CSE512 Spring 2021 - Machine Learning - Homework 1

Name: Sabrina Margetic

Solar ID: 109898930

Net ID Email Address: sabrina.margetic@stonybrook.edu

Names of people whom you discussed the homework with: None (Only the TA)

1 Question 1:

1.1

$E[X] = \int Xf(X)dX$
$E[X] = \int max(x_1, 2x_2)f(x_1, x_2)dX$
We can say that $f(x_1, x_2)$ is equal to $f(x_1)f(x_2)$ because of independence.

Their respective pdf's are:
$f(x_1) = \frac{1}{b-a} = \frac{1}{1-0} = 1$
$f2x_2) = \frac{1}{b-a} = \frac{1}{1-0} = 1$

$\therefore E[X] = \int max(x_1, 2x_2)dX$
We can now split this maximum function into two in order to account for two different scenarios, one in which x_1 is max, and one in which $2x_2$ is max.
$\therefore E[X] = \int_0^1 [\int_0^{\frac{x_1}{2}} max(x_1, 2x_2)dx_2 + \int_{\frac{x_1}{2}}^1 max(x_1, 2x_2)dx_2]dx_1$
$= \int_0^1 [\int_0^{\frac{x_1}{2}} x_1 dx_2 + \int_{\frac{x_1}{2}}^1 2x_2 dx_2]dx_1$
$= \int_0^1 [x_1x_2|_0^{\frac{x_1}{2}} + \frac{2x_2^2}{2}|_{\frac{x_1}{2}}^1]dx_1$
$= \int_0^1 [x_1x_2|_0^{\frac{x_1}{2}} + x_2^2|_{\frac{x_1}{2}}^1]dx_1$
$= \int_0^1 [\frac{x_1^2}{2} + 1^2 - (\frac{x_1}{2})^2]dx_1$
$= \int_0^1 [\frac{x_1^2}{2} + 1 - \frac{x_1^2}{4}]dx_1$
$= \int_0^1 [\frac{x_1^2}{4} + 1]dx_1$
$= \int_0^1 \frac{x_1^2}{4}dx_1 + \int_0^1 1dx_1$
$= \frac{x_1^3}{12}|_0^1 + x_1|_0^1$
$= \frac{1}{12} + 1$
$= \frac{13}{12}$

1.2

$Var(X) = E[X^2] - E[X]^2$

$E[X^2] = \int (max(x_1, 2x_2)^2 f(x_1, 2x_2)dX$
$= \int (max(x_1, 2x_2)^2 dX$
Once again, we can split this based on the two scenarios: x_1 is max, and $2x_2$ is max.
$\therefore E[X^2] = \int_0^1 [\int_0^{\frac{x_1}{2}} max(x_1, 2x_2)^2 dx_2 + \int_{\frac{x_1}{2}}^1 max(x_1, 2x_2)^2 dx_2]dx_1$
$= \int_0^1 [\int_0^{\frac{x_1}{2}} x_1^2 dx_2 + \int_{\frac{x_1}{2}}^1 4x_2^2 dx_2]dx_1$
$= \int_0^1 [x_1^2x_2|_0^{\frac{x_1}{2}} + \frac{4x_2^3}{3}|_{\frac{x_1}{2}}^1]dx_1$
$= \int_0^1 [(x_1^2)(\frac{x_1}{2}) + [\frac{4}{3} - \frac{4(\frac{x_1}{2})^3}{3}]dx_1$

$= \int_0^1 [\frac{x_1^3}{2} + \frac{4}{3} - \frac{4(\frac{x_1^3}{8})}{3}]dx_1$
$= \int_0^1 [\frac{x_1^3}{2} + \frac{4}{3} - \frac{(\frac{x_1^3}{2})}{3}]dx_1$
$= \int_0^1 [\frac{x_1^3}{2} + \frac{4}{3} - \frac{x_1^3}{6}]dx_1$
$= \int_0^1 \frac{x_1^3}{3}dx_1 + \int_0^1 \frac{4}{3}dx_1$
$= \frac{x_1^4}{12}|_0^1 + \frac{4}{3}|_0^1$
$= \frac{1}{12} + \frac{4}{3}$
$= \frac{17}{12}$

$\therefore Var(X) = \frac{17}{12} - \frac{13}{12} = \frac{5}{12}$
where E[X] was obtained from part 1.

