Due Date: 13/05/2022 22/04/2022

CENG 463 HOMEWORK II

In this homework, you are asked to implement a **two-layer neural network** with a sum-of-squares error (RSS) to solve a rating prediction problem on a **Drug Review Dataset**. In this network, there are linear activations (i.e., no non-linear transformation) in the output units and **sigmoid activations** in the hidden units. It means the problem is a regression analysis problem in essence. A **code template**[1] is provided to you and the dataset that you will use is available in **UCI Machine Learning Repository**[2].

The homework has 2 parts, one is the implementation, the other one is a report about your implementation. The implementation should keep the main, train, and test functions as they are (you can only modify the **[input_features]** and **[hidden_layers]** parts according to your own design). The whole design, input features, activation functions, loss function, size of hidden layers, etc. are up to you. As long as the model learns (**a minimum of %20 accuracy** is expected at least) you may change the design. However train, test, and main functions and the skeleton that is provided you should stay still.

Using any explicit library to shorten the operations on the network is prohibited (you can only use NumPy, pandas, matplotlib, and other default python libraries). Also, we want you to implement the code on Python 3 and not use absolute paths in your code, you can use relative paths like “./data/drugLibTest_raw.tsv”.

After the implementation we want you to carefully write a report about “What was the reason that your network learned?”, “How would you improve the accuracy more?”, “What were the trade-offs you face with?”, “And the reasoning of why did you design your network in that way” (Reasoning about things like activation function, loss function, number and size of hidden layers, learning rate, etc.). Also, the report should have a plot of loss and accuracy change through time and we want you to comment on the plots. (A detailed reasoning about the result and the shape of the graph). Please remember that report has a grading factor as much as implementation.

P.S.: Since you are almost an engineer, you are expected to write efficient, clean, well-commented codes.

P.P.S.: Submit only a **ZIP** file (NOT RAR) that contains a **PDF** (NOT DOC or DOCX) of your report and a python file which is the completed version of the given code template[1]. Before submitting rename the template file as **CENG463_2022_HW2_*YourStudentNumber*.py**

[1] https://drive.google.com/file/d/1FD-MZl4rPeUlnMwxX3xWKUoKMrAYmO53
[2] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29

GUIDELINES

- The Drug Review Dataset in UCI is already partitioned into the train (75%) and test (25%) sets. You can use the same partitions for training and testing.

- You should only consider the ratings and reviews (benefits review, side effects review, and comments review) although there are 8 attributes[3] in the dataset. You can use comments only or combine multiple types of reviews (the selection of the [input_ features] is up to you).

- You can use the following formula to calculate sigmoids:

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

$S(x)$ = sigmoid function

e = Euler's number

- You can use the following formula to calculate sum-of-squares error (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

---

[3] **Drug Review Dataset Attributes:**
1. urlDrugName (categorical): name of drug
2. condition (categorical): name of condition
3. benefitsReview (text): patient on benefits
4. sideEffectsReview (text): patient on side effects
5. commentsReview (text): overall patient comment
6. rating (numerical): 10 star patient rating
7. sideEffects (categorical): 5 step side effect rating
8. effectiveness (categorical): 5 step effectiveness rating