Лабораторна робота №6

Регресія. Метод найменших квадратів

Мета роботи: ознайомитися з методом найменших квадратів та набути навичок роботи в середовище розробки Python

Література

Документація по бібліотеці Seaborn - https://seaborn.pydata.org
Statsmodels https://www.statsmodels.org/stable/index.html

Зміст роботи

Машинне навчання намагається прогнозувати подальші події шляхом аналізу попереднього досвіду - наприклад, намагається скласти прогноз погоди на завтра, або вгадати курс акцій на біржі, або, скажімо, діагностувати хворобу пацієнта, ґрунтуючись на його попередньої історії хвороби.

Класифікація намагається визначити категорію вхідних даних або наявність, або відсутність якоїсь їх особливості - наприклад, намагається розпізнати написану цифру або визначити, чи міститься на зображенні кіт.

Регресія ж обчислює певне число або вектор - наприклад, завтрашню температуру або ціну на акції Google.

Лінійна регресія (Linear regression) - модель залежності змінної x від однієї або декількох інших змінних (факторів, регресорів, незалежних змінних) з лінійною функцією залежності.

Лінійна регресія відноситься до задачі визначення «лінії максимальної відповідності умовам» через набір точок даних і стала простим попередником нелінійних методів, які використовують для навчання нейронних мереж.

Проста лінійна регресія

Проста лінійна регресія є підходом для прогнозування кількісної відповіді з використанням однієї ознаки. Вона має наступний вигляд:

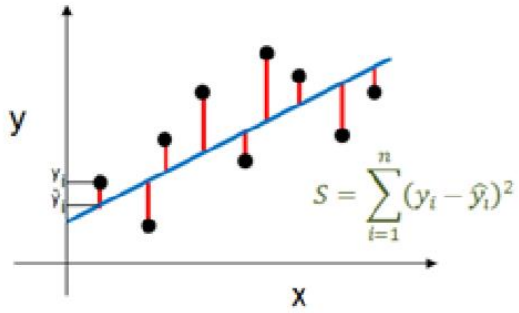
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Де β_0 зрушення (довжина відрізку, що відсікається на координатної осі прямої Y), β_1 - нахил прямої Y, ε_i- випадкова помилка змінної Y в і-м спостереженні.

Разом β_0 і β_1 називаються модельними коефіцієнтами. Щоб створити модель, необхідно дізнатися їх значення. І, як тільки ці коефіцієнти знайдені, можна використовувати модель для прогнозування продажу.

Оцінка ("навчання") модельних коефіцієнтів

Взагалі, коефіцієнти оцінюються з використанням критерію найменших квадратів, що означає, що необхідно знайти лінію (математично), яка мінімізує суму квадратних залишків (або "суму квадратних помилок"):

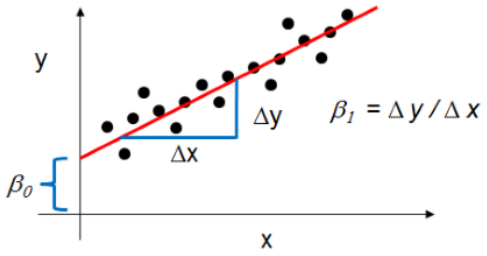


Як модельні коефіцієнти відносяться до лінії найменших квадратів?

β_0 є перехопленням (значення y при x= 0)

β_1 - нахил (зміна y поділена на зміну x)

Графічне зображення цих розрахунків:



Приклад.

В результаті дослідження, бьуло отримано чотири точки (x,y) даних: (1,6), (2,5), (3,7) і (4,10).

