Lab Course Machine Learning

Exercise Sheet 3

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal
Submission deadline : November 25, 2022

General Instructions

1. Perform a data analysis, deal with missing values and any outliers.
2. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
3. Data should be normalized.
4. Train to Test split should be 80-20.
5. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.

1. Linear and Logistic Regression. (15 points)

A. Optimization Routines and Loss Functions

1. In this part of the assignment we learn how to write modular programs and make our code reusable. For this, declare a class named **Optimization** which has 2 inputs X and y as the class variables. Next, implement the following four optimization algorithms in this class.
   a) Stochastic Gradient Descent
   b) Newton's method
2. You will need loss functions and their gradients for the optimization process. Next, implement a class **Loss** which also takes in X and y and computes the following losses and their gradients.
   a) Mean Square Loss
   b) Cross Entropy Loss

   Make the **Loss** class such that you can access it from the **Optimization** class.

B. Linear Regression

In this task, you are given a data set named *regression.csv*. Implement a class **LinearRegression** that has at least two functions, **fit** , **predict** for fitting a linear regression model and predicting the results. You need to use the **Optimization** and **Loss** class inside this. Fit a linear regression model with the following pair:

1. Mean Square Loss and Stochastic Gradient Descent.

In the end, generate the loss trajectory for both training and testing.
While optimising the loss function for Linear Regression or Logistic Regression, one needs to initialise the model parameters. It is well known that deep neural networks do not function if the model parameters are initialised to zero. Why is it so? Does this issue also arise while optimising the loss function for Linear or Logistic Regression? Explain.

C. Logistic Regression

You are given a file *logistic.csv*. This part of the assignment involves a classification task. Implement a class **LogisticRegression** having at least two functions, **fit**, **predict** for fitting the model and getting the predictions. Fit a logistic regression model with the following pair:

1. Cross Entropy Loss and Newton's Method.

Report the test accuracy, plot the confusion matrix and also compute the precision, recall and F-score. Read about these parameters. Suppose model A and model B both have same accuracy, but model B has a higher F-score, which model would be suited? In the end, generate the loss trajectory for both training and testing.

2. Discriminant Analysis (5 points)

In this part of the assignment you will implement linear and quadratic discriminant analysis classifiers on the iris dataset **from scratch**. Again, this should follow an object oriented method of implementation where you need 2 classes **LDA()** and **QDA()** with the associated **fit()** and **predict()** methods. Read this dataset using data = load_iris() command from the sklearn library. Report the test accuracies for both LDA and QDA.