Group Assignment 2

Kirti Barman, Sourav Sharan, Vyoma Shah, Aditya Khadkikar
Group 15 - Introduction to Data Science

November 8, 2024

Exercise 1. Consider a supervised learning problem where we assume that $Y|X$ is Poisson distributed. That is, the conditional density of $Y|X$ is given by

$$f_{Y|X}(y,x) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad \lambda(x) = \exp(\alpha \cdot x + \beta).$$

Here α is a vector (slope) and β is a scalar (intercept). Follow the calculations from Section 4.2.1 to derive a loss function that needs to be minimized with respect to α and β. **Note**: Do we really need the factorial term?

Solution. To derive a loss function that can be minimized with respect to the parameters α (vector of slopes) and β (intercept), as mentioned in Section 4.2.1 we can setup the likelihood for the Poisson distribution and then take the log-likelihood.

Given the assumption that $Y \mid X$ is Poisson-distributed, the conditional probability mass function for Y given X is:

$$f_{Y|X}(y \mid x) = \frac{\lambda(x)^y e^{-\lambda(x)}}{y!}, \quad \lambda(x) = \exp(\alpha \cdot x + \beta).$$

Thus, for a single observation (x_i, y_i), the likelihood is:

$$f_{Y|X}(y_i \mid x_i) = \frac{\lambda(x_i)^{y_i} e^{-\lambda(x_i)}}{y_i!}.$$

For n independent observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the joint likelihood function is the product of the individual likelihoods:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{\lambda(x_i)^{y_i} e^{-\lambda(x_i)}}{y_i!}.$$

To simplify the computation, we take the log of the likelihood function, yielding the log-likelihood function:

$$\log L(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i \log \lambda(x_i) - \lambda(x_i) - \log y_i! \right).$$

Since $\lambda(x_i) = \exp(\alpha \cdot x_i + \beta)$, we substitute this into the log-likelihood:

$$\log L(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i(\alpha \cdot x_i + \beta) - \exp(\alpha \cdot x_i + \beta) - \log y_i! \right).$$

We now try to simplify the loss function. To derive a loss function to minimize with respect to α and β, we observe that the term $\log y_i!$ does not depend on α or β, so it can be ignored for optimization purposes. The remaining terms give us the negative log-likelihood loss:

$$L(\alpha, \beta) = -\log L(\alpha, \beta) = -\sum_{i=1}^{n} \left(y_i(\alpha \cdot x_i + \beta) - \exp(\alpha \cdot x_i + \beta) \right).$$

Thus, the loss function to minimize with respect to α and β is:

$$L(\alpha, \beta) = \sum_{i=1}^{n} \left(\exp(\alpha \cdot x_i + \beta) - y_i(\alpha \cdot x_i + \beta) \right).$$

Note on the Factorial Term

The factorial term $\log y_i!$ is constant with respect to α and β and thus does not influence the optimization of $L(\alpha, \beta)$. Therefore, we can omit it from the loss function.

The final form of the loss function to be minimized with respect to α and β is:

$$L(\alpha, \beta) = \sum_{i=1}^{n} \left(\exp(\alpha \cdot x_i + \beta) - y_i(\alpha \cdot x_i + \beta) \right).$$

This is the loss function used in Poisson regression.

$\square$

Exercise 2. Let $X_1, \ldots, X_n$ be IID random variables from Uniform$(0, \theta)$. Define $\hat{\theta} = \max(X_1, \ldots, X_n)$. First, find the distribution function of $\hat{\theta}$. Then compute the bias, standard error (SE), and mean squared error (MSE) of $\hat{\theta}$.

