Homework 4

Due October 15 at 11 pm

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

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

   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.

   We have the following derivations:

   $$\mathbb{P}\left(\tilde{t} > t_{\frac{1}{2}}\right) = \frac{1}{2} \qquad \text{(From the problem)}$$

   $$\int_{\mathrm{t}}^{\infty} \lambda \mathrm{e}^{-\lambda t} \mathrm{d}t = \frac{1}{2} \qquad \text{(By definition of probability measure)}$$

   $$e^{-\lambda t_{\frac{1}{2}}} = \frac{1}{2} \qquad \text{(intergation)}$$

   $$t_{\frac{1}{2}} = \frac{\ln 2}{\lambda} \qquad \text{(Algebraic operations)}$$

   One way of explaning this is through differential equations. Suppose we have a particle with initial mass x_0 kg, and the rate of decay is $\lambda(kg \times s^{-1})$, within a small amount of time δt we have $\Delta x \approx -\lambda \cdot x \Delta t$. In the limit term, as $\Delta t \to 0$, we have $\frac{dx}{dt} = -\lambda x$, we solve this to get $x(t) = x_0 e^{-\lambda t}, t > 0$. Now the half life is the timestamp $t_{\frac{1}{2}}$ such that $x_0 e^{-\lambda t_{\frac{1}{2}}} = \frac{1}{2} x_0$(the weight is decayed to the half of the initial mass), where $t_{\frac{1}{2}} = \frac{ln2}{\lambda}$, which is what we have derived before using probabilistic perspective.

(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.

$$\begin{aligned}
\mathbb{P}\left(\mathrm{t}_{\frac{1}{2}} < \tilde{t} < \mathrm{t}\right) &= \frac{1}{4} && \text{(From the question)} \\
\mathbb{P}(\tilde{t} < \mathrm{t}) - \mathbb{P}\left(\tilde{t} \leq t_{\frac{1}{2}}\right) &= \frac{1}{4} && \text{(By definition)} \\
\mathbb{P}(\tilde{t} < \mathrm{t}) - \frac{1}{2} &= \frac{1}{4} && \text{(Rearrange)} \\
F_{\tilde{t}}(t) &= \frac{3}{4} && \text{(By definition)} \\
1 - \mathrm{e}^{-\lambda \mathrm{t}} &= \frac{3}{4} && \text{(Rearrange)} \\
\mathrm{e}^{-\lambda \mathrm{t}} &= \frac{1}{4} && \text{(Rearrange)} \\
t &= \frac{2\ln(2)}{\lambda} && \text{(Solve)} \\
t &= 2t_{\frac{1}{2}} && \text{(Substitute from part (a))}
\end{aligned}$$

.

We could explain this as follows:

i. Linear Invariant of the solution of the linear ODE: The input to the system is $x(t_{\frac{1}{2}}) = \frac{1}{2}x_0$ instead of x_0 previously(shifted rightwards by $t_{\frac{1}{2}}$), so by the LTI of the solution we have $x_{new}(t) = x(t - t_{\frac{1}{2}})$ and we know the solution is $x_{new}(t) = x(t_{\frac{1}{2}})e^{-\lambda(t-t_{\frac{1}{2}})}$. Letting $x_{new}(t) = \frac{1}{2} \times \frac{1}{2}x_0$ we get $t = 2t_{\frac{1}{2}}$.

ii. Probablistic Perspective: Imagine the area under the exponential pdf as the total mass that has been decayed. Since $F_{\tilde{t}}(t) = 1 - e^{-\lambda t}$(area under pdf on the interval $[0,\ t]$) is the same as $x_0 = 1$(initial mass) and $x(t) = 1 \times e^{-\lambda t}$(mass at time t). Then the half life is just computing the median splitting point such that the mass of particle that has been decayed is the same as those to be decayed in the future.

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

We compute as follows:

$$\begin{aligned}
\mathbb{P}(\tilde{t} > kt_{\frac{1}{2}}) &= 1 - F_{\tilde{t}}(kt_{\frac{1}{2}}) && \text{(By definition)} \\
&= 1 - (1 - e^{-\lambda k t_{\frac{1}{2}}}) && \text{(By definition)} \\
&= e^{-\lambda k \frac{ln2}{\lambda}} && \text{(From previous part)} \\
&= (\frac{1}{2})^k && \text{(By definition)}
\end{aligned}$$

, from probabilistic perspective we know that this is right since we are dividing the area under pdf by half k times.

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

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

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?

