GRAD C23: Mathematics for Data Science
Hertie School
Fall 2023
Prof. Magazinnik

Problem Set 2

Submitted by: *Giorgio Coppola*
Group: *Shruti Pradeep Kakade, Giorgio Coppola, Monserrat Lopez Perez*

1. **Maximum Likelihood**.

   Synthesized facts:

   - Bangladesh, with 163 million residents, is the densest delta globally, with 25% of its land just seven feet above sea level.
   - The frequency of major floods in its low-lying areas appears to be rising.
   - A question arises: Is this due to climate change?
   - Infrequent climatic occurrences are usually modelled using the Poisson distribution: Xt (number of floods in time t) follows $X_t \sim$ Poisson(λ).

   a. Intuitive understanding: if we assume major floods in Bangladesh occur at an average rate of λ per 5-year period, how plausible is it that we would observe the series [1, 3, 1, 2, 0] over five consecutive periods?

   The following series of events have been recorded (number of floods - period):

$$\begin{bmatrix} 1 & 2000-2004 \\ 3 & 2005-2009 \\ 1 & 2010-2014 \\ 2 & 2015-2019 \\ 0 & 2020-2024 \end{bmatrix}$$

   The Poisson distribution models the number of events (in this case, major floods) occurring in a fixed interval of time or space. Given the observed flood data, we can use the Poisson distribution to determine the likelihood of observing such data for a given flood rate λ.

   To determine the most probable value for λ, we employ Maximum Likelihood Estimation (MLE). MLE aims to find the parameter value (in this case, λ that makes the observed data most probable for a specified model (here, the Poisson distribution). However, we are not yet performing any optimization or maximization; it's simply about understanding the data's fit to the model. Indeed, this is the "likelihood". To derive the likelihood, we multiply the individual probabilities of observing k floods given an average rate λ for each 5-years period of time in the Poisson distribution.

   Poisson distribution:

$$P(X = k; \lambda) = \frac{e^{-\lambda} \cdot \lambda^k}{k!},$$

   where:

- $P(X = k; \lambda)$ is the probability of observing k events.
- λ is the average event rate.
- e is Euler's number (approximately equal to 2.71828).
- k is the number of occurrences.

From our collected data, we know the following events occurred:

$L(\lambda; 1, 3, 1, 2, 0) = P(X = 1; \lambda) \times P(X = 3; \lambda) \times P(X = 1; \lambda) \times P(X = 2; \lambda) \times P(X = 0; \lambda)$

Since the observations are independent, the overall likelihood of observing the entire series of flood events is the product of the individual probabilities. Substituting the expression for the probability mass function of the Poisson distribution:

$$\begin{aligned} L(\lambda; 1, 3, 1, 2, 0) &= \frac{e^{-\lambda} \cdot \lambda^1}{1!} \times \frac{e^{-\lambda} \cdot \lambda^3}{3!} \times \frac{e^{-\lambda} \cdot \lambda^1}{1!} \times \frac{e^{-\lambda} \cdot \lambda^2}{2!} \times \frac{e^{-\lambda} \cdot \lambda^0}{0!} \\ &= \frac{\lambda^7 \cdot e^{-5\lambda}}{12}. \end{aligned}$$

This likelihood function represents how probable our observed data is for different values of λ. The value of λ that maximizes this function would be the Maximum Likelihood Estimate (MLE) for λ.

b. Since the logarithm turns products into sums (which are easier to differentiate), log-transformations are frequently used to facilitate various types of estimation, including MLE.

