Въведение в Машинното обучение

Лабораторно упражнение №4

Регресия.

1. Цел на упражнението.

Целта на лабораторното упражнение е студентите да усвоят темата регресия, като решават задачи с помощта на регресионен анализ с една променлива.

2. Линейна регресия.

Линейната регресия е метод за моделиране на линейността между зависима y (резултативна) променлива и една или повече независими x (факторни) променливи.
Търси се уравнение на правата, която минава "най-близо" до точките от корелационното поле, т.е. най-добре отразява зависимостта между двете променливи. Критерий за "най-близо" – сборът от квадратите на разликите между емпиричните стойности y и техните оценки, които са ординатите на съответните точки от правата, да има минимум.

а) Общо уравнение за линейна регресия

$$\hat{y} = a + bx$$

За намирането на неизвестните коефициенти a и b се прилага методът на най-малките квадрати, при което се стига до системата

$$\begin{cases} \sum y = Na + b\sum x \\ \sum xy = a\sum x + b\sum x^2 \end{cases}$$

От нея се получава решението

$$b = \frac{\sum xy - N\bar{x}\bar{y}}{\sum x^2 - N\bar{x}^2}, \quad a = \bar{y} - b\bar{x}.$$

След определянето на коефициентите a и b се получава регресионният модел. Коефициентът b се нарича регресионен коефициент – той показва с колко единици се изменя зависимата променлива при изменение на факторната променлива с единица. Чрез регресионното уравнение могат да се получат оценките на y за всяка стойност на x:

Обобщаваща информация за големината на отклоненията на фактическите стойности от теоретично очакваните дава показателят стандартна грешка на оценката ($S_{Y/X}$), която се изчислява по формулата:

$$S_{Y/X} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

3. Изграждане на линейна регресия в Excel

Задача 1. В таблицата са дадени данни за резултатите от тест по математика и резултатите от тест за интелигентност на шестима студенти. Да се определи дали съществува зависимост между резултатите от теста по математика и резултатите от теста за интелигентност. (1 точка)

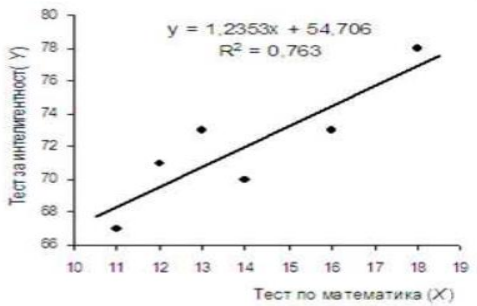
Създайте следната таблица в Excel.

Студент	Тест по математика (X)	Тест по интелигентност (Y)
A	11	67
B	13	73
C	18	78
D	12	71
E	16	73
F	14	70

1. Да се построи линеен регресионен модел.
2. Да се направи прогноза за резултата от теста за интелигентност, ако на теста по математика са получени 15 точки.

Резултати от тест по математика и тест за интелигентност

Студент	Тест по математика (X)	Тест за интелигентност (Y)	XY	X^2	Y^2
A	11	67	737	121	4489
B	13	73	949	169	5329
C	18	78	1404	324	6084
D	12	71	852	144	5041
E	16	73	1168	256	5329
F	14	70	980	196	4900
	ΣX=84	ΣY=432	ΣXY=6090	ΣX^2=1210	ΣY^2=31172
n = 6	$(\Sigma X)^2$=7056 $\bar{X}$ =14	$\bar{Y}$ =72			



$$\hat{y} = bx + a$$

$$b = \frac{n\sum XY - \sum X \cdot \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{6 \cdot 6090 - 84 \cdot 432}{6 \cdot 1210 - 7056} = 1{,}235$$