Since we can observe these value, then the probability for them should not be zero, which means $w \ge \max\{1.25, 0.4, 1.5, 1, 1.2\} = 1.5$.

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

The likelihood function is computed as follows where the green auxiliary line is the constraint surface:

$$\begin{aligned} \mathcal{L}(w) &= \frac{2^5}{w^{10}} \prod_{i=1}^{5} x_i && \text{(By definition)} \\ &= \frac{1.25 \times 0.4 \times 1.5 \times 1 \times 1.2}{w^{10}} && \text{(Plug in the value)} \\ &= \frac{28.8}{w^{10}} && \text{(Simplify)} \end{aligned}$$

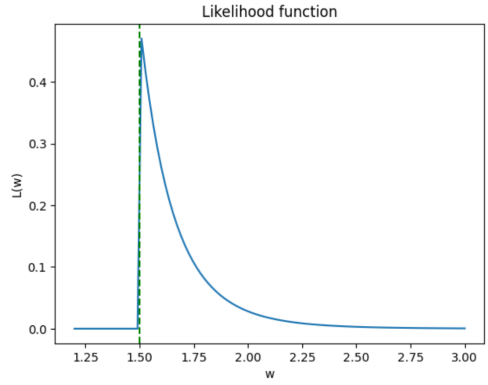
. Also considering the case where $w < 1.5$ we will have the complete definition of this function, which is $\mathcal{L}(w) = \begin{cases} \frac{28.8}{w^{10}} & w \ge 1.5 \\ 0 & 0 \le w < 1.5 \end{cases}$

Codes and plots are as follows:

```
In [1]: import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline
        import pandas as pd

In [50]: w_axis = np.linspace(1.2,3,100)
         plt.plot(w_axis, [(lambda w: 28.8/w**10)(w) if w >= 1.5 else 0 for w in w_axis])
         plt.axvline(1.5, c="g", linestyle="--")
         plt.title("Likelihood function")
         plt.xlabel("w")
         plt.ylabel("L(w)")

Out[50]: Text(0, 0.5, 'L(w)')
```



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

We first compute the log-likelihood function as $log\mathcal{L}(w) = ln(28.8) - 10lnw$, which is a monotonically decreasing function on $w \geq 1.5$, which means $w_{MLE} = \arg\max_{w \geq 1.5} \mathcal{L}(w) = 1.5$.

(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?

Since we have derived in the part (a) that the domain of the likelihood function $\mathcal{L}(w)$ is $[\max\{x_1, x_2, \cdots, x_n\}, \infty)$, and the MLE for $log\mathcal{L}(w)$ is just $\max\{x_1, x_2, \cdots, x_n\}$, so even if the data are generated by w_{true}, the $\max\{x_1, x_2, \cdots, x_n\}$ is not guaranteed to be w_{true}. Moreover, with finite data, the MLE is biased. In this example, we want to compute the pdf of $Y = \max\{x_1, x_2, x_3, x_4\}$ where $x_i \sim f_w(x)$, the computation is as follows:

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leq y) && (3)\\
&= \mathbb{P}(X_1 \leq y, X_2 \leq y, X_3 \leq y, X_4 \leq y) && (4)\\
&= F_X^4(y) && (5)\\
&= \begin{cases} 1 & y \geq w \\ (\int_{-\infty}^{y} \frac{2x}{w^2} dx)^4 = \frac{y^8}{w^8} & 0 < y < w \\ 0 & y < 0 \end{cases} && (6)\\
& && (7)
\end{aligned}
$$

, taking the derivative we get $f_Y(y) = \begin{cases} \frac{8y^7}{w^8} & 0 < y < w \\ 0 & otherwise \end{cases}$. We then compute the expectation of Y, we get $\mathbb{E}[Y] = \int_0^w y \times \frac{8y^7}{w^8} dy = \frac{8}{9}w < w$(w is assumed to be positive in this problem). In this case, the MLE systemically underestimates 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.

We first compute the CDF for X, which is $\int_0^x \frac{2x}{4} dx = \frac{x^2}{4}$. We then compute the inverse of this function, which is $F_w^{-1}(u) = 2\sqrt{u}$. Finally, we sample a data point from uniform distribution on $[0, 1]$, which is given as 0.64 and fit this in the inverse CDF and get $x = 2\sqrt{0.64} = 0.16$.

