# Лабораторна робота №3

## Регресія в Python

***Мета:*** набути навичок роботи в середовище розробки Python та провести регресійний аналіз даних.

***Література***

*Документація по бібліотеці Seaborn -*[*https://seaborn.pydata.org*](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*](https://www.statsmodels.org/stable/index.html)

## Зміст роботи

За допомогою машинного навчання можна прогнозувати подальші події 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.

Машинне навчання

Неконтрольоване навчання

Контрольоване навчання

*класи алгоритмів*

Класифікація

Регресія

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 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.

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

***Опис даних.*** Вхідні дані знаходяться у файлі ***Advertising.txt***, який взятий з книги Introduction to Statistical Learning.

Бібліотеки, які будуть використані:*pandas, matplotlib.pyplot, seabor.n* **Завантажити набір даних.** Так як в таблиці роздільник це кома і заголовок вже є, то ніяких додаткових параметрів вказувати не потрібно.

**Продивитися на перші 5 записів і статистику за ознаками.** Результат:



**Продивитися розміри масиву**

**Побудувати парні графіки**

Для того, що б наочно побачити можливу статистичну залежність в даних необхідно побудувати парні графіки. Зробити це зручно за допомогою бібліотеки *seaborn* в якій є метод *pairplot* який будує попарні залежності ознак з датасету.

Ознайомитися з параметрами методу:

ns.pairplot(df, palette='dict' , x\_vars=('TV', 'Radio', 'Newspaper','Sales '), y\_vars=('TV', 'Radio', 'Newspaper','Sales'), kind='scatter', diag\_kind

='auto') Результат:

 

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

**Побудувати діаграму розсіювання (TV i sales, radio i sales, newspaper i sales) для візуалізації зв’язку.** 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 відповідей:***

* *Чи існує взаємозв’язок між рекламою в газетах, на радіо і ТV та продажами? Наскільки сильні зв’язки?*
* *Яка реклама сприяє продажам?*
* *Який вплив має кожен тип реклами на продажі?*
* *Враховуючи витрати на рекламу на певному ринку, чи можна прогнозувати продаж?*

З графіків можна зробити кілька висновків за даними, щодо того, як впливає реклама в газетах, радіо і ТV на продажі. Видно, що найменший вплив на продажі від реклами в газетах, потім на радіо і нарешті найбільший від реклами на ТV.

**Розрахувати коефіцієнт кореляції даних.**

Для розрахунку кореляції використаємо метод corr() - за замовчуванням це кореляції Пірсона. Результат:



Коефіцієнт кореляції між рекламою на ТV 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 ТV і продажами*

# Проста лінійна регресія

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

*y = β0 + β1x+*

*Де β0**зрушення (довжина відрізка, що відсікається на координатної осі прямої Y), β1 - нахил прямої Y, εi- випадкова помилка змінної Y в i-м спостереженні.*

*Разом β0 і β1 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 критерію найменших квадратів, що означає, що необхідно знайти лінію*

*(математично), яка мінімізує суму квадратів помилок:*



*За методом найменших квадратів невідомі параметри β0 і β1 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?***

*β0 є перехопленням (значення y при x= 0) β1 - нахил (зміна у 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=β0+β1x, яка найкраще підходить для цих точок. Для цього необхідно знайти β0 і β1 і розв’язати систему рівнянь β0+1β1=6β0+2β1=5 β0+3β1=7 β0+4β1=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(β0,β1)=[6−(β0+1β1)]2+[5−(β0+2β1)]2+[7−(β0+3β1)]2+[10−(β0+4β1)]2.

Мінімум визначають через обчислення часткової [похідної](http://fitm.nusta.edu.ua/mediawiki/index.php?title=%D0%A7%D0%B0%D1%81%D1%82%D0%BA%D0%BE%D0%B2%D0%B0_%D0%BF%D0%BE%D1%85%D1%96%D0%B4%D0%BD%D0%B0&action=edit&redlink=1) від S(β0,β1) щодо β0 і β1 і прирівнюванням її до нуля



Це приводить до системи з двох рівнянь і двох невідомих, які називаються нормальними рівняннями. Якщо розв'язати, ми отримуємо β0=3.5 β1=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)=1.12+(−1.3)2+(−0.7)2+0.92=4.2.

**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:



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 = β0 + β1x

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:



### Завдання для самостійного виконання 1

* Використовуючи датасет med.csv провести аналіз даних з застосуванням діаграм розсіювання обраних показників (3-4 параметри)
* Побудувати парні графіки для обраних параметрів

### ***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'}}

* Побудуйте коробчасті діаграми для даних TV, Radio і Newspaper - Зробити прогноз:
	+ Якщо на рекламу на радіо буде витрачено $ 50 000.
	+ Якщо на рекламу в газеті буде витрачено $ 50 000.
* Зробити висновки.

###  Завдання для самостійного виконання 2

* Використовуючи датасет med.csv подубувати графіки лінійної регресії обраних показників (3-4 параметри)

### Завдання для самостійного виконання 3

* Провести дослідження курсу української гривні до курсу долару США за деякий період часу.
* Використати для проведення аналізу архів НБУ курсу долара США до гривни за 2005 -2024 роки.

## 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 ковариація?
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:
	1. від -1 до +1;
	2. від 0 до +1;
	3. від -1 до 0;
	4. від +1 до + 2.
16. Якщо значення коефіцієнта кореляції по модулю близько до 1, то має місце ... кореляція:
	1. середня;
	2. сильна;
	3. слабка.
17. Якщо значення коефіцієнта кореляції по модулю близько до 0, то має місце ... кореляція:
	1. сильна;
	2. слабка;
	3. 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?



1. Визначте тип кореляційного зв’язку?

 

1. Визначте тип кореляційного зв’язку?



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