DESCRIPTION

I have to build a Machine Learning library that contains following algorithms- Linear (and Polynomial) Regression, Logistic Regression, KNN, K-Means Clustering and Neural Networks which are implemented from scratch using Python and following libraries- NumPy, Pandas and Matplotlib.

Various data sets that are provided to me will be passed into this library which will then give an output in the form of accuracy.

FEATURES

➢ Automatic Hyperparameter Optimization

Hyperparameter is a parameter whose value is used to control the learning process. Automatic Hyperparameter Optimization is the problem of automatically choosing a

set of optimal hyperparameters for a learning algorithm such that it yields an optimum model which minimizes a predefined loss function. This is done with no manual effort needed beyond the initial set-up.

➤ <u>Comparison of Different Models</u>

The same datasets will be inputted in each model. The outputs from these models will then be plotted and compared on the basis of different features like accuracy and precision. This will show how different models perform on different datasets.

TECHNOLOGY STACK

➤ Programming Language

- Python

➤ Libraries

- NumPy

- Pandas

- Matplotlib

IMPLEMENTATION DETAILS

I will be implementing the two features which have been mentioned above. Firstly,

➢ <u>Automatic Hyperparameter Optimization</u>

For automatically optimizing the hyperparameters, I will create and use 4 functions namely- Model Creation, Model Training, Output and Evaluation.

1. <u>Model Creation</u> - In this function, I will create the models of all the algorithms separately, the algorithms being Linear (and Polynomial) Regression, Logistic Regression, KNN, K-Means Clustering and Neural Networks. I will define the following functions for each of these algorithms-

   o Plotting the data

   o Cost Function (Regularized for complex datasets)

   o Visualizing Cost Function (in case of Linear Regression)

   o Gradient Checking (in case of Neural Networks)

- o Gradient Descent (Regularized for complex datasets)
- o Plotting the decision boundary (in case of Logistic Regression)

2. Model Training- Since all models are trained in a similar way, the steps for training all the models will remain almost same.
   - o Visualizing the data by plotting it
   - o Randomly initializing the parameters
   - o Implementing code to compute Cost Function
   - o Using Gradient Descent to try to minimize the Cost Function as a function of parameters
   - o Implementing code to compute Regularized Cost Function and Gradient Descent if the datasets are complex

3. Output- In this function, I will use Feature Scaling by Mean Normalization before calculating the

predictions to ensure the features from the input are on similar scale. Then, I will calculate the predictions of the datasets using the hypothesis for each model separately. After which, Training Accuracy will be calculated.

3. <u>Evaluation</u>- I will divide the given datasets into 2 parts- Training and Validation in the ratio 4:1. The bigger part will be used to train the model and the latter will be sent to output function to get accuracy of the model.

➢ <u>Comparison of Different Models</u>

To compare different models, I will create a function called Compare. This function will take 2 models and 1 dataset as input. It will calculate the outputs of the 2 models on the given dataset using the Evaluation function and thus compare the 2 models on the basis of accuracy and precision. It will return which model performs better on the given dataset.

WEEK WISE TIMELINE

WEEK 1

I will write the code for the following algorithms- Linear (and Polynomial Regression), Logistic Regression and KNN.

WEEK 2

I will write the code for the following algorithms- K-Means Clustering and Neural Networks.

WEEK 3

I will write the code for Model Training function, Output function and Evaluation function.

WEEK 4

I will write the code for Compare function. I will test the project for bugs and fix them if found. I will thus compile my project and submit it.

ABOUT ME

➢ PERSONAL DETAILS

- Name: Hansika Gupta

- Place: Shahjahanpur, Uttar Pradesh

- Email id: 21je0376@iitism.ac.in

- Phone: +91 8795622222

- University: Indian Institute of Technology (Indian School of Mines), Dhanbad

- Major: Computer Science and Engineering

- Year of Study: First Year

➢ WHY SHOULD I BE SELECTED FOR THIS PROJECT ?

I am very passionate about Machine Learning. I have finished the Coursera course on Machine Learning taught by instructor Andrew Ng till week 5 including 2 of the coding assignments. I even gave the quizzes at the end of each week in which I scored a 100%.

I will be committed towards this project and dedicate most of my time in its completion in the

specified time period according to the Week Wise Timeline.

I would love to work further in the field of Machine Learning and learn as much as I can.

➢ COMMITMENT

I will be able to devote around 5 hours daily on an average to this project. I will finish the project in the given time period.