MAD 2020-21, Assignment 4

Bulat Ibragimov, Kim Steenstrup Pedersen

hand in until: 20.12.2021 at 23:59

General comments: The assignments in MAD must be completed and written individually. You are allowed (and encouraged) to discuss the exercises in small groups. If you do so, you are required to list your group partners in the submission. The report must be written completely by yourself. In order to pass the assignment, you will need to get at least 40% of the available points. The data needed for the assignment can be found in the assignment folder that you download from Absalon.

Submission instructions: Submit your report as a PDF, not zipped up with the rest. Please add your source code to the submission, both as executable files and as part of your report in appendix. To include it in your report, you can use the `lstlisting` environment in LaTeX, or you can include a "print to pdf" output in your pdf report. In some exercises we will ask you to include a code snippet as part of your solution text - a code snippet is only the most essential lines of code needed for solving the problem, this does not include import statements, other forms of boiler plate code, as well as plotting code.

Maximum Likelihood

Exercise 1 (2 points. (from 27.11.) Maximum Likelihood). Let the random variables $X_1, \ldots, X_n$ be i.i.d. following the geometric distribution Geo(θ) with PMF $p_\theta(x) = (1-\theta)^{x-1}\theta$ and that the values $x \in \mathbb{Z}_+$ (set of positive integers excluding zero), which then also holds for each of $X_1, \ldots, X_n$. Prove that

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^n x_i}$$

is the maximum likelihood estimator (MLE) for the parameter θ. (Hint: Sometimes it is easier to maximize the logarithm of a function instead of the function itself.)

Deliverables. Include the steps of your proof of the maximum likelihood estimator for the parameter θ, including showing that the estimator is indeed a maximum of the likelihood function.

Solution:

A geometric distribution has the pmf $p_\theta(x) = (1-\theta)^{x-1}\theta$. Since $X_1, X_2, ..., X_n$ are independent: $p_\theta^{\text{joint}}(x_1, ..., x_n) = \prod_{i=1}^n p_\theta(x_i)$. We maximize the logarithm of this function. This does not change the location of the maximum parameter but makes differentiation easier.

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log p_\theta^{\text{joint}}(x_1, x_2, ..., x_n) &= \frac{\partial}{\partial\theta}\log\left(\prod_{i=1}^n p_\theta(x_i)\right) \\
&= \frac{\partial}{\partial\theta}\sum_{i=1}^n \log(p_\theta(x_i)) = \frac{\partial}{\partial\theta}\sum_{i=1}^n [(x_i - 1)\log(1-\theta) + \log(\theta)] \\
&= \frac{\partial}{\partial\theta}[n\log(\theta) - n\log(1-\theta) + \sum_{i=1}^n (x_i \log(1-\theta))] \\
&= \frac{n}{\theta} + \frac{n}{1-\theta} - \frac{\sum_{i=1}^n x_i}{1-\theta} \qquad (1) \\
&= \frac{n - \theta\sum_{i=1}^n x_i}{\theta(1-\theta)} = 0 \\
\Rightarrow \quad \hat{\theta}_n &= \frac{n}{\sum_{i=1}^n x_i}
\end{aligned}$$

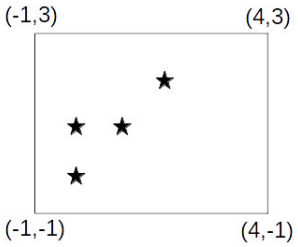
To make sure this is a maximum we compute the second derivative, by starting from (1).

$$\frac{\partial^2}{\partial\theta^2}\log p_\theta(x_1, x_2, ..., x_n) = -\frac{n}{\theta^2} + \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2} < 0$$

I know this is less than zero because: $x_i \in [1, \infty)$, meaning that $\sum_{i=1}^{n} x_i \geq n$. **Instructions**:
0.5 point for writing up the MLE problem, 1 point for solving for the parameter, and 0.5 point for checking that it is a maximum.

