Homework Assignment #4

21600004

Kang Seok-Un

Introduction

In this homework, implements plotting histogram, kernel density estimation(KDE), and Gaussian distribution with MLE by using given a file weight-height.csv(https://www.kaggle.com/mustafaali96/weight-height)

Excepting histogram, however, must implement my own, not using the library.

Each of the six must be implemented;
height all, height male, height female, weight all, weight male, weight female

In this report, the experiment is conducted in the following order.

1. Histogram

2. Gaussian distribution with MLE

3. KDE

Experiment

Before starting, it was necessary to know what is this data. Table 1 shows the number and average population variance for each data.

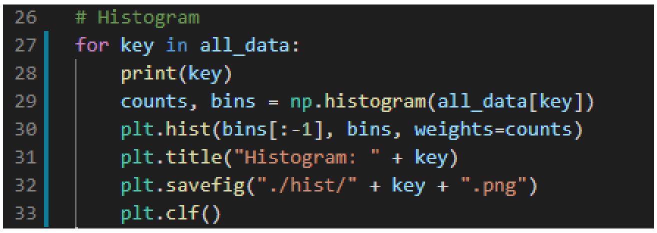
Table 1. Information of Data

Name	# of data	μ	σ^2
Height All	10,000	66.368	14.802
Height Male	5,000	69.026	8.197
Height Female	5,000	63.709	7.2685
Weight All	10,000	161.440	1030.849
Weight Male	5,000	187.021	187.021
Weight Female	5,000	135.860	361.782

1. Histogram

The histogram is the simplest form of non-parametric density estimation.

I used the 'numpy.hist()' existing in the numpy libarary to plot the histogram of the dataset as figure 1. Figure 3. shows the results of plotting each data set in histogram.



```python
# Histogram
for key in all_data:
    print(key)
    counts, bins = np.histogram(all_data[key])
    plt.hist(bins[:-1], bins, weights=counts)
    plt.title("Histogram: " + key)
    plt.savefig("./hist/" + key + ".png")
    plt.clf()
```

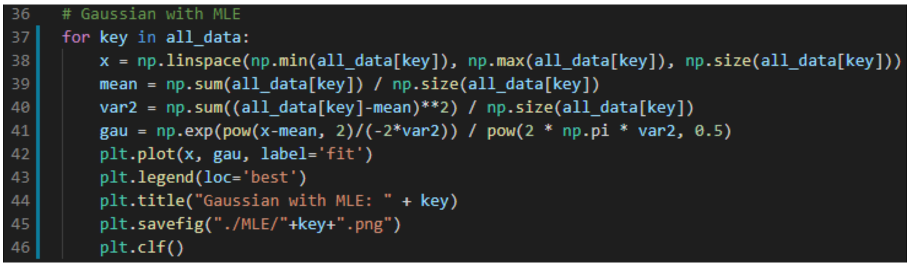
Figure 1. Implementation code about Histogram

2. Gaussian Distribution with MLE

To implement Gaussian distribution with MLE, I use formula (1).

$$\mathrm{N}(\mathrm{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2 \qquad (1)$$

In line 38 of figure 2, the x-axis is set as the range between the minimum and maximum values of the dataset. Next, the mean and population variance of the dataset is calculated to use Equation (1). Line 41 is a part implemented by transferring formula (1) to code. Lines 42 to 46 are the parts where plotting is performed using the values obtained earlier.



```
36    # Gaussian with MLE
37    for key in all_data:
38        x = np.linspace(np.min(all_data[key]), np.max(all_data[key]), np.size(all_data[key]))
39        mean = np.sum(all_data[key]) / np.size(all_data[key])
40        var2 = np.sum((all_data[key]-mean)**2) / np.size(all_data[key])
41        gau = np.exp(pow(x-mean, 2)/(-2*var2)) / pow(2 * np.pi * var2, 0.5)
42        plt.plot(x, gau, label='fit')
43        plt.legend(loc='best')
44        plt.title("Gaussian with MLE: " + key)
45        plt.savefig("./MLE/"+key+".png")
46        plt.clf()
```