Solution. Given that $X_1, X_2, \ldots, X_n$ are IID random variables from a Uniform distribution on $[0, \theta]$, the distribution function of $\hat{\theta} = \max(X_1, \ldots, X_n)$ can be derived as follows. Since $X_i \sim \text{Uniform}(0, \theta)$, the cumulative distribution function (CDF) of each X_i is:

$$F_{X_i}(x) = \frac{x}{\theta}, \quad 0 \leq x \leq \theta.$$

To find the distribution of $\hat{\theta} = \max(X_1, \ldots, X_n)$, we can use the fact that the probability $P(\hat{\theta} \leq x)$ is the probability that all $X_i \leq x$ for $i = 1, \ldots, n$. So,

$$P(\hat{\theta} \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x).$$

Since the X_i are i.i.d., this becomes:

$$P(\hat{\theta} \leq x) = P(X_1 \leq x)^n = \left(\frac{x}{\theta}\right)^n, \quad 0 \leq x \leq \theta.$$

Thus, the CDF of $\hat{\theta}$ is:

$$F_{\hat{\theta}}(x) = \left(\frac{x}{\theta}\right)^n, \quad 0 \leq x \leq \theta.$$

Calculating the bias

The bias of an estimator $\hat{\theta}$ is defined as:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

To find $E[\hat{\theta}]$, we need the expectation of $\hat{\theta}$. We can find this using the probability density function (PDF) of $\hat{\theta}$.

The PDF $f_{\hat{\theta}}(x)$ can be obtained by differentiating the CDF:

$$f_{\hat{\theta}}(x) = \frac{d}{dx} F_{\hat{\theta}}(x) = \frac{d}{dx}\left(\frac{x}{\theta}\right)^n = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1}, \quad 0 \leq x \leq \theta.$$

Now, we can compute $E[\hat{\theta}]$:

$$E[\hat{\theta}] = \int_0^\theta x f_{\hat{\theta}}(x)\, dx = \int_0^\theta x \cdot \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1} dx.$$

Simplifying:

$$E[\hat{\theta}] = \frac{n}{\theta^n} \int_0^\theta x^n\, dx.$$

Integrating x^n:

$$E[\hat{\theta}] = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta.$$

Thus, the bias is:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}.$$

Standard Error (SE) of $\hat{\theta}$

The standard error of $\hat{\theta}$ is the square root of its variance:

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

To compute $\mathrm{Var}(\hat{\theta})$, we need $E[\hat{\theta}^2]$.

$$E[\hat{\theta}^2] = \int_0^\theta x^2 f_{\hat{\theta}}(x)\,dx = \int_0^\theta x^2 \cdot \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1} dx.$$

Simplifying:

$$E[\hat{\theta}^2] = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx = \frac{n}{\theta^n}\cdot\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2.$$

Then, the variance of $\hat{\theta}$ is:

$$\mathrm{Var}(\hat{\theta}) = E[\hat{\theta}^2] - (E[\hat{\theta}])^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2.$$

Simplifying again:

$$\mathrm{Var}(\hat{\theta}) = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right).$$

Therefore, the standard error is:

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

Mean Squared Error (MSE) of $\hat{\theta}$

The MSE of $\hat{\theta}$ is defined as:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

From the bias and variance we calculated:

$$\mathrm{MSE}(\hat{\theta}) = \theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) + \left(-\frac{\theta}{n+1}\right)^2.$$

By Simplifying:

$$\mathrm{MSE}(\hat{\theta}) = \theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} + \frac{1}{(n+1)^2}\right).$$

□

Exercise 3. Consider the continuous distribution with density

$$p(x) = \frac{1}{2}\cos(x), \quad -\frac{\pi}{2} < x < \frac{\pi}{2}.$$

(a) Find the distribution function $F(x)$.

(b) Find the inverse distribution function $F^{-1}(x)$.

(c) To sample using an Accept-Reject sampler (Algorithm 1), we need to find a density g such that $p(x) \leq Mg(x)$ for some $M > 0$. Find such a density g and the value of M.

Solution. (a) The distribution function $F(x)$ can be found using the integral of $p(x)$ from $-\pi/2 \to x$.

$$\int_{-\pi/2}^{x} p(x)\, dx = \int_{-\pi/2}^{x} \frac{1}{2} cos(x)\, dx = \frac{1}{2} sin(x)$$

$$= \frac{1}{2} sin(x) - \frac{1}{2} sin(-\pi/2)$$

$$F(x) = \frac{1}{2} sin(x) + \frac{1}{2}$$

domain?

(b) The inverse distribution function $F^{-1}(x)$ is found by setting the x variable to be in the position of the y-variable. Next, we solve for y, to find the inverse function.

$$x = \frac{1}{2} sin(y) + \frac{1}{2}$$

$$x - \frac{1}{2} = \frac{1}{2} sin(y)$$

$$2x - 1 = sin(y)$$

$$y = sin^{-1}(2x - 1) = F^{-1}(x)$$

domain?

