Лабораторна робота №3

Лінійні моделі регресії в Python

Мета: набути навичок роботи в середовище розробки Python та провести найпростішого аналізу даних.

Література

Документація по бібліотеці Seaborn - https://seaborn.pydata.org
seaborn.pairplot()- https://seaborn.pydata.org/generated/seaborn.pairplot.html
seaborn.boxplot() - https://seaborn.pydata.org/generated/seaborn.boxplot.html
Statsmodels https://www.statsmodels.org/stable/index.html

Зміст роботи

Машинне навчання – це потужний інструмент, який може використовуватися для прогнозування майбутніх подій шляхом аналізу попереднього досвіду. Наприклад, скласти прогноз погоди на завтра, або вгадати курс акцій на біржі, або діагностувати хворобу пацієнта, ґрунтуючись на його попередньої історії хвороби.

Класифікація може визначити категорію вхідних даних або наявність, або відсутність якоїсь їх особливості. Наприклад, намагатися розпізнати написану цифру або визначити, чи міститься на зображенні кіт.

Регресія 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.

Завдання 1. Дослідити залежність продажів від витрат на рекламу на телебаченні, радіо та в газеті.

Опис даних.

Вхідні дані знаходяться у ***Advertising.txt***, що представляють собою набір даних з книги Introduction to Statistical Learning.

Бібліотеки, які будуть використані:

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

Дані потрібно завантажити в Pandas Dataframe. Так як в таблиці роздільник це кома і заголовок вже є, то ніяких додаткових параметрів вказувати не потрібно.

Далі, потрібно подивитися на перші 5 записів і статистику за ознаками. Для цього потрібно викликати метод *head()* об'єкта DataFrame. Як аргумент можна передати число рядків, які потрібно показати.

Результат:

	Unnamed: 0	TV	Radio	Newspaper	Sales
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9

```
df.shape
```

shape - розміри масиву, його форма. Це кортеж натуральних чисел, що показує довжину масиву по кожній осі. Для матриці з n рядків і m стовпів, shape буде (n, m).

Результат:

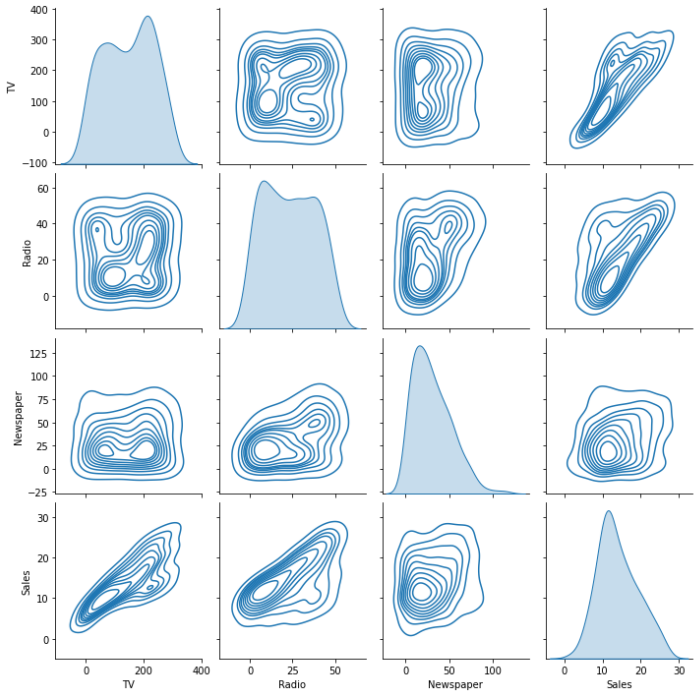
```
(200, 5)
```

Для того, що б наочно побачити можливу статистичну залежність в даних необхідно побудувати парні графіки. Зробити це зручно за допомогою бібліотеки *seaborn* (*https://seaborn.pydata.org*) в якій є метод *pairplot* який будує попарні залежності ознак з датасету (ознаки - це колонки).

Ознайомитися з параметрами методу:

```
sns.pairplot(df, palette='dict' , x_vars=('TV', 'Radio', 'Newspaper','Sales'), y_vars=('TV', 'Radio', 'Newspaper','Sales'), kind="kde", diag_kind='auto')
```

Результат:



На діагоналі представлено розподіл відповідної ознаки.