Exercise 2 (**4 points. 4-dimensional Maximum Likelihood).** During night, a prisoner sits in a completely dark room and faces a dark wall. He knows that the wall has a window, but he neither knows the window's size nor its exact position. The only thing that he sees are four stars, i.e., light points, at positions $(0, 0)$, $(0, 1)$, $(1, 1)$ and $(2, 2)$, which obviously must be within the window. He then wants to infer the boundary of the window by maximum likelihood. Assume that the points $(X_1, Y_1), \ldots, (X_4, Y_4)$ representing stars visible in the window are independent and identically distributed (i.i.d.) by a uniform distribution with the parameters $\theta := (x_{\min}, x_{\max}, y_{\min}, y_{\max})$. That is, the probability density function (PDF) for distribution of stars seen through the window is

$$f_\theta(x, y) = \begin{cases} c & \text{if } x_{\min} \leq x \leq x_{\max} \text{ and } y_{\min} \leq y \leq y_{\max} \\ 0 & \text{otherwise.} \end{cases}$$



a) Find the correct value for $c \in \mathbb{R}$ (Hint: Recall the defining properties of a probability density function).
b) Compute the likelihood for the two sets of parameter values $\theta_1 = (-1, 4, -1, 3)$ (see illustration to the right) and $\theta_2 = (-2, 5, -3, 6)$.
c) Find the maximum likelihood estimator (MLE) for the parameter set, $\hat{\theta}^{\text{ML}} = (\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max})$ given the observed stars at positions $(0, 0)$, $(0, 1)$, $(1, 1)$ and $(2, 2)$. (Hint: Consider for which cases the likelihood > 0, and use the fact that the PDF is uniform).

Deliverables. a) Include the derivation steps and the result, b) write the expression for the likelihood function as well as computing its values for the two sets of parameters, c) include argumentation for the result.
Solution:

a) To ensure that the pdf integrates to one:

$$1 = \iint f_\theta(x, y)\, dxdy = \int_{y_{min}}^{y_{max}} \int_{x_{min}}^{x_{max}} c\, dxdy = c(x_{max} - x_{min})(y_{max} - y_{min})$$
$$= \text{volume under the graph}$$
$$\Rightarrow \quad c = \frac{1}{(x_{max} - x_{min})(y_{max} - y_{min})}$$

b) The likelihood of the points (0,0), (0,1), (1,1) and (2,2) given the window parameters $\theta_1 = (-1, 4, -1, 3)$ equals

$$L\left((0,0),(0,1),(1,1),(2,2);\theta_1\right) = f_{\theta_1}(0,0) f_{\theta_1}(0,1) f_{\theta_1}(1,1) f_{\theta_1}(2,2) = c^4 = \frac{1}{20^4}$$

and the likelihood given $\theta_2 = (-2, 5, -3, 6)$ equals

$$L\left((0,0),(0,1),(1,1),(2,2);\theta_2\right) = f_{\theta_2}(0,0) f_{\theta_2}(0,1) f_{\theta_2}(1,1) f_{\theta_2}(2,2) = c^4 = \frac{1}{63^4}.$$

c) Decreasing the size of the window increases the likelihood; but only until one of the observations lies outside the window: then, the likelihood equals zero. The maximum likelihood estimator $\hat{\theta}^{\text{ML}} = (\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max})$ thus satisfies

$$\begin{aligned} \hat{x}_{min}((X_1, Y_1), \ldots, (X_4, Y_4)) &= \min(X_1, \ldots, X_4) = 0, \\ \hat{x}_{max}((X_1, Y_1), \ldots, (X_4, Y_4)) &= \max(X_1, \ldots, X_4) = 2, \\ \hat{y}_{min}((X_1, Y_1), \ldots, (X_4, Y_4)) &= \min(Y_1, \ldots, Y_4) = 0, \\ \hat{y}_{max}((X_1, Y_1), \ldots, (X_4, Y_4)) &= \max(Y_1, \ldots, Y_4) = 2 \end{aligned}$$

with $c = 1/4$ using the expression from a).