Figure 2. Implementation code about Gaussian Distribution with MLE

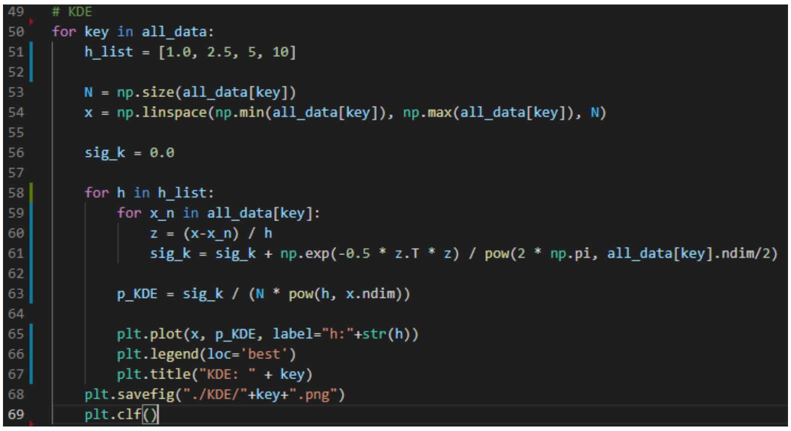
3. KDE

To implement KDE, I use above equation (2) and (3).

$$\mathrm{P}_{\mathrm{KDE}}(x) = \frac{1}{Nh^D}\sum_{n=1}^{N} K\left(\frac{x-x_n}{h}\right) \qquad (2)$$

$$\mathrm{K}(\mathrm{z}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\exp(-\frac{1}{2}z^T z) \qquad (3)$$

Like Gaussian distribution with MLE, the x-axis is set as the range between the minimum and maximum values of the dataset. the parameter z of function K is calculated at line 60 in figure 3. And line 62 and 63, equation (2) and (3) is implemented.



```
49    # KDE
50    for key in all_data:
51        h_list = [1.0, 2.5, 5, 10]
52
53        N = np.size(all_data[key])
54        x = np.linspace(np.min(all_data[key]), np.max(all_data[key]), N)
55
56        sig_k = 0.0
57
58        for h in h_list:
59            for x_n in all_data[key]:
60                z = (x-x_n) / h
61                sig_k = sig_k + np.exp(-0.5 * z.T * z) / pow(2 * np.pi, all_data[key].ndim/2)
62
63            p_KDE = sig_k / (N * pow(h, x.ndim))
64
65            plt.plot(x, p_KDE, label="h:"+str(h))
66            plt.legend(loc='best')
67            plt.title("KDE: " + key)
68        plt.savefig("./KDE/"+key+".png")
69        plt.clf()
```

