Assignment 2

CS6140 Machine Learning

Instructions:

1. **You are required to implement the algorithm from scratch. However, you can use output of standard APIs (sk-learn, scipy etc) to compare your solutions.**
2. **You may use libraries for visualizations and basic operations such as NumPy for matrix operations and scikit-learn or Pandas for data handling.**
3. **Submit a separate file for each task.** It may contain an image of your handwritten derivations, combined with code and outputs. Generously annotate your code with comments and explanations to convey your understanding and work.
4. Name the file as ML_Assignment_2YourName_YourID.pdf
5. Do not zip the files.
6. Do not upload/share your solution to internet. Violation of this rule will lead to zero grade.

Task 1: Regression Task

Dataset: Boston Housing Dataset

You can download the dataset from Kaggle.

Steps:

1. **Data Preprocessing**:
    - Load and clean the data.
    - Normalize the features if necessary. Apply appropriate transformations (i.e. OHE)
2. **Implement Ordinary Least Squares (OLS) Regression**:
    - Derive the OLS normal equation and Implement the **fit** and **predict** functions..
    - Train the model on the training set.
    - Evaluate the model on the test set.
3. **Implement Ridge and Lasso Regression**:
    - Implement the **fit** and **predict** functions.
    - Train the model on the training set with different values (0.5, 1, 1.5, 2) of the regularization parameter (λ).
    - Evaluate the model on the test set.
4. **Comparison and Analysis**:
    - Compare your solutions with standard APIs in terms of model parameters and appropriate metrics (e.g., Mean Squared Error, R^2 score).
    - Analyze the effect of the normalization, transformation and regularization parameter for Ridge and Lasso Regression.
    - Provide reasoning for difference in solution.

Task 2: Classification Task

Dataset: Breast Cancer Wisconsin (Diagnostic)

You can download the dataset from UCI Machine Learning Repository.

Steps:

1. **Data Preprocessing**:
   - Load and clean the data.
   - Normalize the features if necessary.
   - Apply appropriate preprocessing suitable for classification problem.
   - Split the data into training and test sets.
2. **Implement Gaussian NaiveBayes (GNB) and Gaussian Discriminant Analysis (GDA)**:
   - Use shared co-variance as well as class specific co-variance for GDA
   - Implement the **fit** and **predict** functions.
   - Train the model on the training set.
   - Evaluate the model on the test set.
3. **Implement Logistic Regression**:
   - Derive the Logistic Regression equations and implement the **fit** function using gradient descent.
   - Train the model on the training set.
   - Evaluate the model on the test set.
4. **Implement Perceptron**:
   - Derive the Perceptron learning rule and implement the algorithm.
   - Train the model on the training set.
   - Evaluate the model on the test set.
5. **Comparison and Analysis**:
   - Compare your solutions with standard APIs in terms of model parameters and appropriate metrics (e.g., accuracy, precision, recall, F1 score).
   - Analyze the strengths and weaknesses of each algorithm.
   - Discuss linear separability. How did you check if your data is linearly separable.?

Submission:

- Both notebook **ipynb** and corresponding **pdf** files.