Processing Big Data for Analytics

Workshop – Feature Engineering

Table of Contents

List of python scripts and datasets 2
1. Missing data imputation 2
1.1 Mean/median imputation 2
1.2 Frequent category imputation 2
2. Categorical encoding 2
2.1 One-hot encoding 2
2.2 Label encoding 3
3. Feature Transformation 3
4. Feature Selection 3
5. PCA Example 3
6. Synthetic Data Generation 3
Workshop Submission 4

List of python scripts and datasets

File	Datasets
mean_median_imputation.ipynb	houseprice.csv
frequent_category_imputation.ipynb	houseprice.csv
EncodingExample.ipynb	Salary.csv
FeatureTransformation.ipynb	loan_small.csv
FeatureSelection.ipynb	Students2.csv bank.csv
PBDA-PCA example.ipynb	iris.data
Synthetic Data Generation.ipynb	NA
Day1_Workshop.ipynb	Built in dataset - Boston house price

1. Missing data imputation

We will use Jupyter Notebook(Anaconda3) as the IDE to write and run python scripts.

1.1 Mean/median imputation

Go to Anaconda terminal and open Jupyter Notebook. Open file “mean_median_imputation.ipynb”. In this example, we handle the missing data imputation for numeric data and we use the Mean/median imputation method on the dataset.

Run the script and observe results.

1.2 Frequent category imputation

Open file ”frequent_category_imputation.ipynb”. In this example, we handle the missing data imputation for both numeric and categorical data, we use the mean imputation method for numerical variables and most frequent category imputation for categorical variables.

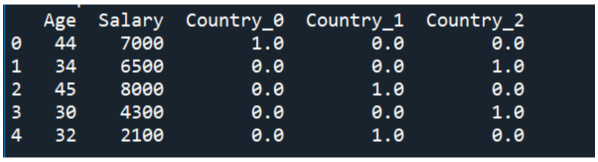
Run the script and observe results.

2. Categorical encoding

2.1 One-hot encoding

Open file “EncodingExample.ipynb”, in this example, we demonstrate how to use one-hot encoding to encode categorical variables.

The column “country” has been encoded into binary number vector:

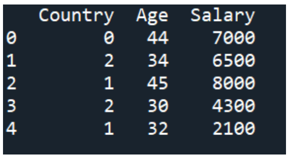


```
   Age  Salary  Country_0  Country_1  Country_2
0   44    7000        1.0        0.0        0.0
1   34    6500        0.0        0.0        1.0
2   45    8000        0.0        1.0        0.0
3   30    4300        0.0        0.0        1.0
4   32    2100        0.0        1.0        0.0
```

2.2 Label encoding

Open file "EncodingExample.ipynb", in this example, we demonstrate how to use label encoding to encode categorical variables.

The column "country" has been encoded into different number:



```
   Country  Age  Salary
0        0   44    7000
1        2   34    6500
2        1   45    8000
3        2   30    4300
4        1   32    2100
```

3. Feature Transformation

Open script "FeatureTransformation.ipynb", in this example, we present a completed feature engineering example including missing data imputation, encoding and feature scaling.

Run this script and observe outputs.

4. Feature Selection

Open the script "FeatureSelection.ipynb". In this example, we demo different feature selection methods including filter-based method, RFE as well as different select method such as SelectKBest and SelectPercentile.

Run this script and observe outputs.

5. PCA Example

Open the script"PCA _example.ipynb". In this example, we apply PCA on the following examples:

- PCA step by step calculation on a 40 samples dataset
- PCA on a sample 2D dataset
- PCA on iris dataset
- PCA on face recognization dataset
- PCA Example - Mathematical calculation

6. Synthetic Data Generation

Open the script" Synthetic Data Generation.ipynb". In this example, we discuss the details of generating different synthetic datasets using Numpy and Scikit-learn libraries. We see how different samples can be generated from various distributions with known parameters.

Workshop Submission

Open script “Workshop_Feature_Engineering.ipynb”, complete the questions. Copy and paste all the code into word document and name the file as “Workshop_Feature_Engineering_yourname.doc”, submit into the submission folder in CANVAS.