Given the original likelihood function:

$$L(\lambda; 1, 3, 1, 2, 0) = \frac{\lambda^7 \cdot e^{-5\lambda}}{12},$$

the log-likelihood is its natural logarithm:

$$\ell(\lambda) = \ln(L(\lambda; 1, 3, 1, 2, 0)) = \ln \frac{\lambda^7 \cdot e^{-5\lambda}}{12}.$$

When applying the logarithm to the likelihood function, we utilize properties of logarithms. In this case we use $\ln(ab) = \ln(a) + \ln(b)$ and $\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$, obtaining:

$$= \ln(\lambda^7) + \ln(e^{-5\lambda}) - \ln(12),$$

and then we simplify, applying $\ln(a^b) = b \cdot \ln(a)$ and $\ln(e^c) = c$, obtaining:

$$= 7\ln(\lambda) - 5\lambda - \ln(12).$$

This is our final log likelihood.

c. Maximizing the likelihood means to answer to the following question: which is the parameter (λ in this case) that maximizes the likelihood to observe this particular set of data?
To find the value of λ that maximizes this function (the Maximum Likelihood Estimate), we would differentiate with respect to λ and set the result to zero.

From previous calculation, we know that the log likelihood function is:

$$\ell(\lambda) = 7\ln(\lambda) - 5\lambda - \ln(12).$$

Differentiating with respect to λ:

$$\frac{d\ell(\lambda)}{d\lambda} = 7 \times \frac{1}{\lambda} - 5 = \frac{7}{\lambda} - 5$$

Setting this derivative to zero to find the maximum:

$$\frac{7}{\lambda} - 5 = 0 \implies \lambda = \frac{7}{5}.$$

d. This implies that the Maximum Likelihood Estimate (MLE) for λ for our Poisson distribution model is $\frac{7}{5}$ or 1.4. We can interpret this value as the most probable average rate of major floods in Bangladesh over a 5-year period, given our model and the observed data. If we wanted to make a prediction based on the given data, we would expect, on average, about 1.4 major floods to occur in Bangladesh every 5 years. This number is derived from the actual observed flood events in the provided data intervals, and therefore it is just a descriptive "best fitting" parameter of the data given our assumed distribution.

e. Assuming 18 major flooding events over the past century, namely over 100 years, the average of floods is:

$$\text{Rate}_{20th} = \frac{18 \text{ floods}}{100 \text{ years}} = 0.18 \text{ floods/year}.$$

A rate of 0.18 floods per year means exactly 0.9 floods over a period of 5 years:

$$\text{Rate}_{20th} \times 5 = 0.9 \text{ floods/5 years}.$$

In contrast, in our previous analysis, the MLE for λ indicated an average of $\hat{\lambda} = 1.4$ floods/5 years or $\hat{\lambda}/5 = 0.28$ floods/year.

The rate of major floods in the first quarter of the 21st century is higher than that of the 20th century. Specifically, there is an increase of 0.5 floods every 5 years or 0.1 floods per year, which is substantially quite significant. This could suggest that potential environmental changes, possibly due to factors like climate change, have influenced the frequency of these events. However, this analysis alone cannot conclusively link the increase in flood frequency to climate change, as it does not say anything about the specific correlation between climate change and extreme whether events in the context of Bangladesh, and least of all it can, without further analysis, establish any causal relationship.

2. **Taylor Series Approximation**

a. Given the PDF for the standard normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The CDF represents the probability that the random variable X takes on a value that is less than or equal to x. The CDF is the result of the integration of a PDF. Indeed, the PDF always integrates to 1 over its entire domain. The integral of a PDF from $-\infty$ to x represents the probability that the random variable X takes on a value less than or equal to x, namely its CDF.

Therefore, the CDF of $F(x)$, is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt.$$

At this point we should take out the constant and apply the partial integration rule. We would quickly notice that we can keep doing so and we will not arrive to a simple closed-form solution in terms of elementary functions.

b. In the previous answer, we were trying to integrate e^{-t^2}, but this term does not have an elementary antiderivative in the real numbers. The Fundamental Theorem of Calculus ensures that if a function is integrable, its antiderivative exists and can be expressed as a definite integral. Remember that the antiderivative of a function $f(x)$ is another function $F(x)$ such that its derivative $F'(x)$ is the initial function $f(x)$. Basically, to find the antiderivative means to answer to the question: "what function has this rate of change?". In this case, a function that has such rate of change exists, but finding this function is not straightforward, as it is not an elementary function. We can still calculate it, but the solution will not be in a simple closed-form. That is why, in practice, values for the CDF of the standard normal distribution are often looked up in a "z-table" or computed using software.

c. To find a second-order Taylor series approximation for the PDF at $x = 1$, we'll first need to determine the first and second derivatives of our function $f(x)$.

