Carnegie Mellon University

Introduction to Machine Learning

Multiple Choice Questions & Project Work: Week 2

Due date: 7/10/23 @ 11:59 P.M.

INDIVIDUAL WORK:
Total Points: 100 (Refer to Chapter 10 of the Prescribed Textbook)

Task 1: Data Exploration (10 Marks)

You are provided with a dataset containing information about **housing prices**. The dataset includes features like the size of the house, number of bedrooms, location, and other relevant attributes. Perform the following tasks:

1. Load and explore the dataset to understand its structure and features.
2. Check for missing values and handle them appropriately.
3. Visualize the distribution of the target variable (house prices) and any other relevant features.

Task 2: Data Preprocessing (10 Marks)

Before training the linear regression model, it's essential to preprocess the data. Complete the following steps:

1. Split the dataset into the feature matrix (X) and the target vector (y).
2. Standardize the numerical features to have a mean of 0 and a standard deviation of 1.
3. Encode any categorical features using one-hot encoding or label encoding as required.

Task 3: Model Training (10 Marks)

Train a Linear Regression model using the pre-processed data from Task 2. Divide the dataset into a training set and a test set (80% training, 20% testing). Fit the model on the training set and then make predictions on the test set. Calculate the following metrics:

1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R-squared (Coefficient of Determination)

Task 4: Model Tuning (10 Marks)

In this task, you will experiment with different hyperparameters to improve the model's performance. Hint: Create polynomial features and assess how it affects the model's performance.

Task 5: Conclusion and Reflection (10 Marks)

Write a brief conclusion up to a maximum of one page in size summarizing the

1. reason why the prediction of median home values using linear regression is important,
2. variables that you believe are relevant to the prediction and how did you determine the same,
3. overall performance of the models and which hyperparameter tuning techniques were most effective, and
4. challenges faced, if any.

Note:

- Marks for each task may vary based on the completeness of the solution and the quality of analysis.
- Proper documentation, clear explanations, and appropriate code comments are essential for a higher score.
- Plagiarism will result in zero marks. Ensure your work is original.
- The assignment can be completed using Python and libraries like NumPy, Pandas, Scikit-learn, and Matplotlib.
- Refer to Chapter 10 of the Prescribed Textbook as noted at the end of class on Friday 6/30/2023.

Good luck with your assignment!