Homework 4

Due October 10 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Half life) The half life of a radioactive substance is a way to quantify how rapidly the substance decays. Given a fixed quantity of the substance, the half time is the time that it takes for it to be reduced to half (i.e. half of the radioactive particles have decayed). It is not immediately apparent why the time should be the same for any quantity. Here we show that it is (probabilistically), as long the particles decay following an exponential distribution.

   (a) Let $\tilde{t}$ be a random variable with a pdf of the form

   $$f_{\tilde{t}}(t) := \begin{cases} \lambda \exp(-\lambda t), & \text{if } t \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

   where λ is a fixed constant. We define the half life $t_{1/2}$ as the number that satisfies $\mathrm{P}(\tilde{t} > t_{1/2}) = 1/2$. Compute $t_{1/2}$ in terms of λ. Then explain intuitively why this is a reasonable definition for the half life.

   (b) Compute t such that $\mathrm{P}(t_{1/2} < \tilde{t} < t) = 1/4$, and express it in terms of only $t_{1/2}$. Explain why the result is consistent with the intuitive meaning of half life.

   (c) Compute $\mathrm{P}(\tilde{t} > k t_{1/2})$ for any integer k. Again, explain why the result is consistent with the intuitive meaning of half life.

2. (Triangular pdf) We are interested in fitting a model with a parametric pdf equal to

   $$f_w(x) = \begin{cases} \frac{2x}{w^2}, & \text{for } 0 \leq x \leq w, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

   where the parameter w is nonnegative.

   (a) The observed values are 1.25, 0.4, 1.5, 1, 1.2. What are the possible values of the parameter w?

(b) Compute the likelihood function corresponding to these data and sketch it.

(c) What is the maximum likelihood estimate of w?

(d) Assume that the data are indeed generated by the parametric model with $w := w_{\text{true}}$. Does the ML estimate systematically underestimate or overestimate the true parameter?

(e) Generate a sample from a random variable with this parametric distribution, where $w := 2$, using a uniform sample from the interval $[0, 1]$ equal to 0.64.

3. (Planet) An astrophysicist determines that a good model for the pdf of the temperature in a newly discovered planet is

$$f_{\tilde{t}}(t) := \frac{\lambda \exp(-\lambda |t|)}{2}, \tag{3}$$

where t can be any real number (in particular it can be negative or positive).

(a) Compute the cdf of $\tilde{t}$.

(b) Compute the maximum-likelihood estimate of λ from the following data: 5, -50, -1, 100

(c) What is the pdf of $\tilde{t}$ conditioned on the event $\tilde{t} > 0$?

4. (Tempature) The tables in *train.csv* and *test.csv* record the daily maximum temperature (TMAX) of Seattle.

(a) Estimate the pdf of TMAX with the following models on the training set. Compare the pdf with a normalized histogram in the test set. Which model performs better visually?

- Estimating the parameter of Gaussian distribution with MLE;
- Non-parametric KDE with the Gaussian kernel at different bandwidths (e.g. 1, 2, 5).

(b) Repeat the experiment only on July and August data. Which model performs better visually? Compare the results with (a) and explain your findings.