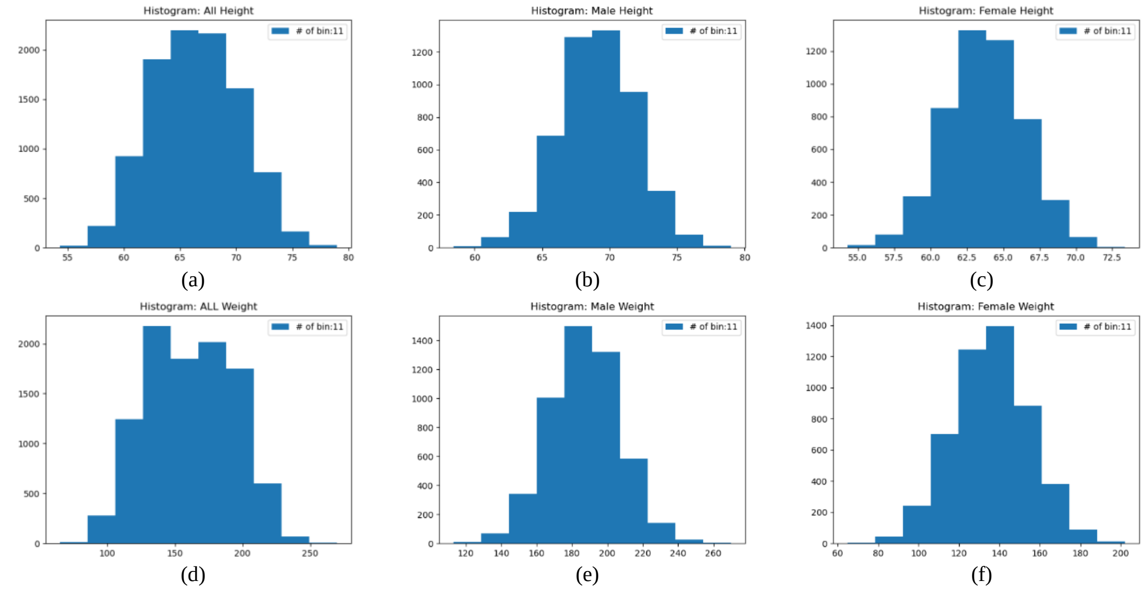
Figure 3. Implementation code about KDE

Result

1. histogram

The size of the bin of each dataset is 11. When looking at the height-related histogram (figure 4; a, b, and c), it can be seen that the average of males is greater than females. Likewise, when looking at the weight-related histogram (figure 4; d, e, and f), it can be seen that the average of males is greater than the average of females.

Men's height is generally larger than women's, indicating that the results came out well.



(a)

(b)

(c)

(d)

(e)

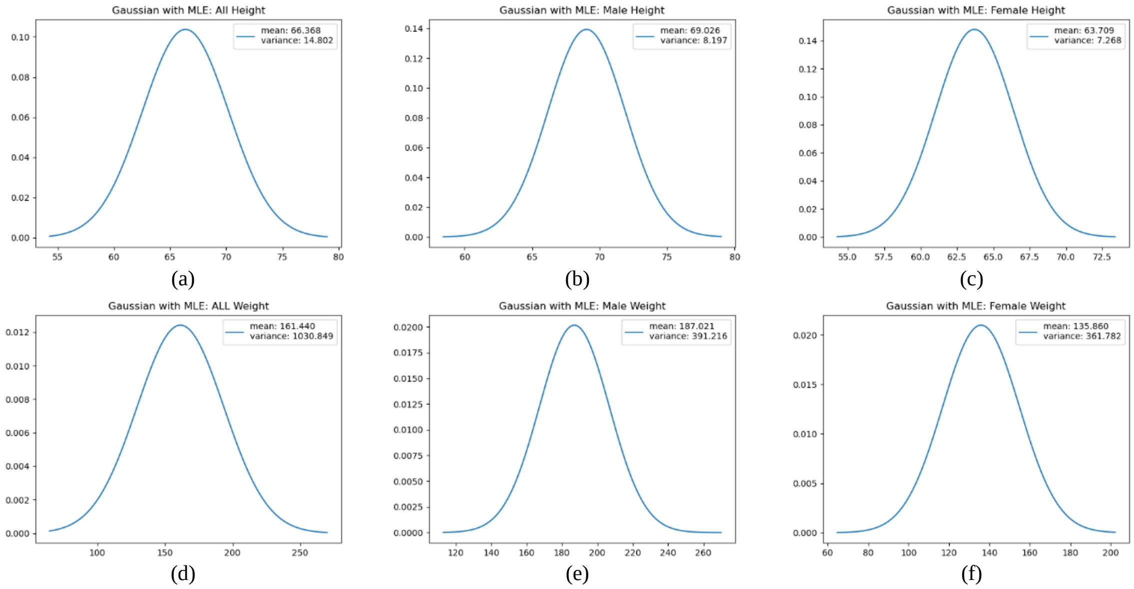
(f)

Figure 4. histogram: (a) Height All, (b) Height Male, (c) Height Female, (d) Weight All, (e) Weight Male, (f) Weight Female

2. Gaussian Distribution with MLE

When looking at the height-related Gaussian distribution (figure 5; a, b, and c), it can be seen that the average of males is greater than females. Likewise, when looking at the weight-related Gaussian distribution (figure 4; d, e, and f), it can be seen that the average of males is greater than the average of females.