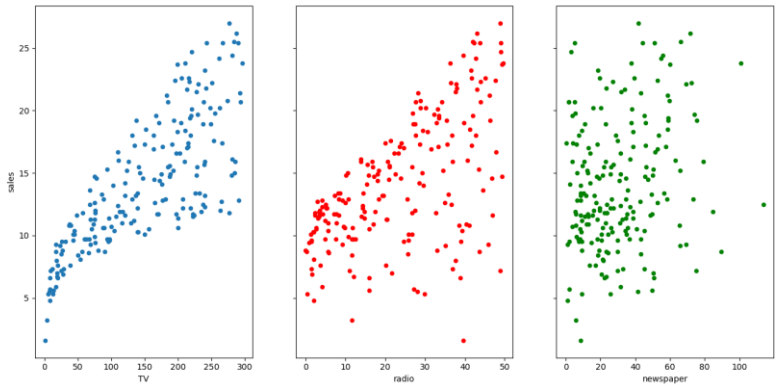
Можна створити скорочену версію:

```
sns.pairplot(df, kind="kde",
             x_vars=["TV", "Radio", "Newspaper"],
    y_vars=["Sales"],)
plt.show()
```

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 рекламою*** *в газетах, на радіо і TV* ***та продажами? Наскільки сильні зв'язки?***

– ***Які реклами сприяють продажам?***

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

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

З графіків, що побудували, вже можна зробити кілька цікавих висновків за даними, щодо того, як впливає реклама в газетах, радіо і TV на продаж. Видно, що найменше впливає реклама в газетах, потім в радіо і нарешті найбільше впливає на продаж реклама на TV.

Далі, можна розрахувати коефіцієнт кореляції даних.

Коефіцієнт кореляції *– показник, який використовують для вимірювання щільності зв'язку між результативними і факторними ознаками у кореляційно-регресійній моделі за лінійної залежності. Чим ближчий цей показник до 0, тим менший зв'язок, чим ближчий він до ±1 – тим зв'язок більш тісніший. Знак «плюс» при коефіцієнті кореляції означає прямий зв'язок між ознаками x і y, знак «мінус» – обернений.*

Коефіцієнт кореляції Пірсона *- це, показник кореляції (лінійної залежності) між двома змінними X та Y. Він широко використовується в науці для вимірювання ступеня лінійної залежності між двома змінними.*

Далі, можна розрахувати кореляцію даних за допомогою методу corr() - за замовчуванням це кореляції Пірсона, яка показує існування лінійної залежності між величинами.

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%).

Також для визначення кореляції можна скористатися тепловою картою:

```
sns.heatmap(df.corr(), cmap="PuBu", annot=True, fmt=".1f")
```

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

Проста лінійна регресія

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

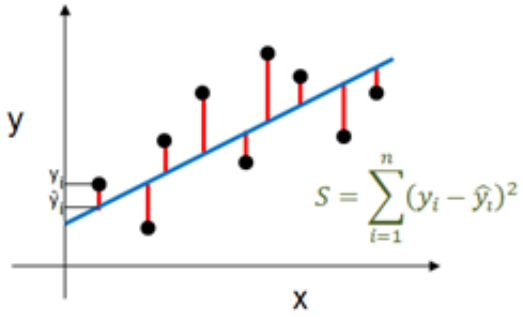
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Де β_0 зрушення (довжина відрізка, 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 значення цих коефіцієнтів. І як тільки ці коефіцієнти знайдені, можна використовувати модель для прогнозування продажів.

Оцінка ("навчання") модельних коефіцієнтів

Взагалі, коефіцієнти оцінюються з використанням алгоритму найменших квадратів, що означає, що необхідно знайти лінію (математично), яка мінімізує суму квадратів помилок:



За алгоритмом найменших квадратів невідомі параметри β_0 і β_1 лінійної регресії знаходяться із умов мінімізації суми квадратів відхилень, тобто із умов мінімізації функції.

Якщо детально розглянути діаграму можна побачити:

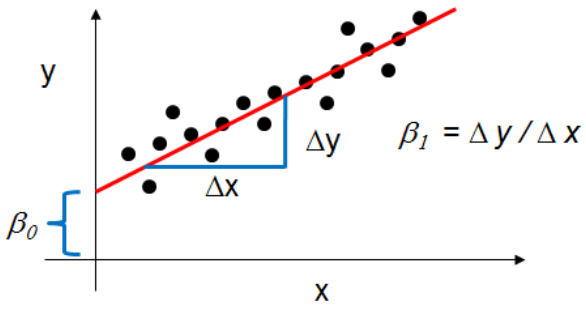
- *чорні крапочки це значення x і y, що спостерігаються;*
- *синя лінія - лінія найменших квадратів;*
- *червоні лінії є відстанями між спостережуваними значеннями і лінією найменших квадратів.*

