Лабораторна робота №7

Лінійна регресія. Метод найменших квадратів. Інтерполяція

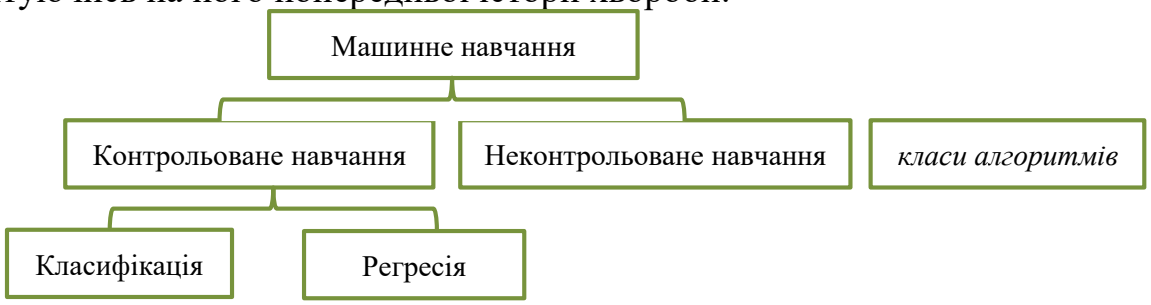
Мета роботи: Опрацювати поняття «лінійна регресія» і дослідити метод найменших квадратів та набути навички роботи в середовищі Python.

Час виконання: 4 години

Зміст роботи

Завдання 1. Ретельно опрацювати теоретичні відомості з лекційного курсу

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.



Класифікація може визначити категорію вхідних даних або наявність, або відсутність якоїсь їх особливості. Наприклад, намагаться 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:

$y = \beta 0 + \beta 1x + \varepsilon$

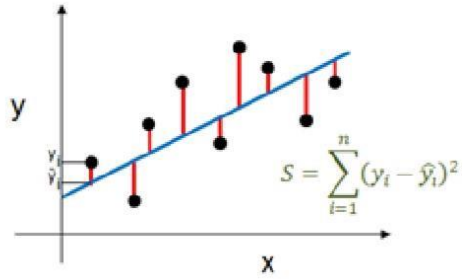
Де $\beta 0$ зрушення (довжина відрізка, що відсікається на координатної осі прямої Y), $\beta 1$ - нахил прямої Y, ε_i- випадкова помилка змінної Y у i-м спостереженні.

Разом $\beta 0$ і $\beta 1$ 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

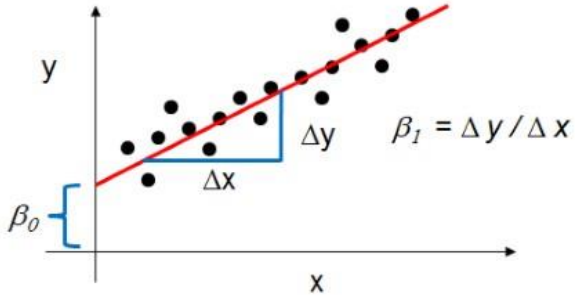
Взагалі, коефіцієнти оцінюються з використанням критерію найменших квадратів, що означає, що необхідно знайти лінію

(математично), яка мінімізує суму квадратів 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, 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 $\beta0=3.5$ $\beta1=1.4$

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 2. Експериментально отримані N-значень величини Y при значеннях величини X. Відшукати параметри функції за методом найменших квадратів.

Побудувати графіки, де в декартовій системі координат нанести 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	6	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
11	X	28	14	54	16	22	15
	Y	-15	10	4	5	11	28
12	X	13,33	21	63,75	20,87	40,42	30,27
	Y	10,48	21,03	23,02	41,25	27,16	51,5
13	X	16	27	38	19	100	72
	Y	12	35	39	41	60	55

14	X	6,5	4,4	3,8	3,5	3,1	3,0
	Y	-5	-4	0,7	1,25	3	5
15	X	8	7	6	5	4	3
	Y	2	4	6	8	10	12

Інтерполяція.

Нехай на деякій множині задана система функцій $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_m(x)$, які в подальшому будемо вважати досить гладкими (наприклад, безперервно диференціюються) функціями. Назвемо цю систему основною.

Функції виду:

$$Q_m(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_m\varphi_m(x), \quad (1)$$

де c_0, c_1, ..., c_m - постійні коефіцієнти, називаються узагальненими многочленами порядку m. Зокрема, якщо основна система складається з цілих невід'ємних степенів змінної x, т. е. $\varphi_0(x) = 1$, $\varphi_1(x) = x$, ..., $\varphi_m(x) = x_m$, то

$$Q_m(x) = c_0 + c_1x + \ldots + c_mx^m \quad (2)$$

є звичайний поліном ступеня m.

Якщо

$$\varphi_0(x) = 1,$$
$$\varphi_1(x) = \cos x,$$
$$\varphi_2(x) = \sin x,$$
$$\ldots \quad (3)$$
$$\varphi_{2m-1}(x) = \cos mx,$$
$$\varphi_{2m}(x) = \sin mx,$$

то

$$Q_m(x) = a_0 + a_1\cos x + b_1\sin x \ldots + a_m\cos mx + b_m\sin mx \quad (4)$$

називається тригонометричним поліномом (або тригонометричним многочленом) порядку m.

Завдання про наближення функцій ставиться таким чином: цю функцію f (x) потрібно замінити узагальненим многочленом Qm (x) заданого порядку m так, щоб відхилення (в сенсі σ або (Δi, Δh)) функції f (x) від узагальненого многочлена Qm (x) на зазначеній множині {x} було найменшим. При цьому многочлен Qm (x) в загальному випадку називається апроксимується.