$$a = \bar{Y} - b\bar{X}$$

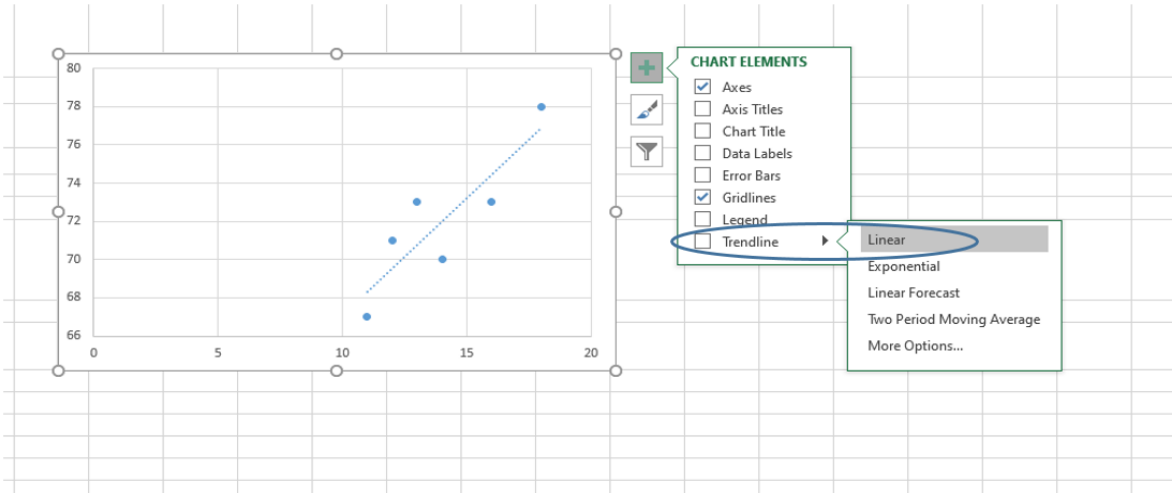
$$a = 72 - 1{,}235 \cdot 14 = 54{,}706$$

$$\hat{y} = 1{,}235x + 54{,}706$$

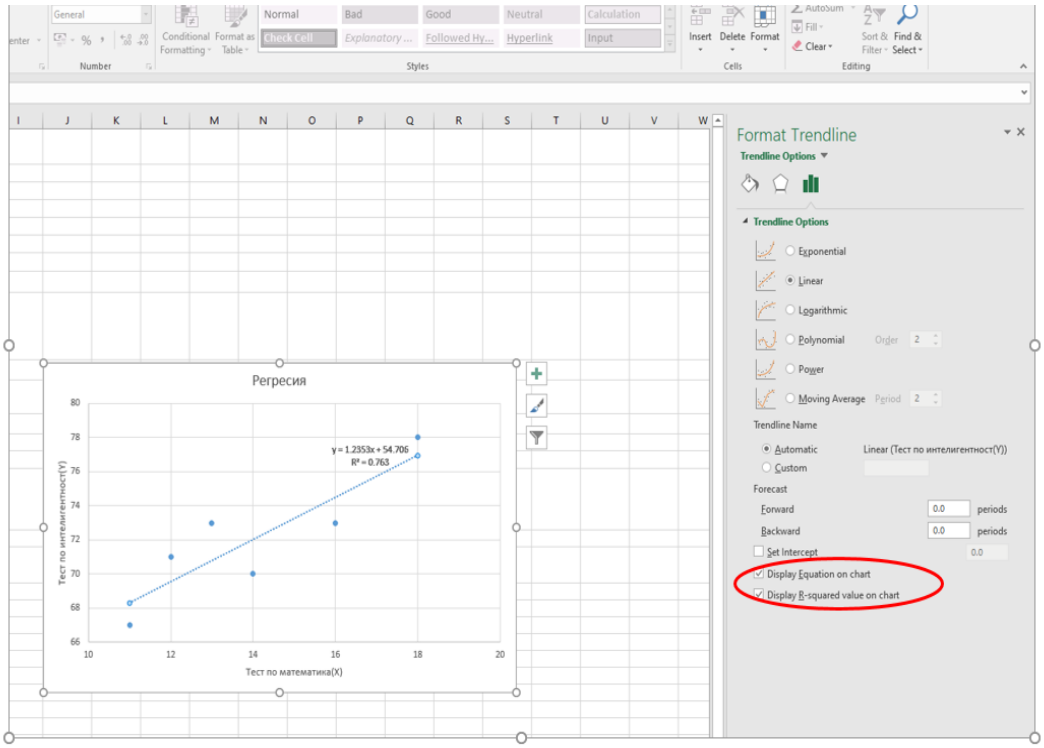
Предвиждане: Ако студент е получил на теста по математика 15 точки, то на теста за интелигентност се очаква да получи **1,235.15 + 54,706 = 73** точки.

За построяване на графиката данните се селектират от таблицата следните колоните →Тест по математика(X) и Тест по интелигентност(Y) и се избира Insert/Scatter.

- С десен бутон на мишката се маркира серията от данни в графиката след което се избира Add Trendline от контекстното меню.



- В прозореца Format Trendline въз основа на графичния вид на серията от данни потребителят избира най-подходящ вид на линия на тези, които Excel предлага. Поставят се отметки в Display Equation on Chart и Display R-squared value on chart.



Задача 2: В таблицата са дадени данни за посещаемостта на търговски център при различни температури на въздуха в различни дни от седмицата. Да се

определи дали съществува зависимост между температурата на въздуха и посещаемостта на търговски обект. (2 точки)

1. Да се построи линеен регресионен модел.
2. Да се направи прогноза за броя клиенти, ако температурата на въздуха е 14 градуса.

