CSE343/ECE343: Machine Learning
Assignment-1 Linear and Logistic Regression, ML in Practice, Empirical Risk Minimization
Max Marks: 25 (Programming: 15, Theory: 10) **Due Date**: 13/09/2024, 11:59 PM

Instructions

- Keep collaborations at high-level discussions. Copying/Plagiarism will be dealt with strictly.
- Late submission penalty: As per course policy.
- Your submission should be a single zip file **2020xxx_HW1.zip** (Where *2020xxx* is your roll number). Include **all the files (code and report with theory questions)** arranged with proper names. A single **.pdf report** explaining your codes with results, relevant graphs, visualization and solution to theory questions should be there. The structure of submission should follow:

  2020xxx_HW1
  |– code_rollno.py/.ipynb
  |– report_rollno.pdf
  |– (All other files for submission)

- Anything not in the report will **not** be graded.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours at least **two days before the deadline.**
- Your code should be neat and well-commented.
- **You have to do either Section B or C.**
- **Section A is mandatory.**

---

1. (10 points) Section A (Theoretical)

(a) (2 marks) You are developing a machine-learning model for a prediction task. As you increase the complexity of your model, for example, by adding more features or by including higher-order polynomial terms in a regression model, what is most likely to occur? Explain in terms of bias and variance with suitable graphs as applicable.

(b) (3 marks) You're working at a tech company that has developed an advanced email filtering system to ensure users' inboxes are free from spam while safeguarding legitimate messages. After the model has been trained, you are tasked with evaluating its performance on a validation dataset containing a mix of spam and legitimate emails. The results show that the model successfully identified 200 spam emails. However, 50 spam emails managed to slip through, being incorrectly classified as legitimate. Meanwhile, the system correctly recognised most of the legitimate emails, with 730 reaching the users' inboxes as intended. Unfortunately, the filter mistakenly flagged 20 legitimate emails as spam, wrongly diverting them to the spam folder. You are asked to assess the model by calculating an average of its overall classification performance across the different categories of emails.

(c) (3 marks) Consider the following data where y(units) is related to x(units) over a period of time: Find the equation of the regression line and, using the regression

x	y
3	15
6	30
10	55
15	85
18	100

Table 1: Table of x and y values

equation obtained, predict the value of y when x=12.

(d) (2 marks) Given a training dataset with features X and labels Y, let $\hat{f}(X)$ be the prediction of a model f and $L(\hat{f}(X), Y)$ be the loss function. Suppose you have two models, f_1 and f_2, and the empirical risk for f_1 is lower than that for f_2. Provide a toy example where model f_1 has a lower empirical risk on the training set but may not necessarily generalize better than model f_2.

2. (15 points) **Section B (Scratch Implementation)**
Implement Logistic Regression in the given dataset. You need to implement Gradient Descent from scratch, meaning you cannot use any libraries for training the model (You may use libraries like NumPy for other purposes, but not for training the model). Split the dataset into 70:15:15 (train: test: validation). The loss function to be used is Cross-entropy loss.
Dataset: **Heart Disease**

(a) (3 marks) Implement Logistic Regression using Batch Gradient Descent. Plot training loss vs. iteration, validation loss vs. iteration, training accuracy vs. iteration, and validation accuracy vs. iteration. Comment on the convergence of the model. Compare and analyze the plots.

(b) (2 marks) Investigate and compare the performance of the model with different feature scaling methods: Min-max scaling and No scaling. Plot the loss vs. iteration for each method and discuss the impact of feature scaling on model convergence.

(c) (2 marks) Calculate and present the confusion matrix for the validation set. Report precision, recall, F1 score, and ROC-AUC score for the model based on the validation set. Comment on how these metrics provide insight into the model's performance.

(d) (3 marks) Implement and compare the following optimisation algorithms: Stochastic Gradient Descent and Mini-Batch Gradient Descent (with varying batch sizes, at least 2). Plot and compare the loss vs. iteration and accuracy vs. iteration for each method. Discuss the trade-offs in terms of convergence speed and stability between these methods.

(e) (2 marks) Implement k-fold cross-validation (with k=5) to assess the robustness of your model. Report the average and standard deviation for accuracy, precision, recall, and F1 score across the folds. Discuss the stability and variance of the model's performance across different folds.

(f) (3 marks) Implement early stopping in your best Gradient Descent method to avoid overfitting. Define and use appropriate stopping criteria. Experiment with different learning rates and regularization techniques (L1 and L2). Plot and compare the performance with and without early stopping. Analyze the effect of early stopping on overfitting and generalization.

OR

3. (15 points) **Section C (Algorithm implementation using packages)**

Split the given dataset into 80:20 (train: test) and perform the following tasks:
Dataset: **Electricity Bill Dataset**

(a) (2.5 marks) Perform EDA by creating pair plots, box plots, violin plots, count plots for categorical features, and a correlation heatmap. Based on these visualizations, provide at least five insights on the dataset.

(b) (1 marks) Use the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce the data dimensions to 2 and plot the resulting data as a scatter plot. Comment on the separability and clustering of the data after dimensionality reduction.

(c) (2.5 marks) Perform the necessary pre-processing steps, including handling missing values and normalizing numerical features. For categorical features, use LabelEncoding. Apply Linear Regression on the preprocessed data. Report Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 score, Adjusted R2 score, and Mean Absolute Error (MAE) on the train and test data.

(d) (2 marks) Perform Recursive Feature Elimination (RFE) or Correlation analysis on the original dataset to select the 3 most important features. Train the regression model using the selected features. Compare the results (MSE, RMSE, R2 score, Adjusted R2 score, MAE) on the train and test dataset with the results obtained in part (c).

(e) (2 marks) Encode the categorical features of the original dataset using One-Hot Encoding and perform Ridge Regression on the preprocessed data. Report the evaluation metrics (MSE, RMSE, R2 score, Adjusted R2 score, MAE). Compare the results with those obtained in part (c).

(f) (2 marks) Perform Independent Component Analysis (ICA) on the one-hot encoded dataset and choose the appropriate number of components (try 4, 5, 6, and 8 components). Compare the results (MSE, RMSE, R2 score, Adjusted R2 score, MAE) on the train and test dataset.

(g) (1.5 marks) Use ElasticNet regularization (which combines L1 and L2) while training a linear model on the preprocessed dataset from part (c). Compare the evaluation metrics (MSE, RMSE, R2 score, Adjusted R2 score, MAE) on the test dataset for different values of the mixing parameter (alpha).

(h) (1.5 marks) Use the Gradient Boosting Regressor to perform regression on the preprocessed dataset from part (c). Report the evaluation metrics (MSE, RMSE, R2 score, Adjusted R2 score, MAE). Compare the results with those obtained in parts (c) and (g).