भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 306: Machine Learning Lab
Practice Assignment 1

Instructions: This is only for practice. Complete it by 12:00 PM today. Your completion will be reviewed by the Teaching Assistants.

1. Please download The salary dataset regarding the prediction of salary from the year of experience.

   (a) Write a python program to read the dataset and display the number of features for the prediction of salaries, the number of patterns, the range of salaries in the data set.

   (b) Create a scatter plot to visualize the relationship between salaries and year of experiences.

   (c) Write a program to randomly split the dataset in X:Y ratio.
   Here, the values of X and Y are as follows (X+Y = 100% always):
   X = 10:10:90 (i.e., initial: 10%, increment by 10%, maximum is 90%)
   Y = 90:10:10 (i.e., initial: 90%, decrement by 10%, minimum is 10%)

   (d) Write a program to calculate the standard deviation, variance of salaries and year of experiences separately and covariance between between salaries and years of experience.

   (e) Write a program to calculate Pearson's correlation coefficient (manually without using any inbuilt function) between salaries and years of experience and interpret the result.

2. Write a program to create two different synthetic datasets, each having two variables (one dataset having negative linear correlation and the other dataset having positive linear correlation between variables). Create a scatter plot to visualize the relationship between X and Y. Report and interpret the value of covariance and Pearson's correlation coefficient between them.

3. Write a program to create two different synthetic datasets, each having two variables (one dataset having high variance and the other dataset having low variance). Create a scatter plot to visualize the relationship between X and Y. Report and interpret the value of covariance and the Pearson's correlation coefficient between them.

4. Write a program to create a synthetic dataset with two variables (having non-linear relation). Create a scatter plot to visualize the relationship between X and Y. Report and interpret the value of covariance and the Pearson's correlation coefficient between them.

5. Write a program to create a synthetic dataset with two variables X and Y, for 20 observations, where X is defined as the amount of fertilizers used (in tons) , and Y is defined as the amount of crops produced (in tons). Given, the variables have a linear relationship between them, then execute the followings:

(a) Create a scatter plot to visualize the relationship between X and Y. Label the axes, add a title, and use different colors for different classes/ targets (if applicable).

(b) Calculate the mean, median, standard deviation and variance of the whole dataset (say population), and display the result.

(c) Create a sample with observations having crop production greater than the mean crop production (w.r.t. entire dataset). Calculate the mean, median, standard deviation and variance of the sample, and display the result.

6. (a) Write a program to create a synthetic dataset for car reselling price with three numerical variables as below:

    1. price: Target variable (in dollars/ Rs.).
    2. mileage: Independent variable.
    3. Age of the car: Independent variable (can be calculated as ((current year)- (year of manufacture)), where current year is 2025.

   (b) Write a program to calculate the Pearson's correlation coefficients for all the variable pairs and display it. Determine the pairs with highest (positive / negative) correlations and interpret the results.