1.3

$Cov(X, x_1) = Cov(max(x_1, 2x_2), x_1)$
$= E[max(x_1, 2x_2)x_1] - E[max(x_1, 2x_2)]E[x_1]$

$E[x_1] = \int_0^1 x_1 f(x_1)dx_1 = \int_0^1 x_1 dx_1 = \frac{x_1^2}{2}|_0^1 = \frac{1}{2}$

$E[max(x_1, 2x_2)x_1] = \int\int max(x_1, 2x_2)x_1 dx_1 dx_2$
We can split this up based on the two scenarios once again.
$\therefore E[max(x_1, 2x_2)x_1] = \int_0^1 [\int_0^{\frac{x_1}{2}} max(x_1, 2x_2)x_1 dx_2 + \int_{\frac{x_1}{2}}^1 max(x_1, 2x_2)x_1 dx_2]dx_1$
$= \int_0^1 [\int_0^{\frac{x_1}{2}} x_1^2 dx_2 + \int_{\frac{x_1}{2}}^1 2x_2 x_1 dx_2]dx_1$
$= \int_0^1 [x_1^2 x_2|_0^{\frac{x_1}{2}} + x_1 x_2^2|_{\frac{x_1}{2}}^1]dx_1$
$= \int_0^1 [\frac{x_1^3}{2} + x_1[1 - (\frac{x_1}{2})^2]]dx_1$
$= \int_0^1 [\frac{x_1^3}{2} + x_1[1 - \frac{x_1^2}{4}]]dx_1$
$= \int_0^1 [\frac{x_1^3}{2} + x_1 - \frac{x_1^3}{4}]dx_1$
$= \int_0^1 \frac{x_1^3}{4}dx_1 + \int_0^1 x_1 dx_1$
$= \frac{x_1^4}{16}|_0^1 + \frac{x_1^2}{2}|_0^1$
$= \frac{1}{16} + \frac{1}{2}$
$= \frac{9}{16}$

$\therefore Cov(X, x_1) = \frac{9}{16} - (\frac{1}{2})(\frac{13}{12}) = \frac{1}{48}$
where E[x] was obtained from part 1.

2 Question 2:

2.1

The function get_mean_and_variance(X, y) is written in hw1.py, but relies on data preprocessing that happens in data_preprocessing.py .

The formula used for mean is: $\frac{1}{N}\sum_{x=1}^{N} x_i$
The formula used for variance is: $\frac{1}{N-2}\sum_{i=1}^{N}(x_i - \mu)^2$
where x_i is a data point, mu is the mean and N is the total number of data points.

2.2

To run code with the specific data provided for this part, run "python -c 'import hw1; hw1.question2_functions()'" in the terminal to obtain the mathematical and graphical results from subsection parts a and b. This function does not take in any arguments.

2.2.1

Using the provided data, the values of mu0, var0, mu1, and var1 are:
$mu0 = [50.51685393, \quad 0.47191011]$
$var0 = [233.97994033, \quad 0.25061893]$
$mu1 = [63.38, \quad 0.4]$
$var1 = [2.99587347e+02, \quad 2.44897959e-01]$

Where the first column is age and the second is gender. Gender is also changed so that 0 correlates to Male and 1 correlates to Female.

2.2.2

As you can see from Fig.1 the survival curve is centered more closely to the younger age range, with a mean age of 51. While the death curve is centered at an age of 63, a difference of 7 years. We also see that the survival curve is narrower in comparison with a variance of about 234 compared to the death variance of about 300. The greater variance of deaths most likely accounts for the fact that infants and the elderly are high risk groups.