Дни в месеца	Температура(X)	Брой клиенти(Y)
1	6	52
2	8	54
3	10	66
4	8	60
5	22	98
6	18	80

4. Изграждане на линейна регресия с Pyhon

За да се изгради модел на линейна регресия в Python, може да се използва библиотеката scikit-learn (съкращение за "Scientific Kit for Learning"), която предлага различни модели за машинно обучение, включително модела на линейна регресия.

LinearRegression е клас в библиотеката scikit-learn за машинно обучение в Python, който се използва за изграждане на модели на линейна регресия.

LinearRegression моделът може да бъде използван за предсказване на стойности на целевата променлива, като се използват една или повече предиктори (feature-и) с линейни взаимоотношения спрямо целевата променлива. Моделът минимизира разликата между реалните стойности на целевата променлива и тези, които се предсказват от модела.

За да се използва LinearRegression в Python, първо трябва да се импортира от scikit-learn библиотеката:

from sklearn.linear_model import LinearRegression

След това се създава обект от тип LinearRegression:

model = LinearRegression()

Този обект може да се използва за трениране на модела чрез метода fit(), който приема два аргумента: X и y. X е матрицата с признаците (features) и y е масивът с целевите стойности:

model.fit(X, y)

След като моделът е обучен, може да се използва за предсказване на нови стойности чрез метода predict():

predictions = model.predict(X_new)

5. Методи на класа LinearRegression от библиотеката scikit-learn на Python

fit(X, y) е метод на класа LinearRegression от библиотеката scikit-learn, който се използва за обучение на линейни регресионни модели. Той приема два аргумента: X - матрица от признаци (features) и y - вектор от целевите стойности (target values). Методът използва X и y за да обучи модела на линейната регресия.

score(X, y) е метод на класа LinearRegression и се използва за измерване на качеството на модела на линейната регресия. Той приема два аргумента: X - матрица от признаци и y - вектор от целевите стойности. Методът връща коефициент на детерминацията (r^2), който представлява процента на вариацията във вектора от целеви стойности, който може да бъде обяснен от модела. Коефициентът на детерминация, също известен като R-squared (R^2), е статистическа мярка за оценка на качеството на модела на регресия. Той измерва колко добре линейната регресионна линия приближава реалните стойности на зависимата променлива. Коефициентът на детерминация е число между 0 и 1, като по-високата стойност означава по-добро приближение на модела към реалните стойности.

predict(X) е метод на класа LinearRegression, който се използва за предсказване на целевите стойности на нови наблюдения на базата на обучения модел. Той приема един аргумент - матрицата от признаци на новите наблюдения (X) и връща вектор от предсказани целеви стойности.

6. Атрибути на класа LinearRegression от библиотеката scikit-learn на Python

model.coef_ и **model.intercept_** са атрибути на класа LinearRegression в библиотеката scikit-learn. Когато използваме метода fit() на този клас, той се обучава върху дадения набор от данни и извлича параметрите на линейната регресия, които след това могат да се достъпят чрез тези атрибути.

coef е коефициентът на наклона на линията на регресия (slope), който отразява колко силно е свързана променливата x с променливата y. Например, ако коефициентът на наклона е положителен, това означава, че когато x нараства, y също ще нараства. Ако коефициентът на наклона е отрицателен, това означава, че когато x нараства, y ще намалява.

intercept е точката на пресичане на линията на регресия с y-оста, която отразява стойността на y, когато x е равно на нула.

Тези атрибути позволяват на потребителя да прогнозира стойността на зависимата променлива, като използва формулата на линейната регресия.

За повече информация:
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

7. Допълнителни пояснения към кода на задача 3

а) за редове 5, 6, 7 и 8

```
5. shopping_data = pd.DataFrame({
6. 'x': [5, 7, 10, 15, 20, 25],  # продължителност на престоя в минути
7. 'y': [8, 10, 13, 18, 22, 25]  # колко пари са вложени за покупки в магазина
8. })
```

DataFrame е основна структура от данни в библиотеката pandas в Python. Той се използва за съхранение на многомерни данни в табличен формат, като същевременно осигурява функционалности за манипулация на данните.

Колоните 'x' и 'y' на DataFrame съдържат поредица от числени данни, които представят продължителността на престоя в минути ('x') и сумата от покупките в долари ('y') на клиенти в магазин По този начин данните са представени в табличен формат, който може да бъде обработен с помощта на библиотеката pandas.

