|  |  |
| --- | --- |
| [Machine Learning]  [2024-1] |  |
| Homework 1 |  |
| [Due Date] 2024.04.04  Student ID :  Name :  Professor : Juntae Kim | logo-placeholder |

1. Write python codes to solve each of the following problem, and attach the result and description. (20 pts)

1-1. Numpy:

, ,

Compute where mean of each column of X

Compute where

|  |
| --- |
| Code |
|  |
| Result(Captured images) |
|  |

1-2. Pandas: Read data from From boston.csv (Boston Housing Price dataset), make a dataframe by selecting data with CRIM values < 1.0. Then from this data, compute “MEDV” column’s mean, and show the distribution of “MEDV” using a Histogram.

|  |
| --- |
| Code |
|  |
| Result(Captured images) |
|  |

1-3. Matplotlib : For plot the graph of in red color. The noise is normal distribution random value with mean 0, standard deviation 0.1

|  |
| --- |
| Code |
|  |
| Result(Captured images) |
|  |

1-4. Scikit-learn: Following is a dataset for student score vs. study hours and sleep ours. Use scikit-learn LinearRegression to learn a prediction model, and predict the score of a student with study hours = 2 and sleep ours = 2.

|  |  |  |
| --- | --- | --- |
| Study hours | Sleep hours | Score |
| 7.0 | 9.0 | 91 |
| 3.5 | 4.0 | 54 |
| 8.5 | 2.5 | 77 |
| 1.0 | 9.5 | 26 |
| 5.0 | 5.0 | 65 |

|  |
| --- |
| Code |
|  |
| Result(Captured images) |
|  |

2. Explain what Supervised Learning, Unsupervised Learning, and Reinforcement Learning are, and describe the differences. (10 pts)

|  |
| --- |
| Your Answer |
|  |

3. Describe the concept of “overfitting”, and explain how you can prevent overfitting in supervised learning. (10 pts)

|  |
| --- |
| Your Answer |
|  |

4. is m x n matrix, is m x 1 vector, is n x 1 vector. .

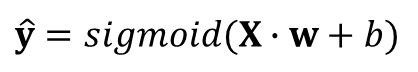
Let

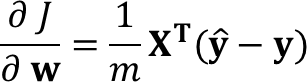
Show that (10 pts)

|  |
| --- |
| Your Answer |
|  |

5. For

Compute followings by hand: (20 pts)





|  |
| --- |
| Your Answer |
|  |

6. The heart\_disease.csv dataset represents 13 attributes of a patient and the presence of heart disease. Meaning of attributes are as below. The ‘num’ is the target value, 0 means no disease, 1~4 means different types of disease.

* age: age in years
* sex: sex (1 = male; 0 = female)
* cp: chest pain type
* trestbps: resting blood pressure (in mm Hg on admission to the hospital)
* chol: serum cholestoral in mg/dl
* fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
* restecg: resting electrocardiographic results

(0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)

* thalach: maximum heart rate achieved
* exang: exercise induced angina (1 = yes; 0 = no)
* oldpeak = ST depression induced by exercise relative to rest
* slope: the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
* ca: number of major vessels (0-3) colored by flourosopy
* thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
* num: diagnosis of heart disease

Change the dataset for binary classification (change 1~4 values of ‘num’ to 1), then perform logistic regression and show 1) the cost function graph, 2) learned model, 3) training accuracy of the model, 4) prediction result for the patient with attribute values of [61, 0, 3, 154, 210, 1, 0, 130, 0, 1.5, 2, 2, 3].

Do NOT use scikit learn library. (30 pts)

|  |
| --- |
| Code |
|  |
| Result(Captured images) |
|  |

**Note**

1. Submit the file to e-class as pdf

2. Specify your pdf file name as “hw1\_<StudentID>\_<Name>.pdf”

Ex) hw1\_2000123456\_홍길동.pdf