CSE343/CSE543/ECE363/ECE563: Machine Learning (PG)
Monsoon 2022

Assignment-1 (130 points)

Release Time: September 09, 2022; 7:00 pm Submission Time: September 21, 2022; 09:00 am

---

Instructions

- This assignment should be attempted individually. All questions are compulsory.
- **Theory.pdf**: For conceptual questions, either a typed or hand-written *.pdf* file of solutions is acceptable.
- **Code Files**: For programming questions, the use of any one programming language throughout this assignment is acceptable. However, the answer key will be provided in python only. For python, either *.ipynb* or *.py* file is acceptable. For other programming languages, submit the files accordingly. Make sure the submission is self-complete & replicable i.e., you are able to reproduce your results with the submitted files only.
- **Regarding Coding Exercises**: You can use modules from sklearn or statsmodels or any similar library for writing the code. Use random seed wherever applicable to retain reproducability.
- **Report.pdf**: Create a *.pdf* report of programming questions that contains your applied approach, preprocessing, assumptions, analysis, visualizations, etc.. Anything not in the report will not be evaluated. Alternatively, a well-documented *.ipynb* file with answers to all the questions may be submitted as a part of both code file and report.
- **File Submission**: Submit a *.zip* named A1_RollNo.zip (e.g., *A1_PhD22100.zip*) file containing *Theory.pdf*, *Report.pdf*, and Code files.
- **Submission Policy**: Turn-in your submission as early as possible to avoid late submissions. Expect **No Extensions**. Besides, submission within 10 min of the passing of the deadline will incur 20% penalty in the total marks of this assignment. Beyond this, late submissions will not be evaluated and hence will be awarded zero marks.
- **Clarifications**: Symbols have their usual meaning. Assume the missing information. Use Google Classroom for any queries. In order to keep it fair for all, no email queries will be entertained. You may attend office/TA hours for personal resolutions. No queries will be answered in Google Classroom comments after 9 pm on September 18, 2022.
- **Compliance**: The questions in this assignment are structured to meet the Course Outcomes CO1, CO2, and CO4, as described in the course directory.
- **Institute Plagiarism Policy Applicable.** Both programming and theoretical questions will be subjected to strict plagiarism check.
- There could be multiple ways to approach a question. Please explain your approach briefly in the report.

1. **Data Analysis and Visualization** **(50 points)**

   (a) Image Data

   (b) Tabular Data: Regression Problem

   (c) Tabular Data: Classification Problem

   (d) Time Series Data

   Refer to the file *'Data_visualisation.ipynb'* and add code to this starter code file.

   Link to *'Data_visualisation.ipynb'* and Datasets :
   https://drive.google.com/drive/folders/1SPmKvecB7M1wxmz7uYJQhWn8Lq7__Bik?usp=sharing

2. **Linear Regression** **(40 points)**

(a) **Pseudo-inverse**: Explain what is the pseudo-inverse of a matrix. (2 points)
Write the expression for pseudo-inverse to find solution to:

i. Under-determined system of equations (2 points)

ii. Over-determined system of equations (2 points)

(b) **Numerical problem on pseudo-inverse**: Solve the following system of linear equations:

$$x_1 + 3x_2 = 17$$

$$5x_1 + 7x_2 = 19$$

$$11x_1 + 13x_2 = 23$$

This is a pen and paper problem. Please explain the steps in detail. (8 points)

(c) i. Write the closed form expression (using normal equations) to solve a Linear Regression problem. (2 points)

ii. Why do we prefer iterative methods like Gradient descent rather than using closed form solutions to solve a Linear Regression problem. (3 points)

(d) **Coding Exercise**: dataset : https://archive.ics.uci.edu/ml/datasets/airfoil+self-noise

i. Visualize the data-set. (6 points)

ii.
- After the necessary data preparation, make a linear regression model to predict the target variable. (5 points)
- Briefly explain the following losses : RMSE, MSE, MAE. (3 points)
- Write a function from scratch to find any one of these loss functions. (3 points)
- Also check the value of this loss using sklearn library. (2 points)
- Report the accuracy and R^2 score of your model for both training and test data. (2 points)

3. **Classification/ Logistic Regression** **(40 points)**

(a) **Coding Exercise**: https://archive.ics.uci.edu/ml/datasets/Secondary+Mushroom+Dataset

i. Visualize the data set. (5 points)

ii. Impute the missing values. (3 points)

iii. Check correlation among the predictor variables and point out the redundant predictor variables if any. (5 points)

iv. Handle categorical variables using one-hot encoding or dummy encoding. (2 points)

(b) **Coding Exercise**: dataset: https://archive.ics.uci.edu/ml/datasets/glass+identification

i. Visualize the data-set. (4 points)

ii.
- After the necessary data preparation, make a logistic regression model to predict the target variable. (10 points)
- Report the accuracy and other metrics of the model (like precision, recall, F1 score). (4 points)
- Which metric do you think is more relevant here? Explain. (2 points)

(c) Derive an expression for gradient descent update rule for logistic regression using 'tanh' function as the decision boundary in place of 'sigmoid' function. (5 points)