CSE 4554
Machine Learning Lab

Lab Mid

Dr. Hasan Mahmud
Professor, Department of CSE
Md. Tanvir Hossain Saikat
Junior Lecturer, Department of CSE

November 18, 2024

Contents

1 Resources

1.1 Dataset

For task 2 (logistics regression) - the dataset is MNIST dataset which has 28 * 28 sized images from (0-9). We have modified the dataset to have only 0 & 1.

1.2 Google Colab

Google Colab Link For Task 2 (Logistics Regression)

2 Tasks

2.1 Task 1 (Basic Python, Data Cleaning, Pre-processing & Linear Regression)

This Sunday MIT (Mirpur Institute of Technology) CSE Dept. will go to their annual departmental picnic. As a student from CSE Dept. you are also going with much excitement. But, you have a very important task to do. You are given the responsibility to create a linear model which will help your senior brothers make a good estimate of cgpa, number of projects and number of interships to make a good salary. So, now, you will create a dataset of graduated students with their name, cgpa, no of projets, internship and salaries with 1000 rows & 5 columns. Below is the description of each of the columns. The first row of the dataset will contain the name of the column. [Note: The seed value for generating the random numbers will be your id]

1. First Column:
   Name: Name
   Datatype: String
   Assignment: Random

2. Second Column:
   Name: CGPA
   Datatype: Float
   Range: 2.4 - 4.0
   Assignment: Random

3. Third Column:
   Name: Number of Projects
   Datatype: Integer
   Range: 0-20
   Assignment: Random

4. Fourth Column:
   Name: Number of Internship
   Datatype: Integer
   Range: 0-20
   Assignment: Random

5. Fifth Column:
   Name: Salary (In BDT)
   Datatype: Float
   Range: 30000 - 100000
   Assignment: Follow a specific condition for assignment with some outliers based on CGPA, Projects and Internship

Now, in my undergraduate, I had a cgpa of 3.92, 6 projects and one internship, based on your dataset trained on a linear regression model what would be my salary?
NOTE: You have to pre-process the dataset, you can use linear regression model implementation from your classroom submission

Submit the dataset file (.csv), google colab pdf & the google colab link

2.2 Task 2 (Logistics Regression)

When you run all the cells of the given google colab link you will see the accuracy of the model on binary MNIST is very poor.

1. Improve the test accuracy to more than 90% without increasing the number of epochs.
2. Modify the code to handle multi-class classification (0-9)

Submit google colab pdf & the colab code

2.3 Task 3 (Comparing Different Models)

Now use the binary MNIST dataset to evaluate the logistics regression, decision tree & naive bayes classifier models from sklearn.

1. Find out all the accuracy metrics (studied in the class) and mention what decision we can take from them as to why there need to be multiple accuracy metrics at the first place.
2. Identify the strengths and weaknesses of each of the models for the dataset with proper justification from the result
3. Modify decision tree to handle multi-class classification

Submit google colab pdf & the colab code

3 Submission Guidelines

Submit all the required files with the naming convention - ID_CSE4554_Lab_MID