Department of Electrical & Computer Engineering

CPE 551A – Engineering Programming: Python

Fall 2021 – Mini-projec3

Yousef Abdelmalek – Hatim Alhazmi

Notes:

- Read all of the instructions and all of the questions before beginning the mini-project.
- Submit python file (.py).
- Add comments as much as you can into your file
- Don’t submit PDF or doc files
- Totaling 100 points.

This python project is a good start machine learning using dataset.

The dataset has 2 columns “YearsExperience” and “Salary” for 30 employees in a company. So in this project, we will train a Simple Linear Regression model to learn the correlation between the number of years of experience of each employee and their respective salary.

The objectives of this project are:

1. How to load dataset into Python environment
2. How to split the dataset into training dataset and testing dataset in your model.
3. How to apply linear regression model to the training dataset.
4. Predict the test set.
5. Visualizing the training and test set.
6. How to make predictions for a value that does not exist in the dataset.

Procedures on how to complete this project:

Step 1: Load the Dataset

Below is the code for loading the dataset. We will be using the pandas dataframe.

Here X is the independent variable which is the “Years of Experience” and y is the dependent variable which is the “Salary”

```python
import pandas as pd

dataset = pd.read_csv('Salary_Data.csv')

X = dataset.iloc[:, :-1].values # which simply means take all rows and all columns except last one

y = dataset.iloc[:,1].values # which simply means take all rows and all columns except last one
```

Note: This step assumes that pandas dataframe is downloaded on your machine. You may see the below link on how to install pandas in pycharm:

https://stackoverflow.com/questions/45548875/how-to-install-pandas-in-pycharm

Step 2: Statistics operation:

In this step you need to find:

1. The average salary for the 30 employees in the company.
2. Maximum salary for the 30 employees in the company
3. Minimum salary for the 30 employees in the company.

Calculate the above statistics using Python functions? [10 points]

Step 3: Split dataset into training set and test set:

In this step, we need to divide the dataset into training dataset and test dataset. The training dataset for training the model and the test data set for checking the performance of the model on the test dataset.

Test set will contain 10 observations and training set will contain 20 observations

Let’s use the train_test_split method from library model_selection to achieve this task

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=1/3)
```

If train_test_split is not installed on your local copy, you will need to install.

```
pip install -U scikit-learn
```

You can also visit this link for additional information on how to install this packages:

https://stackoverflow.com/questions/40704484/importerror-no-module-named-model-selection

You can also visit this link for additional information on how to add sklearn into Python interpretter:
https://stackoverflow.com/questions/32675024/getting-pycharm-to-import-sklearn

Compute the number of records of X_train, X_test, y_train, y_test? [10 points]

Step 4: Fit Simple Linear Regression model to training set:

Here you will use the LinearRegression class from the library sklearn.linear_model.

First: you need to create an object of the LinearRegression class

Second: call the fit method passing the X_train and y_train

[10 points]

```
from sklearn.linear_model import LinearRegression

regressor = LinearRegression()

regressor.fit(X_train,y_train)
```

Step 5: Predict the test set:

Using the regressor we trained in the previous step, you will now use it to predict the results of the test set and compare the predicted values with the actual values.

```
y_pred = regressor.predict(X_test)
```

In this step you can compare and see how well your model did comparing to the actual data.

Compare y_pred and y-test and comment on the difference between them. [10 points]

Step 6 — visualizing the training set:

Install matplot: Pycharm -> file->settings->python interpreter-> search for matplotlib

import matplotlib.pyplot as plt

First plot the actual data points of training and test set — X_train and y_train.

```
        plt.scatter(X_train, y_train, color = ‘red’)
```

Next plot the regression line — which is the predicted values for the X_train.

```
        plt.plot(X_train, regressor.predict(X_train), color=’blue’)
```

Use plt.title function to add a title to the plot. The title of the plot is ‘Salary vs. Experience (Training set)’

Use plt.xlable to add x-axis to the plot. The x-axis name is ‘Years of Experience’

Use plt.ylable to add x-axis to the plot. The x-axis name is ‘Salary’

Use plt.show function to show the plot

[20 points]

Step 7— visualizing the test set:

First plot the test dataset points X_test and y_test :

Second plot the regression line — `X_train, regressor.predict(X_train)`

[20 points]

Step 8— Make new predictions:

Make brand new predictions for data points that do not exist in the dataset. For example for a person with 15 years experience [20 points].