Introduction to Data Science 1MS041 - Assignment 2

Holger Swartling & Nir Teyar

November 4, 2024

Contents

1 Question 1

1.1 Question:

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 number (intercept).

Follow the calculations from Section 4.2.1 to derive a loss that needs to be minimized with respect to α and β.

Note: Do we really need the factorial term?

1.2 Answer:

Suppose we have n i.i.d. samples (X_i, Y_i) for $i = 1, \ldots, n$. The likelihood function for observing these samples is:

$$L(\alpha, \beta) = \sum_{i=1}^{n} f_{Y|X}(Y_i, X_i) = \sum_{i=1}^{n} \frac{\lambda(X_i)^{Y_i} e^{-\lambda(X_i)}}{Y_i!},$$

where $\lambda(X_i) = e^{\alpha \cdot X_i + \beta}$.

The log-likelihood function is:

$$\ln L(\alpha, \beta) = \sum_{i=1}^{n} \left(Y_i \ln \lambda(X_i) - \lambda(X_i) - \ln(Y_i!)\right).$$

Since $\lambda(X_i) = e^{\alpha \cdot X_i + \beta}$, we can substitute to get:

$$\ln L(\alpha, \beta) = \sum_{i=1}^{n} \left(Y_i(\alpha \cdot X_i + \beta) - e^{\alpha \cdot X_i + \beta} - \ln(Y_i!)\right).$$

The negative log-likelihood, which we aim to minimize, is:

$$-\ln L(\alpha, \beta) = \sum_{i=1}^{n} \left(-Y_i(\alpha \cdot X_i + \beta) + e^{\alpha \cdot X_i + \beta} + \ln(Y_i!)\right).$$

The term $\ln(Y_i!)$ does not depend on α or β, so it can be ignored when minimizing the negative log-likelihood. Therefore, the loss function to minimize with respect to α and β is:

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^{n} \left(-Y_i(\alpha \cdot X_i + \beta) + e^{\alpha \cdot X_i + \beta}\right).$$

2 Question 2:

2.1 Question

1. Let $X_1, \ldots, X_n$ be IID from Uniform$(0, \theta)$. Let $\hat{\theta} = \max(X_1, \ldots, X_n)$.

   First, find the distribution function of $\hat{\theta}$. Then compute the **bias**$(\hat{\theta})$, **se**$(\hat{\theta})$ and $\mathbf{MSE}_n(\hat{\theta})$.

2.2 Answer:

2.2.1 Part 1:

1. **Determine the Distribution of** $\hat{\theta}$: Since each $X_i \sim \text{Uniform}(0, \theta)$, the cumulative distribution function for any X_i is $F_{X_i}(x) = \frac{x}{\theta}$ for $0 \leq x \leq \theta$.

   For the maximum, $\hat{\theta} = \max(X_1, \ldots, X_n)$, we need to find $P(\hat{\theta} \leq x)$:

   $$P(\hat{\theta} \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = P(X_i \leq x)^n = \left(\frac{x}{\theta}\right)^n, \quad 0 \leq x \leq \theta.$$

   Thus, $\hat{\theta}$ has the cumulative distribution function:

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

   ### 2.2.2 Part 2:

2. **Compute the Expected Value** $E(\hat{\theta})$**:** The expected value of $\hat{\theta}$ can be calculated using its PDF, which we obtain by differentiating the cumulative distribution function:

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

   Now, calculate $E(\hat{\theta})$:

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

   ### 2.2.3 Part 3:

3. **Compute the Bias of** $\hat{\theta}$**:** Bias is defined as $E(\hat{\theta}) - \theta$.

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

2.2.4 Part 4:

Compute the Variance of $\hat{\theta}$:

The variance is given by $\text{Var}(\hat{\theta}) = E(\hat{\theta}^2) - \left(E(\hat{\theta})\right)^2$. We first need $E(\hat{\theta}^2)$:

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

Variance is now given by:

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

2.2.5 Part 5:

Compute the Mean Squared Error (MSE) of $\hat{\theta}$: The MSE is given by $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \left(\text{Bias}(\hat{\theta})\right)^2$.

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

Simplifying this expression gives:

$$\text{MSE}(\hat{\theta}) = \frac{\theta^2}{(n+1)^2}.$$

some calculation mistake

3 Question 3

3.1 Question:

Consider the continuous distribution with density

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

3.1.1 Question a

Find the distribution function $F(x)$.

