PSG College of Technology

Department of Applied Mathematics and Computational Sciences

MSc TCS IX Semester – Data Mining

DM Assignment 3 Data Preprocessing contd– 11-09-20

This assignment uses data from the UC Irvine Machine Learning Repository, a popular repository for machine learning datasets. In particular, we will be using the ” Auto MPG Data Set” available from <https://archive.ics.uci.edu/ml/datasets/Auto+MPG>.

1. How many cars and how many attributes are in the data set.

2. How many distinct car companies are represented in the data set? What is the name of the car with the best MPG? What car company produced the most 8-cylinder cars? What are the names of 3-cylinder cars? Do some internet search that can tell you about the history and popularity of those 3-cylinder cars.

3. What is the range, mean, and standard deviation of each attribute? Pay attention to potential missing values.

4. Plot histograms for each attribute. Pay attention to the appropriate choice of number of bins. Write 2-3 sentences summarizing some interesting aspects of the data by looking at the histograms.

5. Plot a scatterplot of weight vs. MPG attributes. What do you conclude about the relationship between the attributes? What is the correlation coefficient between the 2 attributes?

6. Plot a scatterplot of year vs. cylinders attributes. Add a small random noise to the values to make the scatterplot look nicer. What can you conclude? Do some internet search about the history of car industry during 70’s that might explain the results

7. Show 2 more scatterplots that are interesting do you. Discuss what you see.

8. Plot a time series for all the companies that show how many new cars they introduces during each year. Do you see some interesting trends?

9. Calculate the pairwise correlation, and draw the heatmap with Matplotlib. Do you see some interesting correlation?