CSE343/ECE343: Machine Learning
Assignment-1 Linear and Logistic Regression, ML in Practice Max Marks: 25
(Programming: 15, Theory: 10) Due Date: 03/09/2023, 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) If two variables exhibit a strong correlation with a third variable, does this necessarily imply that they will also display a high degree of correlation with each other? Provide a reasoned explanation, supported by an illustrative example.

   (b) (2 marks) What is the defining criteria(s) for a mathematical function to be cate gorized as a logistic function? Briefly explain which of the following functions are valid logistic functions: sinh(x), cosh(x), tanh(x), and signum(x).

   (c) (2 marks) Which validation technique is beneficial for very sparse datasets and why? Briefly explain how it is different from the traditional K-Fold cross-validation technique.

   (d) (2 marks) Find the coefficients of the least square regression line for a set of ‘n’

data points (xi, yi) in slope-intercept form.

(e) (1 mark) The parameters to be estimated in the simple linear regression model $Y = \alpha+\beta x+\epsilon\ \epsilon\ N(0,\sigma)$ are:

(a) α, β, σ

(b) α, β, ϵ

(c) a, b, s

(d) ϵ, 0, σ

(f) (1 mark) In a study of the relationship between X=mean daily temperature for the month and Y=monthly charges on the electric bill, the following data was gathered: X=[20, 30, 50, 60, 80, 90], Y= [125, 110, 95, 90,110, 130]. Which of the following seems the most likely model?

(a) $Y= \alpha + \beta x + \epsilon\ \beta<0$

(b) $Y= \alpha + \beta x + \epsilon\ \beta>0$

(c) $Y= \alpha + \beta_1 x + \beta_2 x^2+\epsilon\ \beta_2 < 0$

(d) $Y= \alpha + \beta_1 x + \beta_2 x^2+\epsilon\ \beta_2 > 0$

2. (15 points) Section B (Scratch Implementation)

Logistic Regression

Implement Logistic Regression on the given Dataset. You need to implement Stochastic gradient descent from scratch i.e. You cannot use any libraries for training the model (You may use numpy, but libraries like SKLEARN are not allowed). Split the dataset in 70:20:10 (train:test:val). Loss function:- Cross-entropy loss

Dataset: Diabetes Healthcare Dataset

(a) (3 marks) Perform Logistic regression on the given dataset. Plot training loss v/s iteration, validation v/s iteration and training accuracy v/s iteration, validation accuracy v/s iteration. Comment on the convergence of the model. Compare and analyze the plots.

(b) (1 marks) Run your implementation for learning rates 1, 0.1, 0.01, 0.001. Comment and give your analysis.

(c) (2 marks) Make the confusion matrix and report the accuracy, precision, recall and F1 score for the above learning rates

(d) (4 marks) Modify your implementation by including L1 (LASSO) and L2 (Ridge Regression) regularization. Implement both regularization functions from scratch and train the model again. Try different values of the regularization parameter and report the best one. Plot similar loss v/s iteration and accuracy v/s iteration graph as before (train and val).

(e) (2 marks) Replace the sigmoid logistic function with tanh (tangent hyperbolic func tion) in your implementation and choose the appropriate loss function. Plot loss vs iteration and accuracy vs iteration. Comment on your observation.

(f) (3 marks) Modify your implementation in part 1 with Mini-Batch Gradient descent and plot all the graphs. Vary the batch size and analyze the plots. Comment on the convergence rate of SGD and Mini-BGD.

OR

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

Implementation of linear regression using libraries:- Split the dataset into 80:20 (train: test)

Dataset: CO2 Emissions Dataset

(a) (2 marks) Visualize the dataset. Make scatter plots, pair plots, box plots, and cor relation heatmap. Distribution plots, i.e. histograms, pie charts etc. for categorical features. Give at least five insights on the dataset.

(b) (2 marks) Use TSNE (t-distributed stochastic neighbour embedding) algorithm to reduce data dimensions to 2 and plot the resulting data as a scatter plot. Comment on the separability of the data.

(c) (2 marks) Perform the necessary preprocessing steps, and for categorical features use label-based encoding. Perform linear regression on the preprocessed data. Report MSE, RMSE, R2 score, Adjusted R2 score, MAE on the train and test data.

(d) (2 marks) Use Principal Component Analysis (PCA) on the original dataset to reduce the number of features and then train the model with the reduced feature dataset. Vary the number of components, i.e. 4, 6, 8, 10 and compare the results (MSE, RMSE, R2 score, Adjusted R2 score, MAE) on the train and test dataset.

(e) (2 marks) Encode the categorical features of the original dataset with one-hot en coding and perform all tasks of part c again, i.e. apply linear regression and report MSE, RMSE etc. Compare the results obtained with part c.

(f) (2 marks) Perform PCA on the one-hot encoded dataset and choose the appropriate number of components (try 5 different values). Compare the results (MSE, RMSE, R2 score, Adjusted R2 score, MAE) on the train and test dataset.

(g) (1.5 marks) Use L1 and L2 regularization while training the linear model (use the preprocessed dataset of part c). Compare the MSE, RMSE, R2 score, Adjusted R2 score, and MAE on the test dataset for both regularization techniques.

(h) (1.5 marks) Use SGDRegressor library to perform linear regression on the prepro cessed dataset of part c. Report the evaluation metrics and compare the results.