COSC 4555/5555 Machine Learning, Spring 2024: Homework 5

Due: Thursday, March 28^{th}, 07:59:59 hrs

Instructions: Read all instructions in this section thoroughly.

Answer all of the following questions. If you have questions or confusion, reach out to the instructor or the TA. *Please show your work for all questions.* The primary purpose of this homework is to practice, understand, and compare and contrast linear and logistic regression.

Submission guideline: This homework submission will include two parts: a programming part, and a reflection part.

For the programming part, you will write your code in the python scripts, and complete the programming part of the assignment as per the instructions. [1]

For the reflection part, you will complete this part by answering it in the **written.pdf**.

You are welcome to use any Python libraries for data munging, visualization, and numerical linear algebra. Examples include Numpy, Pandas, and Matplotlib. You may NOT, however, use any Python machine learning libraries such as Scikit-Learn or TensorFlow except the ones clearly mentioned in the individual problem instructions. If in doubt, email the instructor or the TA.

Note: You will submit only the following files: the completed *written.pdf*, and your code in one or more python files. *Please name the code files appropriately.*

1 Programming Question [40 Points]

In this problem, you will implement Stochastic Gradient Descent (SGD) to optimize both a **logistic regression** model and a **linear regression** model to predict whether a given patient has diabetes or not. In clinical informatics, machine-learning approaches have been widely adopted to predict clinically adverse events based on patient data. If you are interested in applications of machine learning to biomedicine and healthcare, check the MIMIC II clinical database demo set: `https://archive.physionet.org/mimic2/demo/`

1.1 Dataset

For this problem set, we will use the Pima Indians Diabetes Data Set: `Pima-Indians-Diabetes.csv`.

All provided patients are females at least 21 years old of Pima Indian heritage. The data on each patient include:

- The output class variable (HasDiabetes: 0 - normal or 1 - diabetes)
- Number of times pregnant (num_preg)
- Plasma glucose concentration at 2 hours in an oral glucose tolerance test (PlGlcConc)

[1] If you are writing your code in any other programming language, you will create your file accordingly and use the file type according to your programming language. You will provide any written answers, analysis, figures, plots, etc. for the programming exercise in the **written.pdf** file.

- Diastolic blood pressure (BloodP)
- Triceps skin fold thickness (tricept, unit: mm)
- 2-Hour serum insulin (insulin, unit: μU/ml)
- Body mass index (BMI)
- Diabetes pedigree function (ped_func)
- Age (age, years)

1.2 Model Training

Among all 768 patients, we will separate 500 patients as training data and 268 patients as test data. Recall that SGD performs gradient descent using a noisy estimate of the full gradient based on just the current example.

(a) [5 points] Use min-max normalization[2] to normalize all the features based on the training dataset.

(b) [15 points] Write down the equation for the weight update step for both linear regression and logistic regression. That is, how should you update the weights $\mathbf{w}^t$ using the data point $(\mathbf{x}^t, y^t)$, where $\mathbf{x}^t = [x_1^t, x_2^t, ..., x_d^t]$ is the feature vector and $y^t \in \{0, 1\}$ is the label for example (i.e patient) in the dataset at t?

(c) For step sizes $\eta = \{0.8, 0.001, 0.00001\}$, and without regularization, implement SGD for both linear regression and logistic regression. Train the model by making one pass over the dataset. You may assume that the data is randomly ordered. Use only one pass over the data on all subsequent questions as well. Initialize the weight vector $\mathbf{w}$ and the bias w_0 to 0. For each step size:

(i) [5 points] Plot the average loss $\bar{L}$ as a function of the number of steps T, where

$$\bar{L}(T) = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}^t - y^t)^2, \tag{1}$$

and where $\hat{y}^t$ (either 0 or 1) is the predicted label for example $\mathbf{x}^t$ using the weights $\mathbf{w}^{t-1}$. Record the average loss every 100 steps, e.g. [100, 200, 300, ...].

(ii) [5 points] Report the l_2 norm of the weights at the end of the pass. In general, l_2 regularization can be useful if the norm of these weights grows very large over the course of model training. In this assignment, however, you are <u>not</u> expected to perform l_2 regularization.

(iii) [5 points] Use the model weights to predict whether each patient in the test set has diabetes, for every 100 steps. Report the SSE (sum of squared errors) of your prediction. Make sure to use the SSE for the 0/1 prediction.

(d) After 100,000 step:

(i) [2.5 points] Select the best logistic regression model based on training data, and report the weights of following features: BMI, 2-Hour serum insulin level, and Plasma glucose concentration.

(ii) [2.5 points] Provide an interpretation of the effect of the features above for diabetes classification based on these inferred weights.

[2]For more details about min-max normalization, refer to this link `http://sebastianraschka.com/Articles/2014_about_feature_scaling.html`

Reflection Section

1. How much time you took to complete this homework?
2. Write a self-reflection on your learning activities to complete this homework.
3. What different resources did you reach out to complete this homework?
4. How did completing this homework enhance your learning?