Необхідно знайти пряму $y=\beta_0+\beta_1 x$, яка найкраще підходить для цих точок. Для цього необхідно знайти β_0 і β_1 і розвязати систему рівнянь

$\beta_0+1\beta_1=6$
$\beta_0+2\beta_1=5$

$\beta_0+3\beta_1=7$
$\beta_0+4\beta_1=10$

Метод найменших квадратів: розв'язання полягає у спробі зробити якомога меншою суму квадратів похибок між правою і лівою сторонами цієї системи, тобто необхідно знайти мінімум функції

$$S(\beta_0,\beta_1)=[6-(\beta_0+1\beta_1)]^2+[5-(\beta_0+2\beta_1)]^2+[7-(\beta_0+3\beta_1)]^2+[10-(\beta_0+4\beta_1)]^2.$$

Мінімум визначають через обчислення часткової похідної від $S(\beta_0,\beta_1)$ щодо β_0 і β_1 і прирівнюванням її до нуля

$$\frac{\partial S}{\partial \beta_0} = 0 = 8\beta_0 + 20\beta_1 - 56$$

$$\frac{\partial S}{\partial \beta_1} = 0 = 20\beta_0 + 60\beta_1 - 154$$

Це приводить до системи з двох рівнянь і двох невідомих, які називаються нормальними рівняннями. Якщо розв'язати, ми отримуємо

$\beta 0=3.5$
$\beta 1=1.4$

В результаті отримаємо рівняння $y=3.5+1.4x$ яке є рівнянням лінії, яка підходить найбільше. Мінімальна сума квадратів похибок є

$$S(3.5,1.4)=1.1^2+(-1.3)^2+(-0.7)^2+0.9^2=4.2.$$

Завдання 1. Експериментально отримані N-значень величини Y при різних значеннях величини X. Відшукати параметри функції за методом найменших квадратів. Зробити креслення, де в декартовій системі координат побудувати експериментальні точки і графік апроксимуючої функції.

Варіанти завдань:

1	X	0	5	10	15	20	25
	Y	21	39	51	63	70	90
2	X	-1	-1	0	1	2	3
	Y	-1	0	1	1	3	5
3	X	7	12	17	22	27	32
	Y	8	7	6	5	4	3
4	X	2	4	6	8	10	12
	Y	6,5	4,4	3,8	3,5	3,1	3,0
5	X	-5	-4	0	1	3	5
	Y	5,3	20,7	21,7	9,2	55,4	64,3

6	X	3,33	1	63	0,87	0,42	0,27
	Y	0,48	1,03	2,02	4,25	7,16	11,5
7	X	-12	29	0	4	6	8
	Y	-3	0	1	2	9	5
8	X	6	7	8	9	10	12
	Y	2	3	3	4		5
9	X	0,3	1,0	1,5	2,2	3,6	4,5
	Y	5	10	13	16	17	18
10	X	1	6	11	16	21	26
	Y	19	37	49	61	68	90

Завдання 2. Дослідити залежність продажів від витрат на рекламу на телебаченні, радіо та в газеті**.**

Опис даних

Вхідні дані знаходяться у Advertising.csv файлі за посиланням: http://faculty.marshall.usc.edu/gareth-james/ISL/data.html, що представляють собою набір даних з книги Introduction to Statistical Learning.

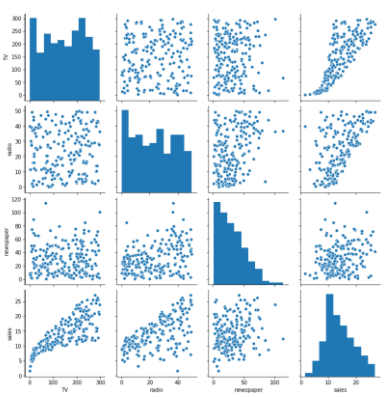
Бібліотеки, що будуть використані:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Завантажити в Pandas Dataframe. Далі необхідно подивитися на перші 5 записів, статистику за ознаками та розмір масиву.

Для того, що б наочно побачити можливу статистичну залежність в даних необхідно побудувати парні графіки. Зробити це зручно за допомогою бібліотеки *seaborn* і метода *pairplot* який будує попарні залежності ознак з датасета (ознаки - це колонки).

