Homework 1: Linear Models CSCE 633

Instructions for homework submission

1. Complete two parts in this homework:
    - Math questions: Include your solution in LaTeX document. Show your work. Submission with embedded photos of handwritten work will not be graded.
    - Programming questions: Complete the given skeleton `Python` code. For questions requiring visualization or analysis, include your solution in the same LaTeX document.
2. Submit your work to `Gradescope` including:
    - A PDF document for written parts: `FirstName_LastName_report.pdf`. LaTeX source code is not required.
    - A completed `Python` code: `FirstName_LastName_code.py`. *(Note: You may only use standard library (`os`, `pickle`, `re`, etc.), `matplotlib`, and library explicitly defined in given skeleton code.)*
    - **There are two separate submission portals on `Gradescope`:** one for code and one for the report. Submitting your work to the wrong portal will result in a loss of marks.
    - Please assign your answer in PDF report to its corresponding question when submitting to Gradescope. Submitting your work without assigning correspoinding question will result in a loss of marks.
3. This is an individual assignment. While you may consult class mates and other resources, your code should be your own and your write up your own. Please cite any such external help in your write up.
4. Start early!
5. Total: 100 points.

Math Questions (16 points)

NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

Problem 1: Gradient Calculation (8 points)

In this question you are required to calculate gradients for 2 scalar functions.

(1) Calculate the gradient of the function $f(x, y) = x^2 + \ln(y) + xy + y^3$. What is the gradient value for $(x, y) = (10, -10)$?

(2) Calculate the gradient of the function $f(x, y, z) = \tanh(x^3y^3) + \sin(z^2)$. What is the gradient value for $(x, y, z) = (-1, 0, \pi/2)$?

Problem 2: Matrix Multiplication (8 points)

In this question you are required to perform matrix multiplication.

(1)

$$\begin{bmatrix} 10 \\ -5 \\ 2 \\ 8 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 & 1 \end{bmatrix} = ?$$

(2)

$$\begin{bmatrix} 1 & -1 & 6 & 7 \\ 9 & 0 & 8 & 1 \\ -8 & 1 & 2 & 3 \\ 10 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 2 & 0 \\ 0 & -1 & 1 \\ -3 & 0 & 4 \\ 3 & 4 & 7 \end{bmatrix} = ?$$

Programming Questions (84 points)

The goal of this section is to create two regression-based models to assess air quality. The pre-split data for this homework can be found on Canvas (`data_train.csv`, `data_test.csv`). For detailed coding instructions, please refer to the provided skeleton code `hw1_ske.py`.

Remember, the test data only contain the features we use for the training, but not the labels. For each row in the test data, you need to use the trained model to predict the corresponding labels. Each row of the data corresponds to a sample, and the columns include the following information:

1. NMHC(GT): hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3
2. C6H6(GT): hourly averaged Benzene concentration in microg/m^3
3. PT08.S2(NMHC): hourly averaged sensor response to NMHC
4. NOx(GT): hourly averaged NOx concentration in ppb
5. PT08.S3(NOx): hourly averaged sensor response for NOx
6. NO2(GT): hourly averaged NO2 concentration in microg/m^3
7. PT08.S4(NO2): hourly averaged sensor response for NO2
8. PT08.S5(O3): hourly averaged sensor response for O3
9. T: temperature in C
10. RH: relative humidity
11. AH: absolute humidity
12. **PT08.S1(CO): TARGET VARIABLE - hourly averaged sensor response for CO**

Note on Submission

Please note that autograder will not call nor grade your `main()` function. Your `main()` function should be used for your local testing only.

(1) Data Processing (4 points)

For this question, complete the `DataProcessor` class. You can add any additional data processing technique as you see fit, as long as the shape of data is not changed.