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

$$f_{\tilde{t}}(t) := \frac{\lambda \exp(-\lambda |t|)}{2}, \tag{8}$$

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

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

We break up into two cases: When $t < 0$, we have:

$$\begin{aligned} F_{\tilde{t}}(t) &= \int_{-\infty}^{t} \frac{\lambda e^{\lambda t}}{2} dt \\ &= \frac{\lambda}{2} \frac{e^{\lambda t}}{\lambda} \Big|_{-\infty}^{t} \\ &= \frac{e^{\lambda t}}{2} \end{aligned}$$

When $t > 0$, we have:

$$\begin{aligned} F_{\tilde{t}}(t) &= \int_{-\infty}^{0} \frac{\lambda e^{\lambda t}}{2} dt + \int_{0}^{t} \frac{\lambda e^{-\lambda t}}{2} dt \\ &= \frac{\lambda}{2} \frac{e^{\lambda t}}{\lambda} \Big|_{-\infty}^{0} + \frac{\lambda}{2} \frac{e^{-\lambda t}}{-\lambda} \Big|_{0}^{t} \\ &= \frac{1}{2} + \frac{1 - e^{-\lambda t}}{2} \end{aligned}$$

Thus, to conclude, we have $F_{\tilde{t}}(t) = \begin{cases} \frac{e^{\lambda t}}{2} & t < 0 \\ \frac{1}{2} + \frac{1-e^{-\lambda t}}{2} & t > 0 \end{cases}$.

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

We first compute the pdf of the $\tilde{t}$, which is $f_{\tilde{t}}(t) = \frac{\lambda e^{-\lambda|t|}}{2}, \forall t \in \mathbb{R}$.
Then the likelihood function is given by $\mathcal{L}(\lambda) = \prod_{i=1}^{4} \frac{\lambda e^{-\lambda|t_i|}}{2} = (\frac{\lambda}{2})^4 e^{-\sum_{i=1}^{4} -\lambda|t_i|}$.
We then compute the log-likelihood function, which is $log\mathcal{L}(w) = 4ln(\frac{\lambda}{2}) - \lambda \sum_{i=1}^{4} |t_i|$.
Taking the derivative w.r.t λ and set it to zero we could have $4 \times \frac{1}{\frac{\lambda}{2}} \times \frac{1}{2} - \sum_{i=1}^{4} |t_i| = 0$ and that $\lambda_{MLE} = \frac{4}{\sum_{i=1}^{4} |t_i|} = \frac{4}{5+50+1+100} = \frac{1}{39}$.

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

We first compute $P(\tilde{t} > 0) = 1 - P(\tilde{t} < 0) = 1 - F_{\tilde{t}}(0) = 1 - \frac{1}{2} = \frac{1}{2}$. We want to compute the conditional pdf, which is given by $f_{\tilde{t}|\tilde{t}>0}(t) = \frac{\frac{\lambda e^{-\lambda t}}{2}}{P(\tilde{t}>0)} = \frac{\frac{\lambda e^{-\lambda t}}{2}}{\frac{1}{2}} = \lambda e^{-\lambda t}$ where $t > 0$.

4. (Temperature) 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).

   From the notes we konw that:

   i. The MLE for Gaussian is $\begin{cases} \mu_{ML} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \sigma^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{ML})^2 \end{cases}$. and the shape of pdf is determined by these parameters.

   ii. The kde pdf is given by the data point

   , we draw these plots as follows, visually, the kde model performs better.

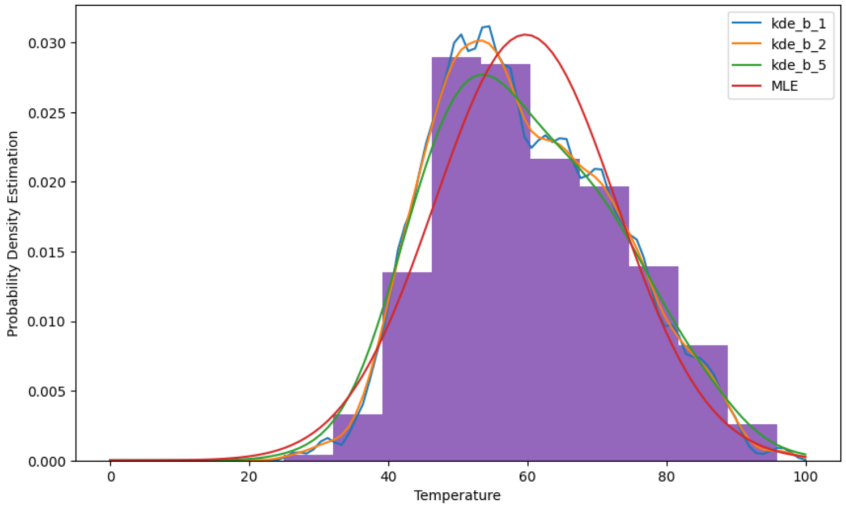


Figure 1: All Month

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

The graphs and codes are as follows, visually, the mle model performs better.