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, було 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 розв'язати 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

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

Використовуємо пакет *Statsmodels* для оцінки модельних коефіцієнтів для заданих 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 $ 50тис.. 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:

```
#потрібно створити 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))
```

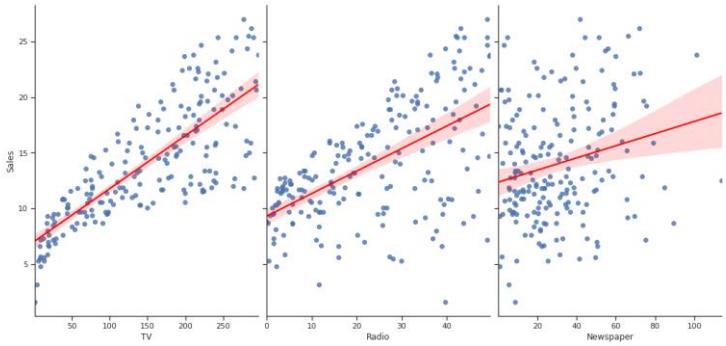
Результат:

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

Завдання 2. Провести додатковий аналіз:

- Побудуйте лінію регресії для продажів в залежності від реклами на TV, Radio і Newspaper за допомогою функції (`sns.pairplot()`), і 95% довірчий інтервал для цієї регресії: y ~ x

Результат:



Для зміни кольору лінії регресії використайте наступний параметр `plot_kws={'line_kws':{'color':'red'}}`

```
sns.pairplot (df, x_vars = ['TV', 'Radio', 'Newspaper'], y_vars = 'Sales',
size = 7, aspect = 0.7, kind = 'reg',
plot_kws={'line_kws':{'color':'red'}})
```

- Побудувати коробчасті діаграми для даних TV, Radio і Newspaper
- Розрахувати оцінки модельних коефіцієнтів для рекламних *в газетах і на радіо.*
- Зробити прогноз:
  - Якщо на рекламу на радіо буде потрачено $ 24 тис..
  - Якщо на рекламу в газеті буде потрачено $ 15 тис..
- Зробити висновки.

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. Які переваги і недоліки методу найменших квадратів?
7. У чому відмінність навчання з вчителем (supervised learning) від навчання без вчителя (unsupervised learning)?
8. В чому полягає завдання регресії в контексті машинного навчання?
9. Опишіть (в загальних рисах) кілька підходів до вирішення завдання регресії.
10. Для чого потрібна кореляція?
11. Що таке коваріація?
12. Доповнити визначення: Кореляційною залежністю називають залежність ...
13. Доповнити визначення: Кореляційний аналіз вивчає ...
14. Доповнити визначення: Коефіцієнтом кореляції називають ...
15. Коефіцієнт кореляції може приймати значення:
    a. від -1 до +1;

b. від 0 до +1;
c. від -1 до 0;
d. від +1 до + 2.

16. Якщо значення коефіцієнта кореляції по модулю близько до 1, то має місце ... кореляція:

a. середня;
b. сильна;
c. слабка.

17. Якщо значення коефіцієнта кореляції по модулю близько до 0, то має місце ... кореляція:

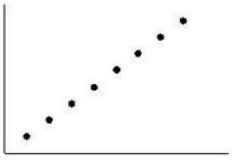
a. сильна;
b. слабка;
c. середня.

18. Якщо високі значеннях однієї змінної пов'язані з високими значеннями іншої, то такий зв'язок називається …

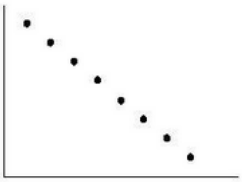
19. Якщо високі значення однієї змінної пов'язані з низькими значеннями іншої, то цей зв'язок називається …

20. При якому з наступних значень кореляції взаємозв'язок найбільш сильна? +0,81; -0,67; -0,86; +1,00 інтерпретація: Існує негативна кореляція між депресією і рівнем фізичної підготовки. Існує позитивна кореляція між обсягом домашньої бібліотеки і середнім балом учня. Існує негативна кореляція між оцінками і боязню іспитів.

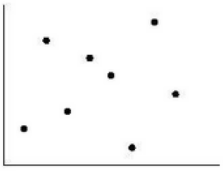
21. Визначте тип кореляційного зв’язку?



22. Визначте тип кореляційного зв’язку?



23. Визначте тип кореляційного зв’язку?



24. У чому полягає різниця між кореляцією Пірсона, Спірмана і Кендала?