Результат:

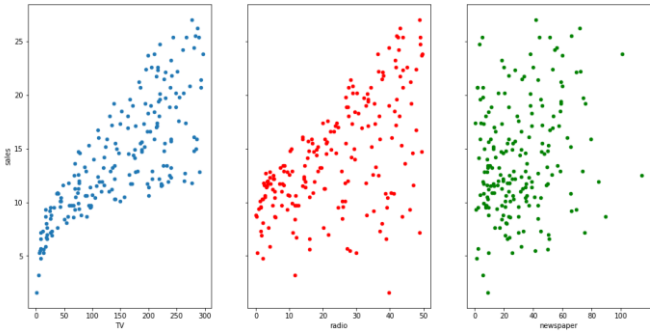


На діагоналі представлено розподіл відповідної ознаки, наприклад, скільки всього було Sales (продаж) після відповідної реклами.

Візуалізувати зв'язок між фактом і відгуком (TV і sales, radio і sales, newspaper і sales) можна за допомогою діаграм розсіювання.

```
fig, axs = plt.subplots(1, 3, sharey=True)
df.plot(kind='scatter', x='TV', y='sales', ax=axs[0], figsize=(16, 8))
df.plot(kind='scatter', x='radio', y='sales', color='red', ax=axs[1])
df.plot(kind='scatter', x='newspaper', y='sales', color='green', ax=axs[2]
)
```

Результат:



Питання, що виникають і потребують рішення :

– ***Чи існує взаємозв'язок між оголошеннями та продажами? Наскільки сильні зв'язки?***

– ***Які типи оголошень сприяють продажам?***

– ***Який вплив має кожен тип продажів?***

– ***Враховуючи витрати на рекламу на певному ринку, чи можна прогнозувати продаж?***

З графіків, що були побудувані, вже можна зробити декілька цікавих висновків за даними, щодо того, як впливає реклама в газетах, на радіо і TV на продажи. Видно, що найменше на продажі впливає реклама в газетах, потім рклама на радіо і нарешті найбільше - реклама на TV.

Далі, можна розрахувати коефіцієнт кореляції даних.

Результат:

	Unnamed: 0	TV	radio	newspaper	sales
Unnamed: 0	1.000000	0.017715	-0.110680	-0.154944	-0.051616
TV	0.017715	1.000000	0.054809	0.056648	0.782224
radio	-0.110680	0.054809	1.000000	0.354104	0.576223
newspaper	-0.154944	0.056648	0.354104	1.000000	0.228299
sales	-0.051616	0.782224	0.576223	0.228299	1.000000

Коефіцієнт кореляції між рекламою на TV і продажами = 0,782224 (78 відсотків), далі йде радіо - 0,576223 (57%) ну і нарешті газети – 0,228299 (22,8%).

Отже, розрахований коефіцієнт кореляції свідчить про наявність значного зв'язку між рекламою на TV і продажами

Використовуємо пакет *Statsmodels* для оцінки модельних коефіцієнтів для рекламних даних:

```
import statsmodels.formula.api as smf

lm = smf.ols(formula='sales~TV', data=df).fit()
print(lm.params)
```

Результат:

```
Intercept    7.032594
TV           0.047537
dtype: float64
```

Приклад:

Припустимо, що є новий ринок, де витрати на рекламу на телебаченні планують у розмірі $ 50,000. Який прогноз продажу можна передбачали на цьому ринку?

Прогноз продаж на новому ринку можна розрахувати вручну:

$$y = \beta_0 + \beta_1 x$$

$$y=7.032594+0.47537*50=9,409444$$

Можна використати *Statsmodels*, щоб зробити прогноз:

```
#потрібностворити DataFrame, оскільки його очікує інтерфейс формули Statsmodels
X_new = pd.DataFrame({'TV': [50]})
print(X_new.head())
#використати модель, щоб зробити прогнози на нове значення
print(lm.predict(X_new))
```

Результат:

```
   TV
0  50
0    9.409426
dtype: float64
```

Завдання 3. Провести дослідження курсу української гривні до долару США за деякий період часу.

Контрольні запитання

1. Для чого застосовується регресійний аналіз?
2. Що таке лінійна регресія?
3. У чому суть методу найменших квадратів?
4. Що таке нахил у рівнянні лінійної регресії?
5. Як розраховуються коефіцієнти рівняння лінійної регресії?
6. Опишіть методи рішення системи лінійних рівнянь?
7. Які переваги і недоліки методу найменших квадратів?