COS 482 & COS 598: Introduction to Data Science
Spring 2023

Homework Assignment 4
Assigned: April 24, 2023
Due: May 5, 2023

Submission instructions:

You must submit your Python code in a single .py file, and include a brief write-up about what you did in your code to complete the tasks in this assignment as well as any relevant information requested in the tasks.

The file(s) must be uploaded to the Brightspace submission site for Homework 4.

Total: 100 points.

Task 0. Download the Spambase Data Set from UCI Machine Learning Repository (0 point).

The Spambase Data Set is a dataset which can be used to train a machine learning model to predict whether a particular email is or is not a spam.

- Go to https://archive.ics.uci.edu/ml/datasets/Spambase.
- Click Data Folder.
- Download spambase.zip, and decompress it.
- The actual data are contained in spambase.data, which is essentially a CSV file with no heading. To understand what each column means, read spambase.DOCUMENTATION and spambase.names for more information (these are plain text files).

Task 1. Process the data (25 points).

Load and split the data into a training (80%) and a test (20%) set. Keep the proportion of the positive examples in the training and the test set approximately the same as that in the original (unsplit) dataset. You can do that by setting `stratify=y`, where `y` is the column of target labels (spam/not spam), when splitting the data using `model_selection.train_test_split` from Scikit-learn (`sklearn`).

Use min-max scaling to scale the features in the training set to values between 0 and 1, and then use the minimums and maximums of the columns *from the training set* to scale the features in the test set.

The end goal of this task is to have four NumPy arrays: `X_train`, `y_train` (for training) and `X_test`, `y_test` (for testing).

Task 2. Train machine learning models to predict whether an email is or is not a spam (25 points).

Train a linear support vector machine and a logistic regression model on the training data. Evaluate their performance on the test data.

- What is the test accuracy of the support vector machine and the logistic regression model?
- Examine the coefficients of the support vector machine. What do they tell you about which features are significant for the prediction of whether an email is a spam?
- Examine the coefficients of the logistic regression model. What do they tell you about which features are significant for the prediction of whether an email is a spam?

Task 3. Train a neural network to predict whether an email is or is not a spam (50 points).

Finally, train three different fully-connected neural networks (with your choices of hidden dimensions and the number of layers) on the training data. Evaluate their performance on the test data. How does the test accuracy of these neural network models compare with that of the support vector machine and the logistic regression model from Task 2?