_2.

Work: To find the maximum likelihood estimates, we need to maximize the log probability with respect to θ_1 and θ_2, subject to the constraint $3\theta_1 + \theta_2 = 1$.

We can use the method of Lagrange multipliers:

$$L(\theta_1, \theta_2, \lambda) = (s_1+s_2)\log\theta_1 + s_3\log\theta_2 + \lambda(3\theta_1 + \theta_2 - 1)$$

Taking partial derivatives and setting them to zero:

$$\begin{aligned}\frac{\partial L}{\partial \theta_1} &= \frac{s_1+s_2}{\theta_1} + 3\lambda = 0 \\ \frac{\partial L}{\partial \theta_2} &= \frac{s_3}{\theta_2} + \lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= 3\theta_1 + \theta_2 - 1 = 0\end{aligned}$$

Solving these equations:

$$\begin{aligned}\theta_1 &= \frac{s_1+s_2}{-3\lambda} \\ \theta_2 &= \frac{s_3}{-\lambda}\end{aligned}$$

Substituting into the constraint equation:

$$3 \cdot \frac{s_1+s_2}{-3\lambda} + \frac{s_3}{-\lambda} = 1$$

Solving for λ:

$$\lambda = -\frac{s_1+s_2+s_3}{3} = -\frac{n}{3}$$

Substituting back:

$$\begin{aligned}\hat{\theta}_1 &= \frac{s_1+s_2}{3n} \\ \hat{\theta}_2 &= \frac{s_3}{n}\end{aligned}$$

Answer: The maximum likelihood estimates are:

$$\hat{\theta}_1 = \frac{s_1+s_2}{3n}, \quad \hat{\theta}_2 = \frac{s_3}{n}$$

Problem 2

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

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

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

Solution

To find the maximum likelihood estimator (MLE) of the parameter β for the exponential distribution, we start with the given i.i.d. sample $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$.

Step 1: Write the Likelihood Function

The joint likelihood function $L(\beta)$ is the product of the individual densities:

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

Simplifying:

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

Step 2: Compute the Log-Likelihood Function

Taking the natural logarithm:

$$\ell(\beta) = \ln L(\beta) = n \ln\left(\frac{1}{\beta}\right) - \frac{1}{\beta} \sum_{i=1}^{n} x^{(i)} = -n \ln \beta - \frac{S}{\beta},$$

where $S = \sum_{i=1}^{n} x^{(i)}$.

Step 3: Compute the First Derivative

Differentiating with respect to β:

$$\frac{d\ell}{d\beta} = -\frac{n}{\beta} + \frac{S}{\beta^2}.$$

Step 4: Find the Critical Point

Setting the derivative to zero:

$$-\frac{n}{\beta} + \frac{S}{\beta^2} = 0.$$

Multiplying by β^2:

$$-n\beta + S = 0.$$

Step 5: Solve for β

Solving for β:

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

Conclusion

The maximum likelihood estimator of β is the sample mean $\bar{x}$:

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

Answer: The MLE of β is the sample mean: $\hat{\beta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$.

Problem 3

Part a)

[10 points] 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.

Solution

We are interested in computing the posterior density $p(\theta \mid \text{Data})$, where θ is the proportion of defective items. Given:

- Prior distribution: $\theta \sim \text{Uniform}[0, 1]$, so $p(\theta) = 1$ for $\theta \in [0, 1]$.
- Data: In a sample of $n = 30$ items, $k = 5$ are defective.

The likelihood function for observing k defective items in n trials is given by the Binomial distribution:

$$p(\text{Data} \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Since we are interested in the posterior density up to a normalization constant that does not depend on θ, we can ignore constants not involving θ. Therefore, the unnormalized posterior density is:

$$p(\theta \mid \text{Data}) \propto p(\text{Data} \mid \theta) \cdot p(\theta) \propto \theta^k(1-\theta)^{n-k}$$

Substituting the given values $k = 5$ and $n = 30$:

$$p(\theta \mid \text{Data}) \propto \theta^5(1-\theta)^{25}$$

This is the posterior density function of θ up to a normalization constant independent of θ.

Part b)

[10 points] 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.

Solution

We aim to derive the posterior distribution $p(\mu \mid D)$ of the mean μ given the data $D = \{x^{(1)}, \dots, x^{(n)}\}$. Given:

- Data: $x^{(i)} \sim \mathcal{N}(\mu, 1)$, for $i = 1, \dots, n$.
- Prior distribution: $\mu \sim \mathcal{N}(0, 1)$.

Using Bayes' theorem:

$$p(\mu \mid D) \propto p(D \mid \mu) \cdot p(\mu)$$

The likelihood function is:

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

The prior distribution is:

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

Multiplying these together and taking the logarithm:

$$\begin{aligned}
\log p(\mu \mid D) &\propto -\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^2 - \frac{1}{2}\mu^2 + \text{constant} \\
&= -\frac{1}{2}\left(\sum_{i=1}^{n}(x^{(i)})^2 - 2\mu\sum_{i=1}^{n}x^{(i)} + n\mu^2 + \mu^2\right) + \text{constant} \\
&= -\frac{1}{2}\left(\sum_{i=1}^{n}(x^{(i)})^2 - 2\mu\sum_{i=1}^{n}x^{(i)} + (n+1)\mu^2\right) + \text{constant}
\end{aligned}$$

Completing the square for μ:

$$\begin{aligned}
&\sum_{i=1}^{n}(x^{(i)})^2 - 2\mu\sum_{i=1}^{n}x^{(i)} + (n+1)\mu^2 \\
&= (n+1)\left(\mu^2 - \frac{2\sum_{i=1}^{n}x^{(i)}}{n+1}\mu\right) + \sum_{i=1}^{n}(x^{(i)})^2 \\
&= (n+1)\left(\mu - \frac{\sum_{i=1}^{n}x^{(i)}}{n+1}\right)^2 + \text{constant}
\end{aligned}$$

Therefore, the posterior density is proportional to:

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

This is the kernel of a normal distribution with mean:

$$\mu_{\text{posterior}} = \frac{\sum_{i=1}^{n}x^{(i)}}{n+1} = \frac{n\bar{x}}{n+1}$$

and variance:

$$\sigma^2_{\text{posterior}} = \frac{1}{n+1}$$

Conclusion:
The posterior distribution of μ given the data D is:

$$\mu \mid D \sim \mathcal{N}\left(\frac{n\bar{x}}{n+1}, \frac{1}{n+1}\right)$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$ is the sample mean.