Therefore, given the original PDF:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

to find the first derivative, $f'(x)$, differentiate with respect to x using the chain rule:

$$f'(x) = \frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right)$$

Since $\frac{1}{\sqrt{2\pi}}$ is a constant, our interest should be focused on $e^{-\frac{x^2}{2}}$: it is a natural exponential elevated to a function, therefore we have to apply the chain rule.

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi}} \times \frac{-2x}{2} \times e^{-\frac{x^2}{2}} \\ &= -\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \end{aligned}$$

To find the second derivative, $f''(x)$, differentiate $f'(x)$ with respect to x using the product rule:

$$f''(x) = \frac{d}{dx}\left(-\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right)$$

Which is a case of $f''(x) = f(x)g(x) = \frac{d}{dx}f(x)g(x) + f(x)\frac{d}{dx}g(x)$:

$$\begin{aligned} f''(x) &= \frac{d}{dx}\left(-\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) \\ &= \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} + \frac{-x}{\sqrt{2\pi}}(-x)(e^{-\frac{x^2}{2}}) \\ &= -\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \\ &= \frac{x^2 e^{-\frac{x^2}{2}} - e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \\ &= \frac{e^{-\frac{x^2}{2}}(x^2-1)}{\sqrt{2\pi}}. \end{aligned}$$

d. Let's evaluate the function and its two first derivatives at point $x = 1$.
Evaluating $f(1)$:

$$f(1) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \approx 0.242$$

Evaluating $f'(1)$:

$$f'(1) = -\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = -\frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \approx -0.242$$

Evaluating $f''(1)$:

$$f''(1) = \frac{e^{-\frac{x^2}{2}}(x^2-1)}{\sqrt{2\pi}} == \frac{e^{-\frac{1}{2}}(1-1)}{\sqrt{2\pi}} = 0.$$

Notice: $f''(1) = 0$, so we can predict that the function has a point of inflection at $x = 1$, as it neither concaves up nor down at this point. Indeed, the second derivative of a function provides information about its concavity. Specifically, if $f''(x) = 0$, the function might switch concavity or stay linear.

e. A Taylor Series produces an approximation of a function at a certain value of x. The higher is the order of the series, the more precise the approximation is. To obtain the second-order Taylor polynomial for $f(x)$ at $x = 1$, we use the Taylor series formula:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

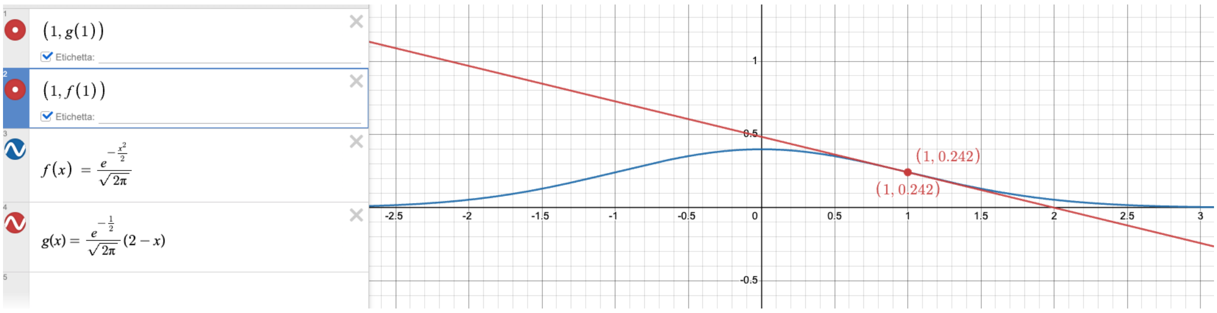
In our case, $a = 1$. Substituting in the values we found:

$$\begin{aligned} f(x) &\approx f(1) + f'(1)(x-1) + \frac{1}{2}f''(1)(x-1)^2 \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}} - \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}(x-1) + 0 \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}(1-(x-1)) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}(2-x) \\ &= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}(2-x) \end{aligned}$$