From Fig. 2 you can see that gender distribution for survival and death is almost the same, they are almost both located at the center. However, it should be noted that both distributions have an average that are closer that of Male. The survival average is .47 and the death average is .40, where once again Male was mapped to 0 and Female was mapped to 1. It is possible that the .47 value is simply due the consequence of the random nature of sample

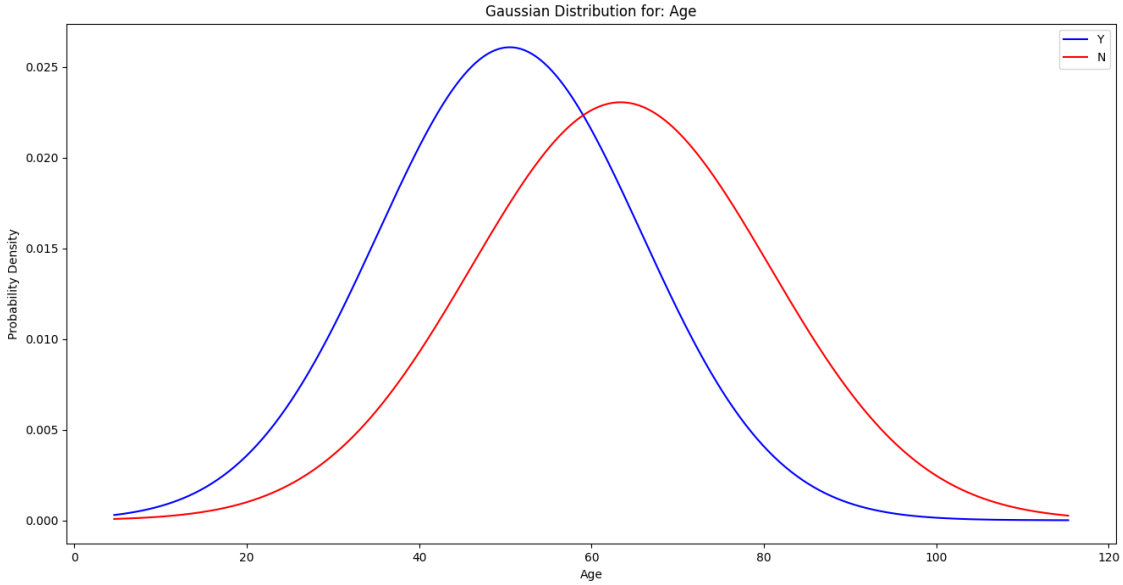

Figure 1: Gaussian age distribution for survival (Y) and death (N).