(c) To find such a density for Accept-Reject based sampling, we take the maximal value of p(x), which is $\pi/2$. Next, keeping this value in mind, in a comparison function, we generate from a uniform distribution given by $Y \sim U(-\pi/2, \pi/2)$. This gives its probability density function as:

$$g = \frac{1}{b-a} = \frac{1}{\frac{\pi}{2} - (-\frac{\pi}{2})} = \frac{1}{\pi}$$

With the above proposal function, we find an M such that $\frac{1}{2} cos(x) \leq M * \frac{1}{\pi}$. The maximum value that $p(x) = \frac{1}{2} cos(x)$ can be is 1, at x = 0, hence this simplifies to:

$$\frac{1}{2} \leq \frac{M}{\pi} \to M \geq \pi/2.$$

□

Exercise 4. Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of IID discrete random variables where:

$$P(Y_i = 0) = 0.1, \quad P(Y_i = 1) = 0.3, \quad P(Y_i = 2) = 0.2, \quad P(Y_i = 3) = 0.4.$$

Define $X_n = \max\{Y_1, \ldots, Y_n\}$. Let $X_0 = 0$ and verify that $X_0, X_1, \ldots, X_n$ is a Markov chain. Find the transition matrix P.

Solution. A sequence of random variables $\{X_n\}$ is a Markov chain if the future state X_{n+1} depends only on the present state X_n and not on the past states, i.e.,

$$P(X_{n+1} = x \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x \mid X_n = x_n).$$

Here, each Y_i is an IID discrete random variable with possible values $0, 1, 2$, and 3, and we define

$$X_n = \max\{Y_1, Y_2, \ldots, Y_n\},$$

where $X_0 = 0$.

1. Verifying the Markov Property for X_n The process X_n represents the maximum value observed among the variables $Y_1, \ldots, Y_n$. Since Y_i can only take values $0, 1, 2$, or 3, the sequence X_n is restricted to these values as well.

   Given that $X_n = x$, the maximum of the first n values of Y_i is x. The next state X_{n+1} depends only on Y_{n+1} (since the maximum can only increase if Y_{n+1} exceeds the current maximum, x), so

   $$X_{n+1} = \max\{X_n, Y_{n+1}\}.$$

   This confirms that X_{n+1} depends only on X_n and Y_{n+1}, which is independent of previous values $Y_1, Y_2, \ldots, Y_n$. Thus, the sequence $\{X_n\}$ satisfies the Markov property.

2. Finding the Transition Matrix P The possible states of X_n are $0, 1, 2$, and 3. We need to find the probabilities of transitioning from one state to another, which will form the entries in the transition matrix P. Let's denote $P(i, j)$ as the probability of moving from state i to state j.

   (a) From State 0: - $P(0, 0)$: The probability that $X_{n+1} = 0$ given that $X_n = 0$ is simply $P(Y_{n+1} = 0) = 0.1$. - $P(0, 1)$: The probability that $X_{n+1} = 1$ given $X_n = 0$ is $P(Y_{n+1} = 1) = 0.3$. - $P(0, 2)$: The probability that $X_{n+1} = 2$ given $X_n = 0$ is $P(Y_{n+1} = 2) = 0.2$. - $P(0, 3)$: The probability that $X_{n+1} = 3$ given $X_n = 0$ is $P(Y_{n+1} = 3) = 0.4$.

   Thus, the row for $X_n = 0$ is:

   $$P(0, :) = [0.1, 0.3, 0.2, 0.4].$$

(b) From State 1: - $P(1,0) = 0$: Once X_n reaches 1, it cannot go back to 0. - $P(1,1)$: The probability that $X_{n+1} = 1$ given $X_n = 1$ is $P(Y_{n+1} \leq 1) = P(Y_{n+1} = 0) + P(Y_{n+1} = 1) = 0.1 + 0.3 = 0.4$. - $P(1,2)$: The probability that $X_{n+1} = 2$ given $X_n = 1$ is $P(Y_{n+1} = 2) = 0.2$. - $P(1,3)$: The probability that $X_{n+1} = 3$ given $X_n = 1$ is $P(Y_{n+1} = 3) = 0.4$.

