First Project - Probably Interesting Data

Anna Seib

March 28, 2021

1 Introduction

For this first project, we want to model two datasets. With the datasets from Kaggle, we formulate ideas on how machine learning can be used, and implement the algorithm to model the data. The two datasets I chose were Salary Data and Avocado Prices Data. For the Salary Data I chose Linear Regression, and for the Avocado Prices Data I chose Gaussian Distribution.

2 Design

2.1 Salary Data

For the salary data, there are two values: salary and the associated years of experience. For this type of data, I thought about doing a histogram, but chose Linear Regression after realizing there were not any "bins", as in numbers that could be classified together in a bin. I implemented the linear regression model by creating a line defined by coefficients estimated from trained data. I wrote Python code to model the distribution, and primarily used the libraries pandas, numpy, and matplotlib.
The calculated values include mean, variance, covariance, coefficients, and the root-mean-square-error, and these values are printed when the program runs. Finally, the values were plot to show the predicted values versus the actual dataset, using the formula for Linear Regression: (Salary = B0 + B1 * Experience, where B0 is the Salary when Experience is 0). In Figure 1, a small sample of the dataset values can be seen.

	YearsExperience	Salary
2	1.1	39343
3	1.3	46205
4	1.5	37731
5	2	43525
6	2.2	39891
7	2.9	56642
8	3	60150

Figure 1: Sample of Salary Data

2.2 Avocado Data

For the avocado dataset, I focused on the average price and number of avocados in a purchase. A sample of the dataset can be seen in Figure 2. I chose Gaussian Distribution because I assumed that most of the results would stay within one standard deviation. I used the empirical cumulative distribution function, since I have a dataset of discrete values. To implement the eCDF, I sorted the array values and found their frequency. The results are seen in Figure 4.

AveragePrice	Total Volume
1.33	64236.62
1.35	54876.98
0.93	118220.22
1.08	78992.15
1.28	51039.6
1.26	55979.78
0.99	83453.76
0.98	109428.33
1.02	99811.42
1.07	74338.76
1.12	84843.44
1.28	64489.17
1.31	61007.1
0.99	106803.39

Figure 2: Sample of Avocado Prices Dataset

2.3 Results

For the Salary Dataset using Linear Regression, the results can be seen in Figure 3. The scattered red dots represent the actual dataset, while the blue line is the predictions.

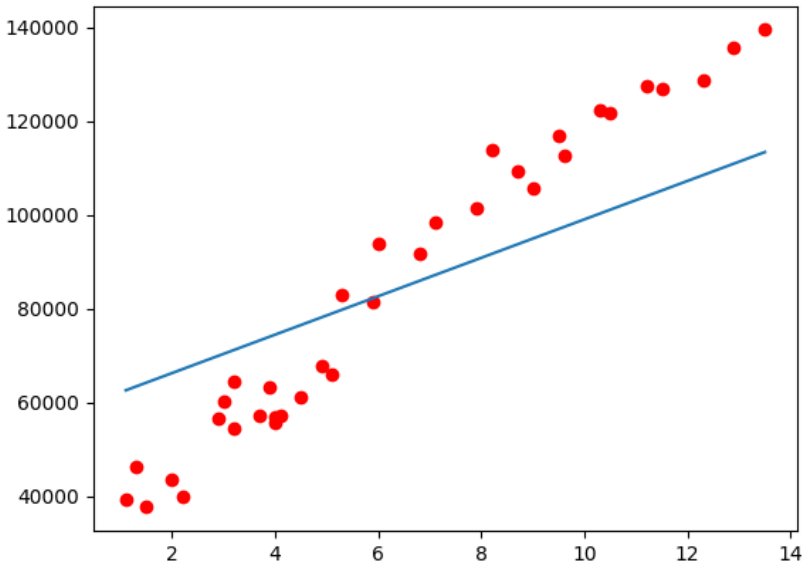

Figure 3: Linear Regression on Salary Data

For the Avocado Dataset using Gaussian Distribution, one of the benefits is being able to adjust the value of the mean and standard deviation to see how the curve changes. I experimented with the mean, and as an example, in Figure 5 the mean is set to 2, versus the calculated 1.40 for Figure 4.

My last analysis was to create the CDF for the Avocado Dataset, which represents the distribution of the values. This can be seen in Figure 6.

