[CSC 5825 Fall 2022]

Homework 1

Due 11:59 pm, September 21, 2022

Full credit: 100 points

Question. Logistic Regression Implementation

In this question, you are asked to implement logistic regression from scratch to dealing with a classification task. The heart failure prediction dataset contains 11 common features (e.g., age, sex, and several medical predictor variables) from 918 observations that can be used to predict whether the patient has heart disease (1) or not (0). You can find more details of the dataset from Kaggle website `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`.

Tasks:

- Split the 918 instances into training and test sets (8:2). (10 points)
- Train your logistic regression classifier on the training set. (60 points)
- After training, test your classifier on the test set, report the confusion matrices (accuracy, precision, recall, and F1-score) as the result. (30 points)

Guidelines:

- Apply data pre-processing and feature engineering. The categorical features (i.e., *Sex, ExerciseAngina, ChestPainType, RestingECG, STSlope*) need to be converted into dummy/indicator or one-hot encoding variables.
- Write a function to calculate the output of sigmoid activation function for a given input x.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

- Implement a logistic regression algorithm based on three basic steps (forward-propagation, back-propagation and gradient descent).
- Randomly initialize the weights W and b.
- Initialize a learning rate and an iteration number.
- Fine-tune different learning rates and iteration numbers using cross-validation, e.g., 5-fold or 10-fold.

Submission Instructions

Homework must be submitted electronically through Canvas website on/before the due date/time. Homework must be typed with LaTeX or Word. The code can be submitted as .py file or .ipynb file. Late homeworks will not be accepted unless with legitimate excuses with documents. Do NOT use models in scikit-learn package directly in the homeworks.