Men's height and weight are generally larger than women's, indicating that the results came out well. In addition, unlike histograms, Gaussian distributions can analyze graphs in terms of probability.



(a) (b) (c)

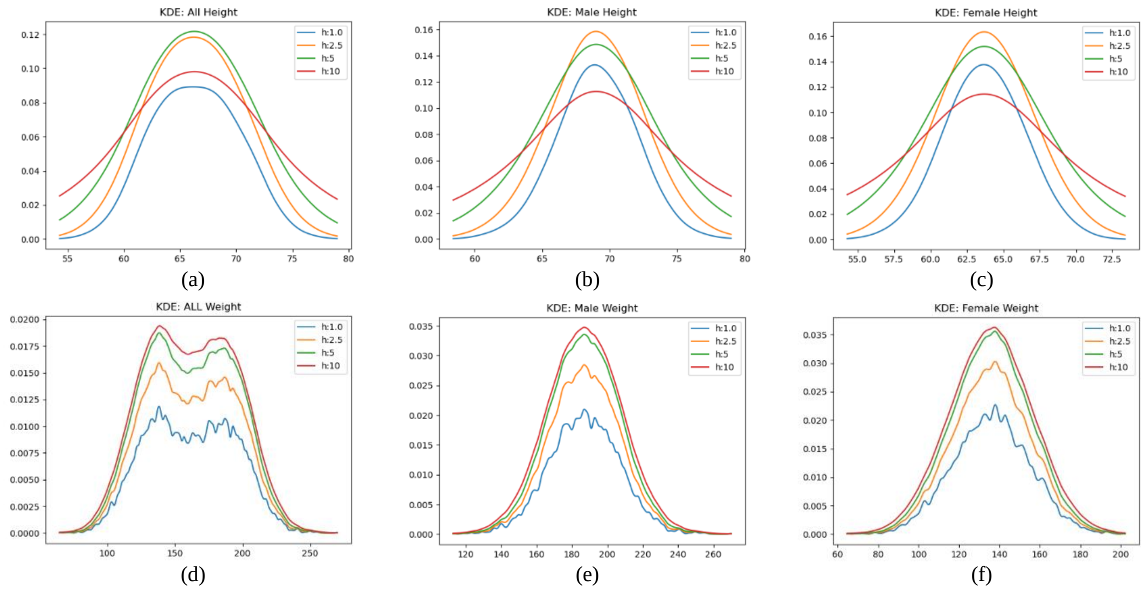
(d) (e) (f)

Figure 5. Guassian Distribution with MLE:
(a) Height All, (b) Height Male, (c) Height Female, (d) Weight All, (e) Weight Male, (f) Weight Female

3. KDE

KDE generates a kernel function centered on the corresponding data value for each observed data. Therefore, rather than the probability distribution function appearing smoothly like Gaussian distribution, an unsmooth graph may be drawn as shown in the weight of Figure 6.

However, it is greatly influenced by the bandwidth value h. Therefore, if look at d, e, and f in Figure 6, can observe that h is smoothed more when it is 10 than when it is 1.



(a) (b) (c)

(d) (e) (f)

Figure 6. KDE: (a) Height All, (b) Height Male, (c) Height Female, (d) Weight All, (e) Weight Male, (f) Weight Female

Conclusion

While performing this task, I feel that there is a big difference between implementing a program for specific data and theoretical study.

In addition, in this task, there was a process of drawing several graphs according to data and method, and I think I was able to understand the difference in results more intuitively by drawing graphs.