Лабораторна робота 4
Побудова лінійної регресії в Python

Мета: ознайомитись з поняттями простої лінійної регресії та роботи з наданими даними для прогнозування в Python.

Передумови / сценарій

У статистиці лінійна регресія - це спосіб моделювання взаємозв'язку між залежною змінною p і незалежною змінною x. У цій лабораторній роботі ви проаналізуєте дані про продажі району та побудуєте просту лінійну регресію для прогнозування річного чистого обсягу продажів на основі кількості магазинів у районі. Завдання до лабораторної роботи знаходиться за посиланням:

https://static-course-assets.s3.amazonaws.com/IoTFBDA201/en/course/files/4.1.2.4%20Lab%20-%20Simple%20Linear%20Regression%20in%20Python.html

Необхідні ресурси

- 1 ПК з доступом до Інтернету
- Бібліотеки Python: `pandas`, `numpy`, `scipy`, `matplotlib`
- Файли даних: store-dist.csv

Частина 1: Імпорт бібліотек та даних

У цій частині ви імпортуєте бібліотеки та дані з файлу `stores-dist.csv`.

Крок 1: Імпортуйте бібліотеки.

На цьому кроці ви імпортуєте такі бібліотеки:

- `matplotlib.pyplot` як plt
- `numpy` як np
- `pandas` як pd

```
# Code Cell 1
```

Крок 2: Імпортуйте дані.

На цьому кроці ви імпортуєте дані з файлу `stores-dist.csv`та переконаєтесь, що файл імпортовано правильно.

```
Code Cell 2
# Import the file, stores-dist.csv
salesDist = pd.read_csv('./Data/stores-dist.csv')
```

```
# Verify the imported data
salesDist.head()
```

Заголовки стовпців `annual net sales`і `number of stores in district` перейменовані для полегшення обробки даних.

- `annual net sales` до продажів
- `number of stores in district` до магазинів

```
# Code Cell 3
# The district column has no relevance at this time, so it can be dropped.
salesDist = salesDist.rename(columns={'annual net sales':'sales','number of stores in district':'stores'})
salesDist.head()
```

Частина 2: Складання даних

Крок 1: Визначте співвідношення.

На цьому кроці ви дослідите кореляцію даних до регресійного аналізу. Ви також скинете будь-які не пов'язані між собою стовпці за необхідності.

```
# Code Cell 4
# Check correlation of data prior to doing the analysis
# # Hint: check lab 3.1.5.5
```

З коефіцієнта кореляції виявляється, що стовпець `district` має низьку кореляцію до `annual net sales`і `number of stores in the district`. Отже, колонна округа не є необхідною як частина регресійного аналізу. Колонка `District` може бути виключена з dataframe.

```
# Code Cell 5
# The district column has no relevance at this time, so it can be dropped.
#sales = salesDist.drop(...)

sales.head()
```

За даними коефіцієнта кореляції, який тип кореляції ви спостерігали між річними чистими продажами та кількістю магазинів у районі?

 Введіть тут свою відповідь.

Крок 2: Створіть сюжет.

На цьому кроці ви створите графік для візуалізації даних. Ви також призначите магазини як незалежну змінну ***x*** а продажі як залежна змінна ***p***.

```
# Code Cell 6
# dependent variable for y axis
y = sales['sales']
# independent variable for x axis
x = sales.stores
```

```
# Code Cell 7
# Display the plot inline
%matplotlib inline

# Increase the size of the plot
plt.figure(figsize=(20,10))

# Create a scatter plot: Number of stores in the District vs. Annual Net Sales
plt.plot(x,y, 'o', markersize = 15)

# Add axis labels and increase the font size
plt.ylabel('Annual Net Sales', fontsize = 30)
plt.xlabel('Number of Stores in the District', fontsize = 30)

# Increase the font size on the ticks on the x and y axis
plt.xticks(fontsize = 20)
plt.yticks(fontsize = 20)

# Display the scatter plot
plt.show()
```

Частина 3: Побудуйте просту лінійну регресію

У цій частині ви будете використовувати numpy для створення лінії регресії для аналізованих даних. Ви також розрахуєте центроїд для цього набору даних. Центрод - це середнє значення для набору даних. Сформована проста лінійна лінійна регресія також повинна проходити через центроїд.

Крок 1: Обчисліть нахил та перетин Y-лінії лінійної регресії.

```
# Code Cell 8
# Use numpy polyfit for linear regression to fit the data
# Generate the slope of the line (m)
# Generate the y-intercept (b)
m, b = np.polyfit(x,y,1)
print ('The slope of line is {:.2f}.'.format(m))
print ('The y-intercept is {:.2f}.'.format(b))
print ('The best fit simple linear regression line is {:.2f}x + {:.2f}.'.format(m,
b))
```

Крок 2: Обчисліть центроїд.

Центроїд набору даних обчислюється за допомогою функції середнього значення.

```
# Code Cell 9
# y coordinate for centroid
y_mean = y.mean()
# x coordinate for centroid
x_mean = x.mean()
print ('The centroid for this dataset is x = {:.2f} and y = {:.2f}.'.format(x_mean
, y_mean))
```

Крок 3: Накладіть лінію регресії та центральну точку на графіку.

```
# Code Cell 10
# Create the plot inline
%matplotlib inline

# Enlarge the plot size
plt.figure(figsize=(20,10))

# Plot the scatter plot of the data set
plt.plot(x,y, 'o', markersize = 14, label = "Annual Net Sales")

# Plot the centroid point
plt.plot(x_mean,y_mean, '*', markersize = 30, color = "r")

# Plot the linear regression line
plt.plot(x, m*x + b, '-', label = 'Simple Linear Regression Line', linewidth = 4)

# Create the x and y axis labels
plt.ylabel('Annual Net Sales', fontsize = 30)
plt.xlabel('Number of Stores in District', fontsize = 30)

# Enlarge x and y tick marks
plt.xticks(fontsize = 20)
plt.yticks(fontsize = 20)

# Point out the centroid point in the plot
plt.annotate('Centroid', xy=(x_mean-0.1, y_mean-5), xytext=(x_mean-3, y_mean-20),
arrowprops=dict(facecolor='black', shrink=0.05), fontsize = 30)

# Create legend
plt.legend(loc = 'upper right', fontsize = 20)
```

Крок 4: Прогнозування

Використовуючи лінійну лінійну регресію, ви можете передбачити річний чистий обсяг продажів на основі кількості магазинів у районі.

```
# Function to predict the net sales from the regression line
def predict(query):
    if query >= 1:
        predict = m * query + b
        return predict
    else:
        print ("You must have at least 1 store in the district to predict the annu
al net sales.")

 # Code Cell 12
# Enter the number of stores in the function to generate the net sales prediction.
```

Який прогнозований чистий продаж, якщо в районі є 4 магазини?
Part 1: *Введіть тут свою відповідь.*