3.1.2 Answer:

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

The distribution function $F(x)$ is given by the cumulative distribution function (CDF):

$$F(x) = \int_{-\frac{\pi}{2}}^{x} p(t)\,dt = \int_{-\frac{\pi}{2}}^{x} \frac{1}{2}\cos(t)\,dt.$$

Integrating, we have:

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

Since $\sin\left(\frac{-\pi}{2}\right) = -1$ which is a constant we make it a part of C

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

To determine C we use the boundary condition $F\left(-\frac{\pi}{2}\right) = 0$:

$$F\left(-\frac{\pi}{2}\right) = \frac{1}{2}\sin\left(-\frac{\pi}{2}\right) + C = 0 \implies \frac{1}{2}(-1) + C = 0 \implies -\frac{1}{2} + C = 0 \implies C = \frac{1}{2}.$$

Thus, the CDF is:

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

domain?

3.1.3 Question b

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

3.1.4 Answer:

To find $F^{-1}(u)$, we set $F(x) = u$ and solve for x:

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

Rearranging gives:

$$\sin(x) = 2u - 1$$

Taking the inverse sine, we have:

$$x = \arcsin(2u - 1)$$

Thus, the inverse distribution function is:

domain?

$$F^{-1}(u) = \arcsin(2u - 1)$$

3.1.5 Question 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 find the value of M.

3.1.6 Answer

The maximum value of $p(x)$ occurs when $x = 0$:

$$p(0) = \frac{1}{2}\cos(0) = \frac{1}{2}.$$

Thus, we have:

$$\max_{x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)} p(x) = \frac{1}{2}.$$

A natural choice for the density $g(x)$ which makes the calculations easy is the uniform distribution over $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, with density:

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

We need to ensure that:

$$p(x) \leq Mg(x) \implies \frac{1}{2}\cos(x) \leq M \cdot \frac{1}{\pi}.$$

The maximum of $g(x)$ is:

$$M \geq \frac{\pi}{2}.$$

The density $g(x)$ and constant M are given by:

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

4 Question 4

4.1 Question:

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

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

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

4.2 Answer:

4.2.1 Part 1: Verify the Markov Property

Let $X_n = \max(Y_1, \ldots, Y_n)$ represent the maximum value observed in the sequence up to the n-th draw.

To demonstrate that $X_0, X_1, \ldots, X_n$ forms a Markov chain, we need to show that the future state of the process depends only on the current state and not on any previous values.

Given $X_n = k$, the possible transitions depend solely on the next drawn value, Y_{n+1}, as follows:

- If $Y_{n+1} \leq k$, then $X_{n+1} = k$ (the "max number" does not change).
- If $Y_{n+1} > k$, then $X_{n+1} = Y_{n+1}$ (the "max number" updates to this higher value).

Since the future state X_{n+1} depends only on X_n and Y_{n+1}, this sequence satisfies the Markov property.

4.2.2 Part 2: Determine the Transition Matrix P

The states for X_n are $\{0, 1, 2, 3\}$, as these are the possible maximum values. To construct the transition matrix P, we need to calculate the probability of transitioning from each current "max number" $X_n = k$ to each possible next "max number" $X_{n+1} = j$.

Case 1: $X_n = 0$

If $X_n = 0$, then the possible transitions are:

- Stay at 0 if $Y_{n+1} = 0$: $P(X_{n+1} = 0 \mid X_n = 0) = P(Y_{n+1} = 0) = 0.1$.
- Move to 1 if $Y_{n+1} = 1$: $P(X_{n+1} = 1 \mid X_n = 0) = P(Y_{n+1} = 1) = 0.3$.
- Move to 2 if $Y_{n+1} = 2$: $P(X_{n+1} = 2 \mid X_n = 0) = P(Y_{n+1} = 2) = 0.2$.
- Move to 3 if $Y_{n+1} = 3$: $P(X_{n+1} = 3 \mid X_n = 0) = P(Y_{n+1} = 3) = 0.4$.