Instructions:
a) 1 point, b) 1 point, c) 2 points

Coin Game

Exercise 3 (Based on Exercises 3.1, 3.2 and 3.3 in the book, 4 points). Consider the coin game discussed in the book and mentioned in lecture L7. You will compute the posterior distribution $p(r|y_N)$ for three different priors:

a) For $\alpha = \beta = 1$, the beta distribution becomes uniform between 0 and 1. In particular, if the probability of a coin landing heads is given by r and a beta prior is placed over r, with parameters $\alpha = 1 = \beta$, then this prior density can be written as:

$$p(r) = 1 \quad (0 \leq r \leq 1).$$

Using this prior and the binomial likelihood, compute the posterior density for r if y heads are observed in N tosses (i.e. recall that the posterior is also a beta distribution due to conjugacy of the beta prior and bionomial likelihood. Thus multiply this prior by the binomial likelihood and manipulate the result to obtain something that looks like a beta density). What are the parameters δ and γ of this posterior beta density?

b) Determine an expression of the posterior beta-density and find its parameters (as in exercise a.) for the following prior, which is also a particular form of the beta density:

$$p(r) = \begin{cases} 2r & 0 \leq r \leq 1 \\ 0 & o.w. \end{cases}$$

What are the values of the prior parameters α and β that result in $p(r) = 2r$? What are the parameters δ and γ of the posterior beta density?

c) Determine an expression of the posterior beta-density and find its parameters (as in exercise a.) for the following prior, again a particular form of the beta density:

$$p(r) = \begin{cases} 3r^2 & 0 \leq r \leq 1 \\ 0 & o.w. \end{cases}$$

What are the prior parameters α and β as well as the posterior parameters δ and γ in this case?

Deliverables. a) The posterior expression, b) the posterior expression and the parameter values α and β of the prior, c) the posterior expression and the parameter values α and β of the prior.

Solution:

In this exercise, you can check your solution by verifying it against the general analytical formula for the posterior density: If your prior $p(r)$ is a beta distribution with parameters α and β:

$$p(r) \propto r^{\alpha-1}(1-r)^{\beta-1}$$

then your posterior $p(r|y_N)$ is a beta distribution with parameters δ and γ:

$$p(r|y_N) \propto r^{\delta-1}(1-r)^{\gamma-1}$$

where the parameters are given by

$$\begin{aligned} \delta &= y_N + \alpha \\ \gamma &= N - y_N + \beta. \end{aligned}$$

a) The posterior is

$$\begin{aligned} p(r|y_N) &\propto P(y_N)|r)p(r) && \text{Bayes rule} \\ &= P(y_N|r) && p(r) = 1 \\ &= \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N} \\ &\propto r^{(y_N+1)-1}(1-r)^{(N-y_N+1)-1}, \end{aligned}$$

which is the (un-normalized) beta density with parameters $\delta = y_N + 1$ and $\gamma = N - y_N + 1$. Note that this is in agreement with the general form of the posterior for $\alpha = \beta = 1$.

NB! Here, the exercise is explicitly that the students should multiply out the posterior and do the manipulation themselves. Plugging in $\alpha = 1$ and $\beta = 1$ into the solution from the lecture should only give a 1/2 feelgood-point for trying; this is not a solution

b) The posterior is

$$\begin{aligned} p(r|y_N) &\propto P(y_N)|r)p(r) && \text{Bayes rule} \\ &= \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N} \cdot 2r \\ &\propto r^{y_N+1}(1-r)^{N-y_N} && \propto r^{(y_N+2)-1}(1-r)^{(N-y_N+1)-1}, \end{aligned}$$

which is the (un-normalized) beta density with parameters $\delta = y_N + 2$ and $\gamma = N - y_N + 1$.