Якщо множина {x} складається з окремих точок x_0, x_1, ..., x_n,

то наближення називається дискретним. Якщо ж {x} є відрізок

$a \le x \le b$, то наближення називається інтегральним.

На практиці часто користуються приближеннями функцій звичайним і тригонометричним поліномами.

В теорії дискретного наближення функцій має місце задача інтерполяції функцій. У разі звичайного полінома завдання інтерполяції формулюється в такий спосіб: для даної функції f(x) знайти поліном $Q_m(x)$ можливо нижчого ступеня m,

що приймає в заданих точках xi (i = 0, 1, 2, ..., n; $x_i \neq x_j$ при $i \neq j$) ті ж значення, що і f (x), т. е. такий, що $Q_m(x_i) = f(x_i)$ (i = 0, 1, 2, ..., n). Такий поліном називають інтерполяційним, а точки x_i (i = 0, 1, 2, ..., n) - вузлами інтерполяції.

Нехай маємо функцію $f(x)$. Розглянемо відрізок $[a,b]$ на якому визначено дану функцію. Зафіксуємо на цьому відрізку $n+1$ значень аргумента x_i. При цьому покладемо

$$f(x_i) = P(x_i), \tag{5}$$

де $P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$. Підставивши даний вираз для $P(x)$ в (1), отримаємо систему з $n+1$ рівнянь з $m+1$ невідомими коефіцієнтами $a_0, a_1 .....a_m$:

$$\begin{aligned} &a_0 + a_1 x_1 + a_2 x_1^{\,2} + ... + a_m x_1^{\,m} = f(x_1) \\ &\text{----------} \\ &a_0 + a_1 x_{n+1} + a_2 x_{n+1}^{\,2} + ... + a_m x_{n+1}^{\,m} = f(x_{n+1}). \end{aligned} \tag{6}$$

Визначник складений з $x_1, x_2 .....x_n$ має назву визначник Вандермонда:

$$W(x_1, x_2, ...x_n) = \begin{vmatrix} 1 & x_1 .....x_1^k \\ 1 & x_2 .....x_2^k \\ ..... & \\ 1 & x_k .....x_k^k \end{vmatrix} \tag{7}$$

Для визначення $P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$ необхідно розв'язати систему рівнянь (2). Але існує більш раціональний шлях.

Розглянемо багаточлен $P_i(x) = \dfrac{(x-x_1)....(x-x_{i-1})(x-x_{i+1}).....(x-x_{n+1})}{(x_i-x_1)....(x_i-x_{i-1})(x_i-x_{i+1}).....(x_i-x_{n+1})}$,

$i = 1,2,....n+1$. При $x = x_i$, $P_i(x) = 1$, при $x = x_j \quad i \neq j$, $P_i(x_j) = 0$

Тому ми можемо представити $P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$ у наступному вигляді:

$$P(x) = \sum_{i=1}^{n+1} f(x_i) P_i(x) \tag{8}$$

Окрім інтерполяційного многочлена Лагранжа (8) використовується також метод розділених різниць з отриманням інтерполяційного многочлена Ньютона.

Як відомо, існує єдиний поліном ступеня не вище n, що приймає в точках x_i (i = 0, 1, 2, ..., n) задані значення. Тому можна покласти n = m. Коефіцієнти a_0, a_1, ..., a_n полінома Q_n (x) можна визначити з системи рівнянь:

$$\mathbf{XA} = \mathbf{Y}, \tag{9}$$

де

$$\mathbf{X} = \begin{pmatrix} 1 & x_0 & x_0^2 & ... & x_0^n \\ 1 & x_1 & x_1^2 & ... & x_1^n \\ ... & ... & ... & ... & ... \\ 1 & x_n & x_n^2 & ... & x_n^n \end{pmatrix}; \quad \mathbf{A} = \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_n \end{pmatrix}; \quad \mathbf{Y} = \begin{pmatrix} y_0 \\ y_1 \\ ... \\ y_n \end{pmatrix}; \tag{10}$$

$y_i = f(x_i)$ (i = 0, 1, 2, ..., n).

Визначник цієї системи лінійних алгебраїчних рівнянь є так званий визначник Вандермонда $\Delta \neq 0$, і, отже, система (10) має єдине рішення. Інтерполяція дає можливість обчислити значення функції $y = F(x)$ між заданими точками x_i - 1 і $x_i (i = 1, 2, ..., n)$.

Завдання № 3:

Виконати інтерполяцію функції, задану в табличній формі в п'яти точках (див. нижче). Розрахунки виконати в середовищі Python.

Вектори даних:

$$x := \begin{pmatrix} 0.1 \\ 0.3 \\ 0.4 \\ 0.6 \\ 0.7 \end{pmatrix} \qquad y := \begin{pmatrix} 3.2 \\ 3 \\ 1 \\ 1.8 \\ 1.9 \end{pmatrix}$$

Алгоритм розв’язку завдання № 3:

1. Заповнення матриці **X**;
2. Отримання коефіцієнтів інтерполяційного полінома;
3. Визначення функції полінома (прийняти поліном степеню 4);
4. Побудова графіка функції для інтерполюючого полінома;
5. Визначити значення функції в проміжних точках зі значеннями 0,2 і 0,5.

Для реалізації обчислювальних алгоритмів рекомендується використання онлайн середовищ тестування (наприклад repl.it, trinket, і.т.д.)

Захист лабораторної роботи передбачає виконання практичних завдань поставлених в роботі, та виконання завдань теоретичного характеру.

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