Thus, the row for $X_n = 1$ is:

$$P(1,:) = [0, 0.4, 0.2, 0.4].$$

(c) From State 2: - $P(2,0) = 0$: Once X_n reaches 2, it cannot go back to 0. - $P(2,1) = 0$: Once X_n reaches 2, it cannot go down to 1. - $P(2,2)$: The probability that $X_{n+1} = 2$ given $X_n = 2$ is $P(Y_{n+1} \leq 2) = P(Y_{n+1} = 0) + P(Y_{n+1} = 1) + P(Y_{n+1} = 2) = 0.1 + 0.3 + 0.2 = 0.6$. - $P(2,3)$: The probability that $X_{n+1} = 3$ given $X_n = 2$ is $P(Y_{n+1} = 3) = 0.4$.

Thus, the row for $X_n = 2$ is:

$$P(2,:) = [0, 0, 0.6, 0.4].$$

(d) From State 3: - $P(3,0) = 0$: Once X_n reaches 3, it cannot go back to 0. - $P(3,1) = 0$: Once X_n reaches 3, it cannot go down to 1. - $P(3,2) = 0$: Once X_n reaches 3, it cannot go down to 2. - $P(3,3) = 1$: The probability that $X_{n+1} = 3$ given $X_n = 3$ is 1, since 3 is the maximum possible value of Y_i.

Thus, the row for $X_n = 3$ is:

$$P(3,:) = [0, 0, 0, 1].$$

(e) Final Transition Matrix P The transition matrix P is:

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0 & 0.4 & 0.2 & 0.4 \\ 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix represents the transition probabilities of the Markov chain $X_0, X_1, \ldots, X_n$.

$\square$

Exercise 5. Let $X_1, \ldots, X_n$ be IID from some distribution F that is unknown. Let $\hat{F}_n$ be the empirical distribution function. Use this to find an estimate of the quantile p of F. Use Theorem 5.28 (Dvoretzky-Kiefer-Wolfowitz inequality) to find a confidence interval for p.

Solution. We want to find a confidence interval for the quantile p of the unknown distribution F using the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.

To start, we define a confidence level $1 - \alpha$, where α is a small probability (0.05 for 95% confidence level). This confidence level represents the probability that the empirical distribution function $\hat{F}_n$ will closely approximate the true distribution function F across all values. According to the DKW inequality, we have:

$$P\left(\sup_x |\hat{F}_n(x) - F(x)| \leq \epsilon\right) \geq 1 - \alpha,$$

where ϵ represents the bound on the maximum deviation between $\hat{F}_n(x)$ and $F(x)$ over all x. We equate this confidence level to the DKW bound, giving:

$$1 - \alpha = 2e^{-2n\epsilon^2}.$$

Solving for ϵ, we find:

$$\epsilon = \sqrt{\frac{\ln(2/\alpha)}{2n}}.$$

This value of ϵ gives the error bound on the approximation of $F(x)$ by $\hat{F}_n(x)$ with confidence $1 - \alpha$. Now, let $\hat{x}_p$ denote the empirical quantile estimate, which is the value such that $\hat{F}_n(\hat{x}_p) = p$. Using the DKW inequality, we know that with probability at least $1 - \alpha$,

$$F(\hat{x}_p - \epsilon) \leq p \leq F(\hat{x}_p + \epsilon).$$

Thus, a $(1 - \alpha) \times 100\%$ confidence interval for the quantile p is:

$$\left[\hat{F}_n^{-1}(p - \epsilon),\ \hat{F}_n^{-1}(p + \epsilon)\right],$$

where $\epsilon = \sqrt{\frac{\ln(2/\alpha)}{2n}}$.

In summary, the confidence interval for the true quantile p is:

$$\left[\hat{F}_n^{-1}(p - \epsilon),\ \hat{F}_n^{-1}(p + \epsilon)\right],$$

with $\epsilon = \sqrt{\frac{\ln(2/\alpha)}{2n}}$.

$\hat{F}_n^{-1}(q)$ denotes the empirical quantile for probability q, which is the $\lceil q \cdot n \rceil$-th order statistic of the sample. This interval has a confidence level of $1 - \alpha$. $\square$