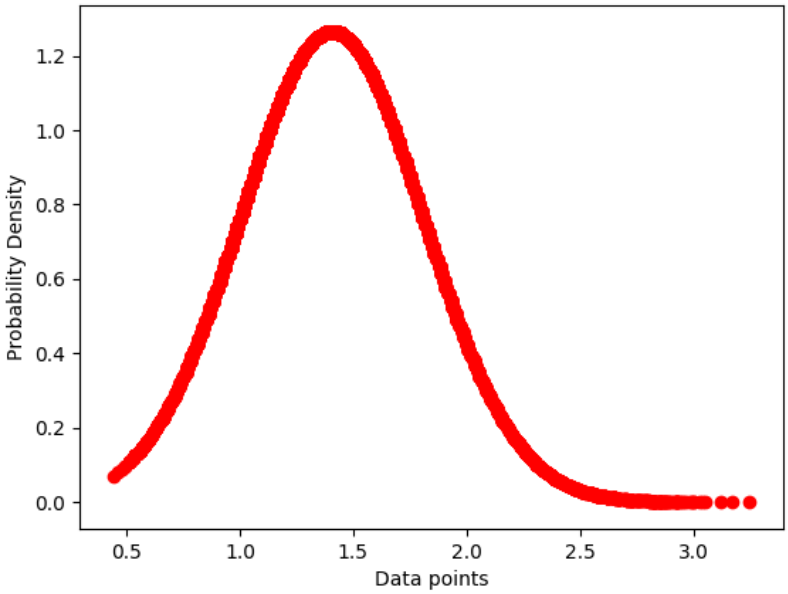

Figure 4: DFT of Avocado Dataset

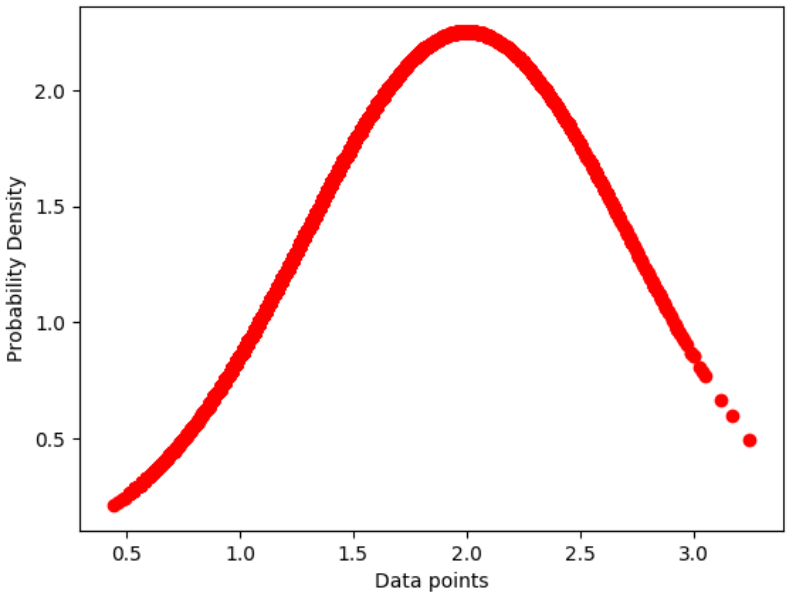

Figure 5: DFT of Avocado Dataset with Mean Adjusted to Two

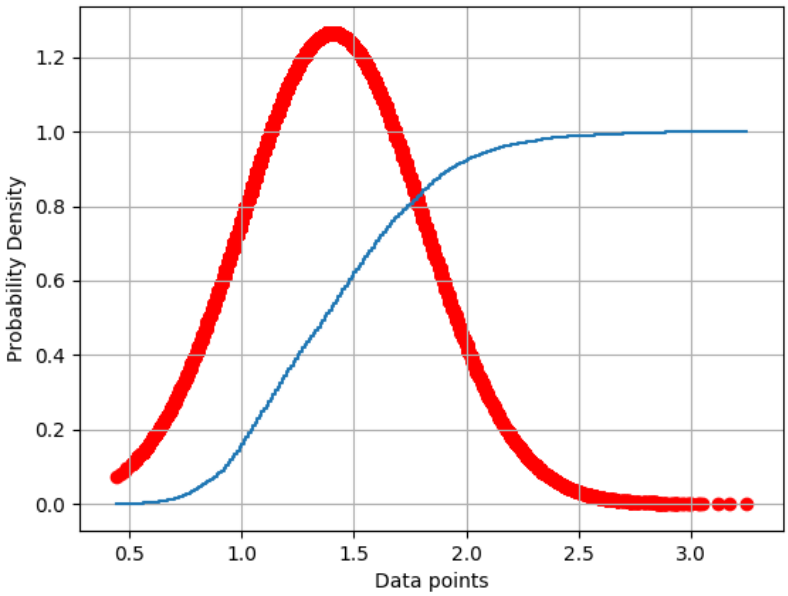

Figure 6: CDF of Avocado Dataset