CS529 Security Analytics Fall 2023

Homework 1

Due: **September 22, 2023, 11:59 PM**

Descriptions & Instructions:

The goal of this homework is to help you become familiar with the implementation of linear regression, logistic regression and K-means. You are provided with the Jupyter Notebook, **HW1.ipynb,** and you are required to complete the coding corresponding to the following questions. *We have provided the code for loading the datasets. You just need to run the corresponding code to download it.* After completing the code segments, please ensure that the entire code in the notebook executes without any errors.

Submit format:

You will compress your Jupyter Notebooks source code into a **zip file** called HW1.zip and submit them on **Brightspace**.

Then you need to submit a report including your solutions to the coding problems. *You can choose to convert your Jupyter Notebook to a pdf report or take screenshots of your code and results*. The report should be submitted on **Gradescope**. The report should clearly rephrase each question, followed by the required answer/figures for that question.

Please mark your solutions to each question correctly while submitting the report on Gradescope.

Failure to follow the instructions will lead to a deduction in points!

Assignment 1: Linear Regression [30 pt]

Task: You are provided with the "diabetes" dataset, which contains medical information about diabetes patients and their corresponding diabetes progression measurements. We have provided the code to load the dataset using the load_diabetes() function from the sklearn.datasets module. Your task is to build a linear regression model to predict diabetes progression based on the provided features.

To accomplish this assignment, please complete the following questions in HW1.ipynb:

1. Split the dataset into training and testing sets by using 80% for training and 20% for testing. [5 pt]
2. Build a linear regression model using the training data. [7 pt]
3. Fit the training data to the build model. [7 pt]
4. Evaluate the model's performance on the testing data and report the Mean Squared Error. You will need to submit a number which corresponds to the testing MSE [6 pt]
5. Visualize predicted values against the actual values for the testing data using a scatter plot. You will need to submit a figure, where x-axis is the predicted value and the y-axis is the true value [5 pt]

What should be included in the PDF report:

6. Fill in the missing code in the Jupyter notebook and add proper comments to indicate the above steps
7. Use the code in the Jupyter notebook to print the testing MSE and draw the scatter figure
8. Covert the Jupyter notebook wit the code, comments, MSE, and figure, into a PDF

Assignment 2: Logistic Regression [40 pt]

Task: You will be implementing the Logistic Regression algorithm in Python using the Breast Cancer Wisconsin dataset. This dataset contains features extracted from digitized images of breast cancer biopsies and is used to predict whether a tumor is malignant (class 1) or benign (class 0). *We have provided the code to load the dataset using the load_breast_cancer() function from the sklearn.datasets module*. Logistic Regression is a widely used algorithm for binary classification tasks.

To accomplish this assignment, please complete the following questions in HW1.ipynb:

1. Implement the Logistic Regression algorithm, which mainly consists of the following steps: [15 pt]
    a. Initialize weights and bias to zeros.
    b. Update weights and bias using gradient descent.
    c. Use the sigmoid function to compute predicted probabilities.
2. Train the model on the normalized feature data (X) and target labels (y). [5 pt]
3. Implement a function to predict binary class labels (0 or 1) using the trained Logistic Regression model. [10 pt]

4. Evaluate the model's performance using precision, recall, and F1-score metrics. You can use precision_score, recall_score, and f1_score from the sklearn.metrics module. You will need to submit three numbers corresponding to the three metrics [10 pt]

What should be included in the PDF report:

1. Fill in the missing code in the Jupyter notebook and add proper comments to indicate the above steps.
2. Use the code in the Jupyter notebook to print the testing precision_score, recall_score, and f1_score.
3. Covert the Jupyter notebook wit the code, comments, and the number.

Assignment 3: K-Means [30 pt]

Task: You will be implementing the K-Means clustering algorithm in Python using the Iris dataset. It contains information about iris flowers with different features like sepal length, sepal width, petal length, and petal width. We have provided the code to load the dataset using the load_iris() function from the sklearn.datasets module. K-Means is an unsupervised machine learning algorithm used for clustering data points into groups.

To accomplish this assignment, please complete the following questions in HW1.ipynb:

1. Implement the K-Means algorithm with the following steps: [10 pt]
   a. Assign each data point to the nearest cluster centroid.
   b. Recalculate the cluster centroids as the mean of the data points assigned to each cluster.
   c. Repeat the above two steps until convergence or a maximum number of iterations is reached.
2. Apply your implemented K-Means algorithm to cluster the Iris dataset. [10 pt]
3. Create scatter plots to visualize the color-code the data points based on their assigned clusters, in the same figure, also visualize the cluster centroids. [10 pt]

What should be included in the PDF report:

1. Fill in the missing code in the Jupyter notebook and add proper comments to indicate the above steps.
2. Based on the code in the Jupyter notebook to plot the figure.
3. Covert the Jupyter notebook wit the code, comments, and the figure.