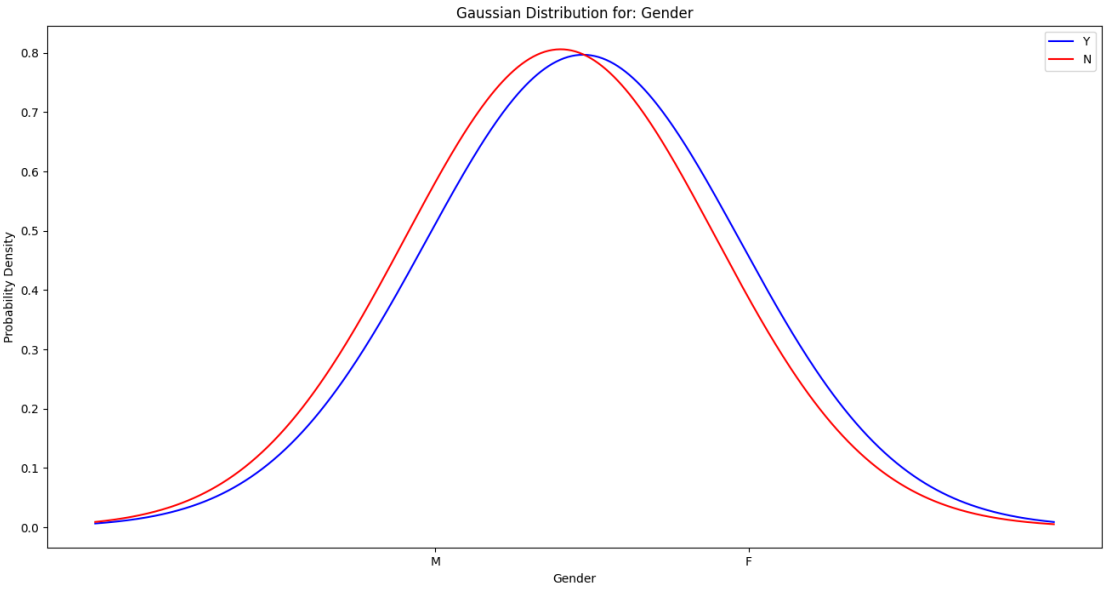

Figure 2: Gaussian death distribution for survival (Y) and death (N).

data. This is also possible for the .40 value, but less likely since there is a percentage difference of 20 % between the two genders. The data indicates that there is a slightly higher death rate amongst males

2.2.3

An important attribute of Gaussian distributions is that they show the location of the most frequent data, the mean, and they are symmetric around such. This allows us to say that a certain percentage of the data occurs within a standard deviation from the mean. However, this is not what we see with gender, there are only two possible values, Male or Female. Therefore, after the distribution has hit the Male and Female marker, there is no more data to accumulate; the concept of a standard deviation away from the mean, is meaningless. Additionally, there is no data located at the mean itself. Once again, only at the Male and Female markers.
While Fig.2 did provide some information, a bar graph could provide the same information.

3 Question 3:

To run code with the specific data provided for this part, run "python -c 'import hw1; hw1.question3_functions()'" in the terminal to obtain the mathematical and graphical results from subsection parts a, b, c and d. This function does not take in any arguments.

3.1

The logic behind the code for this part of the assignment was to change the original data such that the previous 7 days (as a list) was a single input value, with the output as the 8th day death.

3.2

3.2.1

$w = [3.32971812e{-}03, \quad -5.97868846e{-}03, \quad -1.07181360e{-}02, \quad 2.97430033e{-}02, \quad -1.33508726e{-}02, \quad -2.94346183e{-}03, \quad -8.47872012e{-}05, \quad 1.74178344e{+}00, \quad -1.06486597e{+}00, \quad 6.02420242e{-}01, \quad -3.68951686e{-}01, \quad -9.27692517e{-}02, \quad 6.78640538e-01, \quad -4.98447204e-01]$
$b = 58.15728936600499$

3.2.2

As you can see from the graph above, the linear regression model worked extremely well for the data provided. It is almost an exact match.

