[CSC 5825 Fall 2021]

Due. Before Class Sept. 22, 2021 Homework 1

Total Points: 100

Non-Programming Questions:

Question 1. (10 points) Describe one problem in your study area that can be solved by machine learning techniques. Classify your problem in terms of supervised or unsupervised, classification or regression? Explain the unique challenges to standard machine learning methods?

Question 2. (10 points) In equation 2.13 (textbook page 35), we summed up the squares of the differences between the actual value and the estimated value. This loss function is the one most frequently used for continuous output, but it is one of several possible loss functions. Because it sums up the squares of the differences, it is not robust to outliers. Propose a better error function to enable robust regression? Please define all mathematical notations clearly.

Programming Questions:

Question 3. (20 points) Practice using Numpy

In this question, you are asked to practice using Numpy which is commonly used in many machine learning tasks. Numpy supports a large number of dimensional array and matrix operations and provides a large number of mathematical function libraries for array operations. Here are some basic operations you will practice. You will find the Python quick starting guide useful. Please use "Numpy Practise_template.py" or "Numpy Practise_template.ipynb" as your template.

- Import numpy package.
- Create different kinds of arrays and initialize them, then print some specified elements in these arrays.
- Practice some array arithmetical operations in numpy, such as add, subtract, multiply, divide, square and dot.
- Practice some other operations in numpy, such as sum, mean, max, sin and cos functions.

Question 4. (20 points) Practice using Pandas

In this question, you are asked to practice using Pandas which is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool. It aims to be

the fundamental high-level building block for doing practical, real world data analysis in Python. Here are some basic operations you will practice. You will find the Python quick starting guide useful. Please use "Pandas Practise_template.py" or "Pandas Practise_template.ipynb" as your template.

- Import pandas package.
- Create and print Pandas series and dataframe.
- Practice selection methods in Pandas.
- Read a csv file using Pandas and plot data with plot package.

Question 5. (40 points) Logistic Regression

In this question, you are asked to employ logistic regression on a data set. The labels of this data set are 0 and 1. Some basic codes are already given, what you need to do is to complete the Python code based on the different procedures of logistic regression. Please use "Logistic Regression_template.py" or "Logistic Regression_template.ipynb" as your template.

- Read the data in the file Question5.txt, the first two columns as X and the last column as y, visualize the data in a scatter plot with different colors for each class (0 and 1).
- Write a function to calculate the output of sigmoid activation function for a given input t.
$$S(t) = \frac{1}{1 + e^{-t}} \tag{1}$$
- Implement a logistic regression algorithm based on three basic steps (forward-propagation, back-propagation and gradient descent).
- Randomly initialize the weights W and b and initialize a learning rate alpha.
- Initialize an iteration number and repeat the logistic regression algorithm several passes over the entire training data set until converge.
- Visualize the losses generated by the cost function in the training process for each iteration.
- Try different learning rates (alpha) and iteration numbers (iterations) and report the optimal ones.
- Use the final weights W and b to plot the line to separate points on the same plot in the first step.

Submission Instructions

Homework must be submitted electronically through Canvas website on/before the due date/time. Homework assignments are usually due in class at the beginning of lecture on the due date given. 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 functions in scikit-learn package directly in the homeworks.