Programming Assignment #1

Logistic Regression

DUE: Friday February 17th at 11:59 PM

Problem: Implement the **Logistic Regression** algorithm. The code skeleton as well as data sets for this assignment can be found on e-Learning.

Data Set: The data set (in the folder ./data/) is obtained from the UCI Repository and are collectively the MONK's Problem. These problems were the basis of a first international comparison of learning algorithms[1]. The training and test files for are named monks-3.train and monks-3.test. There are six attributes/features (columns 2–7 in the raw files), and the class labels (column 1). There are 2 classes. Refer to the file ./data/monks.names for more details.

a. (**Reporting Error Rates**, 30 points) For the following learning rates {0.01, 0.1, 0.33} and iterations {10, 100, 1000, 10000} fit Logistic Regression Models using the MONKS-3 train set and compute the average training and test errors. Report what the best parameters are for the model.

b. (**Saving the Model**, 50 points) Retrain the models on the best parameters and save it as a pickle file in the format 'NETID_lr.obj'. The object file will be loaded for grading and the class functions will be individually tested.

c. (**scikit-learn**, 10 points) For monks datasets, use scikit-learns's default Logistic Regression Algorithm[2]. Compute the train and test errors using sklearn's Logistic Regression Algorithm. Compare the results with the version you implemented and speculate on why there is a difference in performances between the two algorithms. **Do not change the default parameters**.

d. (**Plotting curves**, 10 points) For each of the learning rates, fit a Logistic Regression model that runs for 1000 iterations. Store the training and testing loss at every 100th iteration and create three plots with epoch number on the x-axis and loss on the *y*-axis.

.

[1] https://archive.ics.uci.edu/ml/datasets/MONK's+Problems
[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html