Written by: Prof. Wu.
DS4420: Machine Learning II

Coding Midterm, Total: 1000 points

Question 1. (200pts)

- According to Forbs, the average income of a young adult between the ages of 20-24 is about $38,324 a year.
- According to GObankRantes.com the average livable wage in the US is $57,200.
- According to ZipRecruiter the starting salary of an **entry level data scientist** is $167,175.

Load the following file of salaries with the equivalent mean distribution of 40,000 entry-level data scientists.

`entry_level_DS_salary.csv.`

Load the file that consists of 40,000 individuals

`young_adult_income.csv`

with the equivalent mean and a standard deviation of $2,000.

Note that the incomes are in the units of thousands, so 1.2 = 1.2k.

1) Identify the distributions $p(x)$ and $p(y)$ that describe the college graduate income and entry-level DS salary. Plot the histogram for both datasets along with $p(x)$ and $p(y)$ over it **on the same plot**, showing that they are a good match.
2) Use the **counting method** on $p(x)$ to identify the probability that you will make a livable wage right out of college (the probability you won't need to move back with your parents).
3) Use the **integration method** on $p(x)$ to calculate the probability that you will make a livable wage right out of college.
4) If you are a really good data scientist and is able to get a DS job out of college. Use **the integration method** to identify the probability that you will make a livable wage and won't have to move back with your parents.
5) The top 10% earner of any age is $167,000. Given that you got a DS job right out of college, what's the probability that you will immediately (right out of college) become the top 10% income earner in the country?
6) Given this analysis, how does it conditionally change the probability that you will work harder in Prof. Wu's class?

Professor's Note

- This example is meant to demonstrate the power of having the ability to predict the likelihood of the future. It allows you to realize how your own actions right now can impact where you end up.
- The future is not written even if you have the probability distribution. Instead, it is completely based on the conditional probability of your choices and actions. Notice how given you got an entry-level DS job, the distribution looked completely different. The idea of conditional distribution is what gives you the power to change the distribution of your own future.
- Data science and Machine learning is the science of prediction. This is an incredible power.
- However, the path to obtaining this power requires an enormous amount of dedication. This is not a power that can be simply given to you. It is a long and difficult journey that must be taken individually. There are no shortcuts.

Question 2. (200pts) The following data consists of data collected from a population of 500 individuals on their habits and where they live. The label predicts the increased or decreased likelihood of them getting lung cancer due to their habits and environments.

```
lung_cancer_data.csv
lung_cancer_label.csv
```

1) Split the data into train, validation, and test.
2) Center and Scale the data.
3) Use gradient descent to train both linear and 2nd-order polynomial models.
    - Show the plot of MSE going down as you step through gradient descent.
    - Print out the final MSE for both models on training, validation, and test.
    - Which model would you pick?
    - Print out the first 20 entries for your predictions vs the truth using the test data. Show you that you did a good job predicting.

Question 3. (200 pts)

- Load `life_expectancy_X.csv`. Assuming that each row is a single sample. The data consists of 1000 individuals and 7 dimensions. The label is age at the point of death. You can interpret the 7 dimensions as

  1) Exercise amount
  2) Amount of supportive relationships
  3) Number of siblings
  4) Alcohol / Drugs / Smoking consumption
  5) Height
  6) Attractiveness
  7) work ethics

- By using these 7 factors, your goal is to predict how long a person will live.

  (I made up the data, so it has no relation to reality, but you would be doing the same analysis if the data was real.)
- Load `life_expectancy_y.csv` as the corresponding label y.
- Remember to use a bias term and assume linear regression.
- Scale the data between 0 and 1, and NOT center and scale. We do this to make sure all x_i values are positive. This allows for an interpretable result by looking at the weights from $w^\top x$. If the corresponding weight is positive, it tells you that the feature is positively contribution to the final result, but if it is negative, it would be negatively influencing the result.

  1) Use your own gradient descent algorithm to perform regression with the equation

     $$\min_w \sum_i^n (w^\top \varphi(x_i) - y_i)^2 - \lambda ||w||_1$$

     - This is a slightly modified regression called LASSO. We are still performing regression, but it allows you to identify the most important features. You would have to find the gradient of the L1 norm too.
     - You are not required to know why we added the L1 norm, but if you are curious, you can find out here:

       `https://youtu.be/bEXXB9G_VRk?si=vGm7qNiQUxogfxOv`

  2) According to LASSO, which factors most positively influence longevity? Which factors negatively impact longevity?
  3) For each feature of A calculate the correlation of each feature to the label and generate a table of correlations. Which features are most positively and negatively correlated to the label?
  4) What are the most important features according to PCA?
  5) Do the most correlated features also match
     - The most important features of PCA?
     - The most important features based on LASSO?
     - Putting the results from PCA, regression, and correlation together, what would you say are the most positive and negative factors that influence life expectancy?

Question 4. (200pts) This file consists of the number of dates 5000 people got last year.

`num_of_dates_a_year_5k.csv`

Part I

1) Use KDE to model this distribution and call it $p_1(x)$.
2) Plot $p_1(x)$ on top of the histogram, showing that they are a good match.
3) Use the numpy integration on $p_1(x)$ to find the probability of getting a value greater than 2.
4) Pretend you are a social scientist, by looking at the plot, come up with a theory why the plot is the way it is.

Part II

1) Randomly pick 2000 samples out of 5000 and save this as a new dataset file

   `num_of_dates_a_year_2k.`

2) Use KDE to model this distribution and call it $p_2(x)$.
3) Plot $p_2(x)$ on top of the histogram, showing that they are a good match.
4) Use the numpy integration on $p_2(x)$ to find the probability of getting a value greater than 2.

Professor's Note

- This example demonstrates that you don't always need the entire dataset to approximate the data distribution.
- Often, if you have a very large dataset and the computer cannot handle it, you can just take a smaller subset and arrive at a similar result.
- A natural question arises, how do you know that you have enough data?
  - Obviously, if you have 1 sample only, that's too few.
  - Obviously, if you have all the samples, then that's too much.
  - But then, when do you have the right amount of samples so that you can use the subset as the replacement?
- You can achieve this if you know how to measure if $p_2(x)$ is very similar to $p_1(x)$.
- The 2 convenient ways to measure the difference between distributions are ”KL divergence” and ”MMD”. Hopefully, you'll get to learn this.

Question 5. (200pts) Given the following integral and its result.

$$\int_0^2 x^2 - 3x + 4 \; dx = \frac{14}{3}$$

1) Use numpy automatic integration and print out your integration result (It should be 14/3).

2) Treat $f(x) = x^2 - 3x + 4$, use numpy sampling from a uniform distribution to approximate the integral.
   - Hint: Treat $p(x)$ simply as a uniform distribution within the integral

$$\int_0^2 f(x)p(x) \; dx$$

3) - (Bonus 20pts) instead of generating uniform distribution samples directly from python, use rejection sampling to generate uniform distribution samples to solve the integral.

4) - (Bonus 20pts) Use importance sampling to solve the integral.