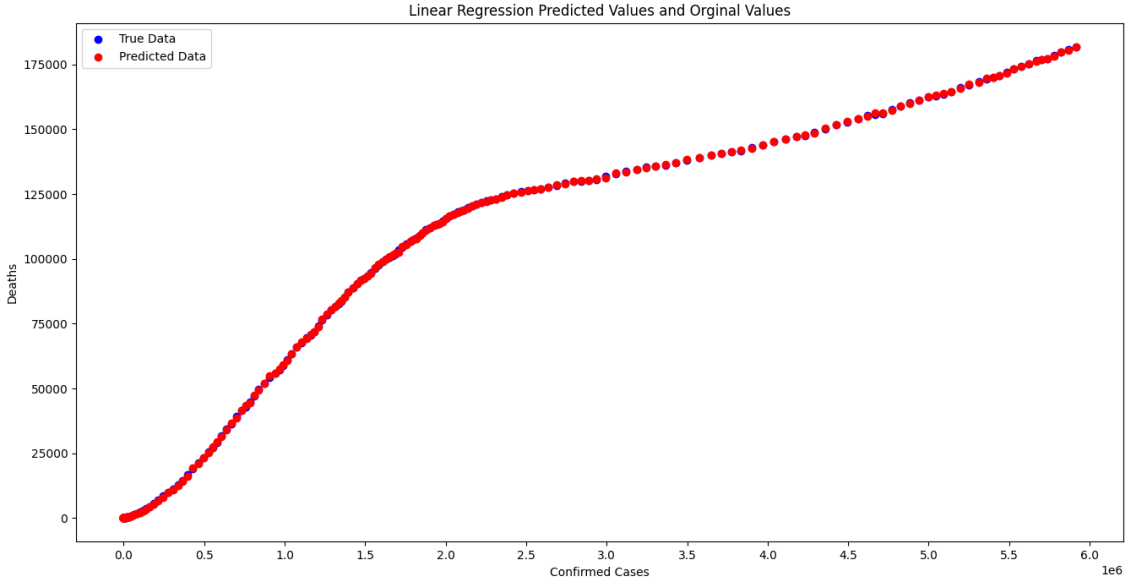

Figure 3: Actual death value versus linear regression predicted death value.

3.2.3

Gaussian distribution parameters for $y_t - \hat{y}_t$:

Mean: 4.705514225528353e-11
Variance: 39949.115786979324

3.2.4

Yes, in this case the Gaussian Distribution is a good model for the data. As you can see in Fig. 4 the histogram produced seems to follow a bell shaped curve. Additionally, the Gaussian curve we model has a mean near to the peak that we observe on this histogram. The Gaussian mean is slightly shifted to the right of this peak due to more data occurring to the right. However, for the most part, this data seems to have a more symmetric nature, which is a requirement for Gaussian Distributions.

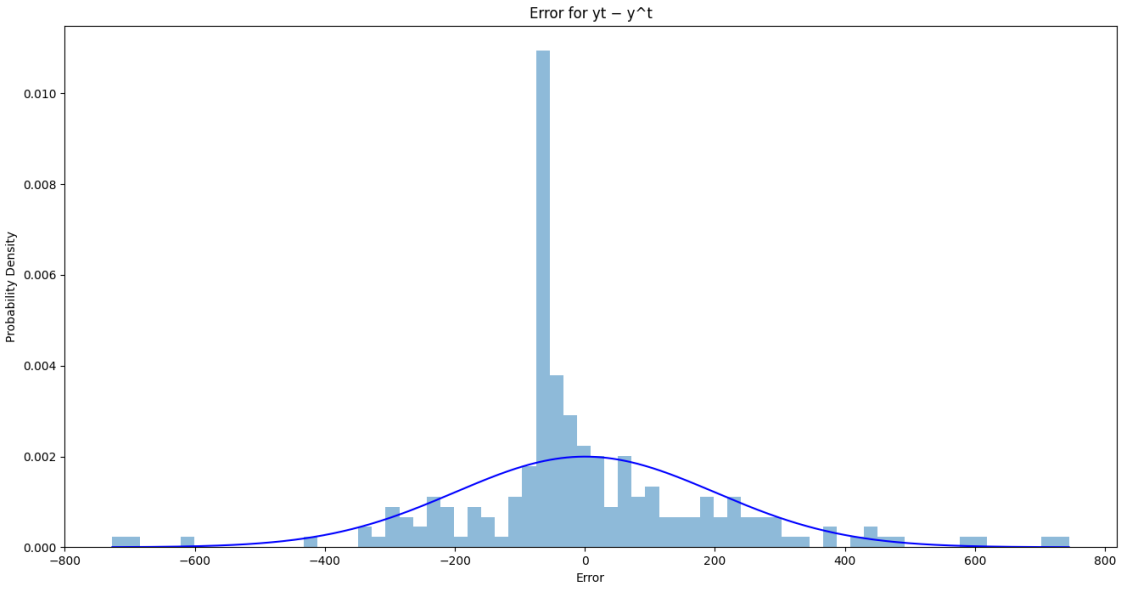

Figure 4: Error between actual death value and linear regression model predicted death value. Graph contains both the Gaussian model and the histogram of such. Both are normalized.