Looking at the prior, we see that it is a beta density with $\alpha = 2, \beta = 1$. We can verify that for these parameters, our computed prior is in agreement with the general analytical prior.

c) The posterior is

$$\begin{aligned} p(r|y_N) &\propto P(y_N)|r)p(r) && \text{Bayes rule} \\ &= \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N} \cdot 3r^2 \\ &\propto r^{y_N+2}(1-r)^{N-y_N} && \propto r^{(y_N+3)-1}(1-r)^{(N-y_N+1)-1}, \end{aligned}$$

which is the (un-normalized) beta density with parameters $\delta = y_N + 3$ and $\gamma = N - y_N + 1$.
Looking at the prior, we see that it is a beta density with $\alpha = 3, \beta = 1$. Again, we can verify that for these parameters, our computed prior is in agreement with the general analytical prior.

Instructions:
a) 1 point b) 1.5 point c) 1.5 points

Probabilistic regression

Exercise 4 (Bayesian Regression, 4 points). In this exercise, we will revisit linear regression from a Bayesian perspective. You will imitate the linear Bayesian regression from Lecture L7, but apply it to the Olympic 100m dataset used in the book (found in the file `men-olympics-100.txt`). Here, you should use the years (found in the first column) as your input x_n and the winner times (second column) as your output t_n. We subtract the first year x_1 (which is 1896) from all x_n to make years relative to the first year and scale by dividing by 4 (the Olympics is usually every 4 years) just as done in the R&G book section 3.8.5 (this is only done for visualization purposes).

a) Assume that the noise in your generative model of the data is i.i.d. normal distributed with zero mean and variance $\sigma^2 = 10$. What is the likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$ of observing the given target values in the vector $\mathbf{t}$ given the model defined by the vector $\mathbf{w}$ when the input data matrix is $\mathbf{X}$?
b) Assume that the prior distribution for your model parameters $\mathbf{w}$ is normally distributed $p(\mathbf{w}) = \mathcal{N}(\mu_0, \mathbf{\Sigma}_0)$, and let the likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$ of observing the data matrix $\mathbf{X}$ given model parameters $\mathbf{w}$ be given by another normal distribution $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mu_l, \mathbf{\Sigma}_l)$. What is the corresponding posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2)$?
c) Set your prior parameters to be $\mu_0 = [0, 0]^T$ and $\mathbf{\Sigma}_0 = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$. Assume that your prior and likelihood are both normally distributed as in b). Implement a function that computes the corresponding posterior probability density and add it to `A4_Ex2.py`.
d) Add code to `A4_Ex2.py` that use your function from c) on the dataset. What is the mean and covariance of the posterior probability density after seeing the entire dataset? Use the function `visualize_model` in `A4_Ex2.py` to visualize the posterior distribution and samples from this distribution in the form of random lines.

Deliverables. a) The mathematical expression of this likelihood and arguments for how you reach this, b) a mathematical expression for the posterior distribution and arguments for how you reach this, c) a code snippet showing your function in the report, d) report the computed posterior mean and covariance for the dataset, and include the plots generated by `visualize_model` in the report and comment on what you see from the plot.
Solution:

a) Since we assume the noise is i.i.d. the likelihood is given by

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 I_N) = \mathcal{N}(\mathbf{X}\mathbf{w}, 10 \cdot I_N)$$

as in chapter 3.8.2 of the book.
b) The posterior is given, as in Chapter 3.8.4 of the book, by $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mu_\mathbf{w}, \mathbf{\Sigma}_\mathbf{w})$, where $\mathbf{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \mathbf{\Sigma}_0^{-1}\right)^{-1}$ and $\mu_\mathbf{w} = \mathbf{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \mathbf{\Sigma}_0^{-1}\mu_0\right)$
c) See `A4_Ex2_solution.py`.
d) See `A4_Ex2_solution.py`. Notice since we are sampling in `visualize_model` the sample plot will vary and the produced lines can appear to be all over the place - this is correct behaviour.

Instructions:
a) 1 point b) 1 point c) 1 points (this is just a function) d) 1 point (putting the function to work on the dataset)