Case 2: $X_n = 1$

If $X_n = 1$, then the possible transitions are:

- Stay at 1 if $Y_{n+1} = 0$ or $Y_{n+1} = 1$: $P(X_{n+1} = 1 \mid X_n = 1) = P(Y_{n+1} = 0) + P(Y_{n+1} = 1) = 0.1 + 0.3 = 0.4$.
- Move to 2 if $Y_{n+1} = 2$: $P(X_{n+1} = 2 \mid X_n = 1) = P(Y_{n+1} = 2) = 0.2$.
- Move to 3 if $Y_{n+1} = 3$: $P(X_{n+1} = 3 \mid X_n = 1) = P(Y_{n+1} = 3) = 0.4$.

Case 3: $X_n = 2$

If $X_n = 2$, then the possible transitions are:

- Stay at 2 if $Y_{n+1} = 0$, $Y_{n+1} = 1$, or $Y_{n+1} = 2$: $P(X_{n+1} = 2 \mid X_n = 2) = P(Y_{n+1} = 0) + P(Y_{n+1} = 1) + P(Y_{n+1} = 2) = 0.1 + 0.3 + 0.2 = 0.6$.
- Move to 3 if $Y_{n+1} = 3$: $P(X_{n+1} = 3 \mid X_n = 2) = P(Y_{n+1} = 3) = 0.4$.

Case 4: $X_n = 3$

If $X_n = 3$, then the possible transitions are:

- Stay at 3 regardless of Y_{n+1}, since 3 is the highest possible value: $P(X_{n+1} = 3 \mid X_n = 3) = 1$.

4.2.3 Transition Matrix P

Based on the calculations above, the transition matrix P is:

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

Each entry P_{ij} in this matrix represents the probability of transitioning from state i to state j.

5 Question 5

5.1 Question:

The quantile p of a distribution F is the value x_p such that:

$$F(x_p) = p,$$

where F is the cumulative distribution function of X, the random variable from which $X_1, X_2, \ldots, X_n$ are IID samples.

x_p is the value below which p proportion of the data lies.

The empirical distribution function $\hat{F}_n$ is an approximation of F based on the sample data $X_1, \ldots, X_n$. It is defined as:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}},$$

where $\mathbf{1}_{\{X_i \leq x\}}$ is an indicator function that is 1 if $X_i \leq x$ and 0 otherwise. $\hat{F}_n(x)$ gives the proportion of sample points less than or equal to x, and it converges to $F(x)$ as $n \to \infty$ by the Law of Large Numbers.

5.2 Answer:

5.2.1 Part 1: Estimating the Quantile p

To estimate the quantile p, we find the value x such that $\hat{F}_n(x) \approx p$. This x serves as an empirical approximation of x_p (the p-th quantile of F).

Define $\hat{x}_p$ such that:

$$\hat{F}_n(\hat{x}_p) = p.$$

You can obtain $\hat{x}_p$ by sorting the sample $X_1, \ldots, X_n$ and finding the data point at the $\lceil np \rceil$-th position.

5.2.2 Part 2: Applying the DKW Inequality

The DKW inequality provides a probabilistic bound on the difference between $\hat{F}_n(x)$ and $F(x)$:

$$P\left(\sup_x \left|\hat{F}_n(x) - F(x)\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

With high probability, the empirical cumulative distribution function $\hat{F}_n(x)$ is close to the true cumulative distribution function $F(x)$.

To construct a confidence interval, we rearrange the inequality to bound the probability that $\hat{F}_n(x)$ differs from $F(x)$ by more than ϵ. For a given confidence level $1 - \alpha$, set:

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

Solving for ϵ, we get:

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

5.2.3 Part 3: Constructing the Confidence Interval for the Quantile p

Since $\hat{F}_n(\hat{x}_p) = p$, we can use the DKW bound to say that, with probability at least $1 - \alpha$:

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

This provides an interval around p for the empirical estimate $\hat{F}_n(\hat{x}_p)$.

In terms of x values, we can interpret this as a confidence interval for x_p based on the points where $\hat{F}_n(x)$ differs from p by no more than ϵ. Therefore, the confidence interval for the quantile p can be approximated as:

$$[\hat{x}_{p-\epsilon}, \hat{x}_{p+\epsilon}],$$

where $\hat{x}_{p-\epsilon}$ and $\hat{x}_{p+\epsilon}$ are the empirical values corresponding to the probabilities $p - \epsilon$ and $p + \epsilon$ in $\hat{F}_n$.

6 Contribution statement

The assignment as a whole was done by both group members. The group members did the assignment first individually and then checked answers with each other upon completion, thus resulting in approximately 50% contribution from each member.