IITM-CS5691 : Pattern Recognition and Machine Learning — Release Date: August 30, 2023
Assignment 1 — Due Date : September 14, 2023, 23:59

Roll No: CS23E001 — Name: Shuvrajeet Das
Collaborators (if any):
References/sources (if any):

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting pdf files (one per question) at Crowdmark by the due date. (Note: **No late submissions** will be allowed, other than one-day late submission with 10% penalty or four-day late submission with 30% penalty! Instructions to join Crowdmark and submit your solution to each question within Crowdmark **TBA** later).
- For the programming question, please submit your code (rollno.ipynb file and rollno.py file in rollno.zip) directly in moodle, but provide your results/answers (including Jupyter notebook **with output**) in the pdf file you upload to Crowdmark.
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material or LLMs (Large Language Models like ChatGPT) for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words (this also means that you cannot copy-paste the solution from LLMs!). Please be advised that *the lesser your reliance on online materials or LLMs* for answering the questions, *the more your understanding* of the concepts will be and *the more prepared you will be for the course exams*.
- Points will be awarded based on how clear, concise and rigorous your solutions are, and how correct your answer is. The weightage of this assignment is 12% towards the overall course grade.

1. (8 points) [EXPLORING MAXIMUM LIKELIHOOD ESTIMATION]
Consider the i.i.d data $\mathbf{X} = \{x_i\}_{i=1}^{n}$, such that each $x_i \sim \mathcal{N}(\mu, \sigma^2)$. We have seen ML estimates of $\mu, \sigma^2$ in class by setting the gradient to zero.

(a) (4 points) How can you argue that the stationary points so obtained are indeed global maxima of the likelihood function?

(b) (4 points) Derive the bias of the MLE of $\mu, \sigma^2$.

**Solution:** The solution of question (a)

The normal distribution can be written as $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Now, we can describe, $\mathcal{L}(X)$ as

$$\begin{aligned}
\mathcal{L}(X) &= \prod_{i=1}^{n} f(x_i, \mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)
\end{aligned}$$

Now for log-likelihood $l(x)$ to be defined by $\log(\mathcal{L}(x))$

$$\begin{aligned}
l(x) &= \log(\mathcal{L}(x)) \\
&= n\log(\sigma) - n\log(\sqrt{2\pi}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2
\end{aligned}$$

Differentiating with $\mu$ and $\sigma$,

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2(-1) = 0 \\
\implies & \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0 \\
\implies & \sum_{i=1}^{n} x_i - n\mu = 0 \\
\implies & \mu = \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma} &= \frac{n}{\sigma} - (-2)\frac{1}{2\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \\
\implies & -n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0 \\
\implies & \sum_{i=1}^{n}(x_i - \mu)^2 = n\sigma^2 \\
\implies & \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2
\end{aligned}$$

Thus the problem can be optimized as a minimization problem as least square forming a convex function to get a global minima

The solution of question (b)

Mathematically, the bias (B) of an estimator $\hat{\theta}$ estimating a parameter $\theta$ is defined as:

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

To derive the bias of the Maximum Likelihood Estimators (MLE) for $\mu$ and $\sigma^2$, we need to calculate the expected values of the MLEs and compare them to the true values of $\mu$ and $\sigma^2$.

Let's start with the MLE for $\mu$, denoted as $\hat{\mu}$. The bias, $B(\hat{\mu})$, is defined as:

$$B(\hat{\mu}) = E[\hat{\mu}] - \mu$$

To calculate the expected value of the MLE for $\mu$, we consider that $\hat{\mu}$ follows a normal distribution with mean $\mu$ and variance $\frac{\sigma^2}{n}$ (this is a known property of the MLE for the mean of a normal distribution):

$$E[\hat{\mu}] = \mu$$

Therefore, the bias of the MLE for $\mu$ is:

$$B(\hat{\mu}) = \mu - \mu = 0$$

Now, let's derive the bias for the MLE of $\sigma^2$, denoted as $\hat{\sigma}^2$. The bias, $B(\hat{\sigma}^2)$, is defined as:

$$B(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2$$

The MLE for $\sigma^2$ is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

To calculate the expected value of $\hat{\sigma}^2$, we can use the properties of sample variances for a normal distribution:

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

Therefore, the bias of the MLE for $\sigma^2$ is:

$$B(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{n-1}{n}\sigma^2 - \frac{n}{n}\sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = \left(\frac{-1}{n}\right)\sigma^2$$

So, the bias of the MLE for $\sigma^2$ is $-\frac{\sigma^2}{n}$.