Python Tutorial 2 Exercises

January 27, 2022

This is the exercise of "Python Tutorial 2" for Prof. Xin Tong's DSO 530 class at the University of Southern California in spring 2022.

1. Write a program to achieve the following things (try to do parts 2) - 5) without looking at the Python Tutorial 2):

1) read in the Wine dataset and add column names just like what we do in Section 3 of Python Tutorial 2 .

2) use train_test_split from sklearn.model_selection to partition this dataset into separate training and test datasets to get X_train1, X_test1, y_train1, y_test1: set test_size to 0.3, set random_state to 1;

3) use train_test_split from sklearn.model_selection to partition this dataset into separate training and test datasets to get X_train2, X_test2, y_train2, y_test2: set test_size to 0.3, set random_state to 2;

4) use train_test_split from sklearn.model_selection to partition this dataset into separate training and test datasets to get X_train3, X_test3, y_train3, y_test3: set test_size to 0.3, set random_state to 1;

5) compare the column means of X_train1, X_train2 and X_train3

2. Write a program to achieve the following things (try to do the problems without looking at the Python Tutorial 2):

1) First create some missing values out of this Wine dataset: replace the first 20 rows of the Alcohol feature by np.NaN in the whole Wine dataset and take the whole dataset with 20 missing values as the starting point.

2) Impute the miss values using the median imputation techiniques.

3) Answer the following question:

Is it a recommended practice to split the dataset which was imputed in step 2) into training and test sets? If not, what would you do if you knew that you would need to split the data into training and test sets?