1. Load the data to `DataProcessor` class using `pandas` library and use `read_csv` function. To inspect your data is correctly loaded, you may use `.head` and `.shape` method for `pandas.DataFrame` class.
2. Does the data have any missing values? How many are missing? Return the number of missing values. (In `pandas`, check out `isnull()` and `isnull().sum()`)
3. Drop all the rows with any missing data. (In `pandas`, check out `dropna()`. `dropna()` accepts an argument `inplace`, check out what it does and when it comes in handy.)
4. Extract the features and the label. The label is `PT08.S1(CO)`.

(2) Exploratory Data Analysis (10 points)

For this question, include your graph and analysis in your LaTeX report.

1. Plot the histograms of all the features in the data. Do all the features have a normal distribution? Do you see any outlier values? Do you need to apply any normalization technique to these values? If so, you can transform your data in `DataProcessor` and explain your thought process in the LaTeX report.
2. Pick 2 features and create a scatter plot to illustrate the correlation between these two features. Is there a high correlation between these features?
3. Compute the Pearson's correlation between all pairs of variables 1-12. Assign the resulting correlation values in a 12x12 matrix C, whose (i; j) element represents the correlation value between variables i and j, i.e., C(i; j) = corr(i; j). Visualize the resulting matrix C with a heatmap and discuss potential associations between the considered variables. Note: You can use the `heatmap` function from `seaborn`.

(3) Linear Regression Implementation (20 points)

For this question, complete the `LinearRegression` class in skeleton code. Include your loss plot in your LaTeX report. You are going to implement a linear regression model **from scratch** to regress the target variable.

1. Initialize the model.
2. Specify the criterion for the model. For linear regression, use **MSE (Mean Squared Error)**.
3. Construct main training loop, record loss, and plot the loss against iterations. The loss scale does not need to match the scale of the final predictions.
4. Make prediction using trained model.
5. Tune hyperparameters to achieve **RMSE (Root Mean Squared Error)** ≤ 71 on reserved test set.

Note: importing model directly from external library will result in compilation error in test environment.

(4) Logistic Regression Implementation (20 points)

For this question, complete the `LogisticRegression` class. Include your loss plot in your LaTeX report. Your are going to implement a logistic regression model **from scratch** for binary classification to predict the target label.

1. Using the column PT08.S1(CO), create a binary label for this dataset where the values more than 1000 correspond to label 1 and the values less than or equal to 1000 correspond to label 0.
2. Specify the criterion for the model. For logistec regression, use BCE (Binary cross entropy).
3. Construct main training loop, record loss, and plot the loss against iterations.
4. Make prediction using trained model.
5. Tune hyperparameters to achieve F1 score ≥ 0.90 and AUROC ≥ 0.90 reserved test set.

Note: importing model directly from external library will result in compilation error in test environment.

Note on Linear Model Implementation

A standard routine for implementing a machine learning model typically consists of the following steps:

- Initialize the model with learnable parameters and set tunable hyperparameters.
- Set up the training loop with defined stopping criteria.
- Compute gradients and update the model weights.
- Save the learned parameters and compute the training loss.
- Tune hyperparameters to achieve optimal performance.

(5) Result Analysis - Cross Validation (20 points)

Perform a 5-fold cross validation on both models. Complete the `ModelEvaluator` class in the skeleton code, and include your analysis in the report.

- For linear regression: Compute RMSE for each validation set across 5 folds. Report average and standard deviation of RMSE values.
- For logistic regression: Compute AUROC and F1 score for each validation set across 5 folds. Report the average and standard deviation of these metrics.

Do you see a big change across different folds? How can you use the coefficient of this model to find the most informative features?

(6) ROC Curve - Logistic Regression (10 points)

Plot the ROC curve for each fold of your logistic regression model and compute the area under the curve. Is this result consistent across the fold? Include your graphs and analysis in the LaTeX report.

Use of AI and External Resources

The purpose of this section is to cite the external resources you used in creating your homework. In particular, if you use any generative AI tools to help generate your code you should then explain:

1. What were the prompts you provided the AI tools?
2. What mistakes did you catch the AI making?
3. What did you do to validate the correctness of the AI generated portions, aside from relying on Gradescope Autograder scoring?