Exercise 4: Part 2

Data Processing

Python Programming Bootcamp by Dr Rohitash Chandra
UNSW, 2021

Ensure that you use functions or OOP to do the following:

Description

1. Data is given in several formats and times you need clean, process and also fix missing values etc. Use Numpy with the Abalone dataset (https://archive.ics.uci.edu/ml/datasets/Abalone) and process it. Visualize the data by making histograms for each feature, and a co-variance matrix of all features. Plot the co-variance matrix as a heatmap using matplotlib. Then, provide a box-plot of the respective features. Report the mean and std of the entire set of features and the outcome variable (ring age).

2. In Abadone dataset, show the histogram ring age (outcome) and then divide the ring age into 4 major groups equally by age intervals of 25%. Report the class distribution of each class. Is this a class imbalance problem?

3. Repeat Part 1 for Adult, Iris and Wine datasets: https://archive.ics.uci.edu/ml/datasets/

4. Represent the outcome variable (class) of the processed datasets using one hot encoding.

5. Use pandas and seaborn and process and make the visualization again for all the above datasets. Note pandas tutorial here: https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html

Resources

1. https://datatofish.com/covariance-matrix-python/
2. https://datatofish.com/plot-histogram-python/
3. https://realpython.com/python-histograms/
4. Follow the example code here: https://www.geeksforgeeks.org/python-pandas-dataframe/
5. https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html