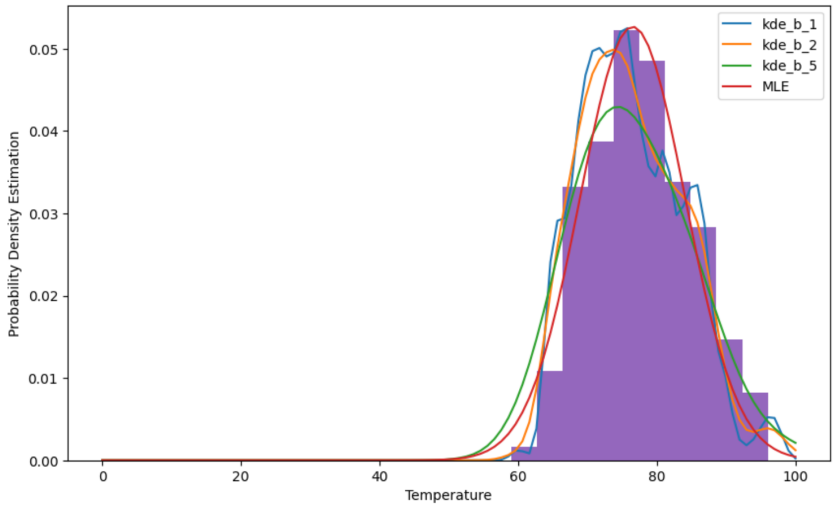


Figure 2: July and August

```
# kde function at an arbitrary data point
def gaussian_kernel(t, center, bandwidth):
    return 1 / (bandwidth * np.sqrt(2 * np.pi))
    * np.exp(- 0.5 * ((t - center)/ bandwidth) ** 2);

# mle gaussian
def gaussian_pdf(t, mu_hat, sigma_hat):
    return (1 / (sigma_hat * np.sqrt(2 * np.pi)))
    * np.exp(-0.5 * ((t - mu_hat) / sigma_hat) ** 2)


# Compute kde for a single value
def kde(t, data_points, bandwidth):
    return (1 / len(data_points))
    * sum([gaussian_kernel(t, data, bandwidth) for data in data_points]);


def train_kde(data_points, bandwidth):
    def trained_kde(t):
        return kde(t, data_points, bandwidth)

    return trained_kde


def train_mle(data_points):
    mu_hat = np.mean(data_points)
    sigma_hat = np.sqrt(np.mean((data_points - np.mean(data_points))**2))

    def trained_mle(t):
        return gaussian_pdf(t, mu_hat, sigma_hat)

    return trained_mle


def plot_kde(axis, trained_kde, label):
    plt.plot(axis, [trained_kde(point) for point in axis], label=label)


def plot_MLE(axis, trained_mle):
    plt.plot(axis, [trained_mle(point) for point in axis], label="MLE")


def plot_on_test(axis,trained_kdes, trained_mle, testing_data):
    for trained_kde in trained_kdes:
        plot_kde(axis, trained_kde, label="kde_b_"+trained_kde.name)
    plot_MLE(axis, trained_mle)
    plt.hist(testing_data, bins = 10, density="true")
    plt.legend()

```

```
50
51  def run_plot_main(training_data, testing_data):
52      plt.figure(figsize=(10, 6))
53      bandwidths = [1,2,5]
54      trained_kdes = []
55      for bandwidth in bandwidths:
56          trained_kde = train_kde(training_data, bandwidth)
57          trained_kde.name = str(bandwidth)
58          trained_kdes.append(trained_kde)
59      trained_mle = train_mle(training_data)
60
61      x_axis = np.linspace(0,100,100)
62      plot_on_test(x_axis,trained_kdes, trained_mle, testing_data)
63
64
65  def run_program_on_month(training_data, testing_data, month="All"):
66
67      if month == "All":
68          run_plot_main(training_data["TMAX"], testing_data["TMAX"])
69      else:
70          training_data = training_data[training_data["month"] == month]["TMAX"]
71          testing_data = testing_data[testing_data["month"] == month]["TMAX"]
72          run_plot_main(training_data, testing_data)
73
74
75  def main():
76      df_train = pd.read_csv("./weather_train.csv")
77      df_test = pd.read_csv("./weather_test.csv")
78      run_program_on_month(df_train, df_test)
79
80  main()
```