Therefore, $g(x) = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}(2-x)$ provides an approximation to the standard normal PDF near $x = 1$. We call $g(x)$ the approximated PDF so that we do not get confused with the previous mathematical manipulations.

f. Below, you find the graph:

Figure 1: The graph shows the plot of the original PDF of the Normal Distribution in green, the second-order Taylor series approximation of the same function at x = 1 in red, the point x = 1, which appears also to be the point of inflection of the original function.



1 $(1, g(1))$

☑ Etichetta:

2 $(1, f(1))$

☑ Etichetta:

3 $f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$

4 $g(x) = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}(2-x)$

5

Notice: the two values at $x = 1$ is exactly the same, showing that this was a very precise approximation.

g. As said earlier, the CDF is given by the integral of the PDF. Since we cannot take the integral of the original PDF, as it has no closed-form solution, we integrate its second-order Taylor series approximation to get an approximation of the CDF at some value x.

For our approximated PDF:

$$g(x) = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}(2 - x).$$

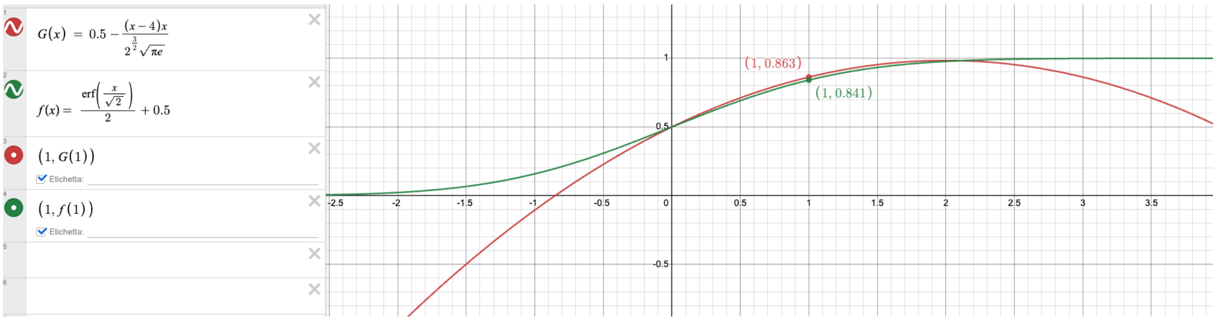
Integrating this PDF approximation function $g(x)$ from $-\infty$ to x will approximate the value of the CDF of X $G(x)$ around $x = 1$.

$$\begin{aligned} G(x) &= \int_{-\infty}^{x} g(t)\,dt \\ &= \int_{-\infty}^{x} \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}(2 - t)\,dt \\ &= -\frac{(t-4)\,t}{2^{\frac{3}{2}}\sqrt{e\pi}} \end{aligned}$$

Notice: the function $G(x)$ is the integral of the second-order Taylor Series approximation $g(x)$. It is a function, giving as output the integral value of whatever $x = n$ we input. The output value will be a good approximation of the actual CDF if $x = 1$ or is around 1. We can check from the plot showed to answer question h.

h. Below, you find the graph:

Figure 2: This figure shows the original CDF of the Normal Distribution in green, and the second-order Taylor approximation of the original function in red at x = 1.



1 $G(x) = 0.5 - \frac{(x-4)x}{2^{\frac{3}{2}}\sqrt{\pi e}}$

2 $f(x) = \frac{\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)}{2} + 0.5$

3 $(1, G(1))$

☑ Etichetta:

4 $(1, f(1))$

☑ Etichetta:

5

6

Notice: the two values at $x = 1$ are almost identical. This shows that our approximation is very close to the real value.