Homework 5

Problem 1

Your task is to find a linear approximation of the function

$$f(x) = \sqrt{1+x}, \quad x \in [0,1].$$

Your homework should contain the following steps:

a) Generate $N = 10000$ random numbers from $[0,1]$:

$$x_1, x_2, \ldots, x_N \in [0,1],$$

and then obtain their labels:

$$y_i = \sqrt{1+x_i}, \quad i = 1, 2, \ldots, N.$$

b) Perform linear regression on your generated data using the closed form solution.

c) Perform linear regression on your generated data using the library `sklearn`.

d) Perform linear regression on your generated data implementing the gradient descent algorithm by yourself.

e) Perform linear regression on your generated data using `pytorch`.

f) Perform linear regression on your generated data using mini-batch gradient descent (implement yourself).

g) Sketch the graphs of all approximations on one graph.

h) Compare all solutions with the first degree Taylor approximation of the function $\sqrt{1+x}$.

i) How will you define polynomial regression inspired from linear regression?

j) Can you implement polynomial regression using linear regression?

Problem 2

Your task is to predict whether a patient has heart disease or not. You can download and see information about data by following this link:

Heart Disease Dataset on Kaggle

You should do the following steps in your homework:

a) Keep 20 percent of the data for testing.

b) Perform logistic regression and try to find the best hyperparameters (using sklearn).

c) Normalize your data using standard normalization and then repeat the previous step.

d) Try to find the best neural net to solve this problem which will have no more than 2 hidden layers (use sklearn).

e) Normalize your data using standard normalization and then repeat the previous step.

f) Perform the step d) using `pytorch`.

g) Compare accuracies of all obtained models.

h) Retrain your best-performing model using PyTorch Lightning. Incorporate a logger to track your experiments and use a learning rate scheduler.

Remarks:

- You must achieve more than 85% accuracy on the test dataset in at least one of the steps above to get full grade.

Useful links:

- Heart Disease Dataset on Kaggle

- sklearn MLPClassifier documentation

- PyTorch Lightning official site