б) За редове 10 и 11

```
10. X = shopping_data.iloc[:, :-1].values
11. y = shopping_data.iloc[:, 1].values
```

"iloc" е метод на pandas DataFrame, който се използва за индексиране и извличане на данни по числовите им индекси в таблицата.
В конкретния код "shopping_data.iloc[:, :-1]" извлича всички редове на DataFrame "shopping_data", като взема всички колони, освен последната. Изразът ":, :-1" означава, че се вземат всички редове на DataFrame-а, а с ":" в първата позиция се указва, че не се прилага филтър на редовете, а с ":-1" във втората позиция се указва, че не се взима последната колона.

След като се извлекат нужните данни от DataFrame, методът "values" преобразува данните в NumPy масив, за да могат да бъдат използвани в машинното обучение.

Изразът "-1" се използва за указване на последната колона в DataFrame. Това е така, защото в този код последната колона съдържа целевата променлива (target variable), която не трябва да бъде включена в характеристиките на модела при обучението. Така, с изключването на последната колона, "X" съдържа само характеристики (features) за обучението на модела.

в) за ред 13

13. *plt.scatter(X, y, color='blue')*

Функцията scatter() от библиотеката matplotlib.pyplot, изобразява графика на разсейване (scatter plot) на данните. Тази функция приема два аргумента - X и y, които са масиви (arrays) или списъци (lists) от еднакъв размер, представляващи съответно стойностите на признака (feature) и целевата променлива (target variable) в линейната регресия.

Цветът на точките в графиката се задава чрез параметъра color, като в този случай е избран синьо (blue). В резултат на това, функцията scatter() изобразява всеки наблюдения като точка на координатната система, където по оста x е поставена стойността на признака, а по оста y е поставена стойността на целевата променлива.

г) за ред 39

39. *print(f'За престой от 12 минути се очаква харченето да е {new_y[0]:.2f} лева.')*

Този код използва f-стрингове (f-strings), за да изведе на конзолата съобщение, което съдържа текст и стойност на променлива, форматирана с определен брой знака след десетичната запетая.

f-стринговете в Python 3.6 и по-нови версии се означават със символа "f" или "F" пред низа. В този случай, f-стрингът започва с "f" и съдържа текстовия низ "За престой от 12 минути се очаква харченето да е ", следван от вмъкване на стойността на променливата "new_y[0]" в скобите с фигурни скоби - "{new_y[0]}".

След като е изведена стойността на променливата, ":.2f" указва, че искаме да се изведат два знака след десетичната запетая. Това означава, че стойността ще бъде закръглена до два знака след запетаята.

В крайна сметка, изведеният текст на конзолата ще включва съобщението "За престой от 12 минути се очаква харченето да е " и стойността на променливата "new_y[0]", която е закръглена до два знака след запетаята.

Задача 3: Дадени са данни за времето за престой на клиенти в магазина и похарчените суми за покупки. Да се построи линеен регресионен модел в Python. Да се изчисли коефициента на детерминация. Да се изведат параметрите на линейната регресия - наклон и свободен член. Да се построи права на регресията. Да се прогнозира колко пари ще похарчи клиент при престой в магазина от 12 минути.

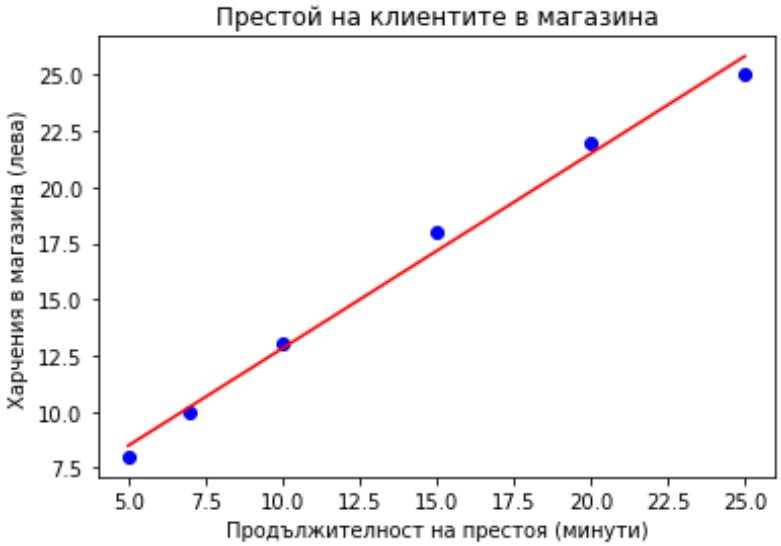
1. *import pandas as pd*
2. *import matplotlib.pyplot as plt*

```
3. from sklearn.linear_model import LinearRegression

4. # Данните за престоя на клиентите в магазина
5. shopping_data = pd.DataFrame({
6. 'x': [5, 7, 10, 15, 20, 25],  # продължителност на престоя в минути
7. 'y': [8, 10, 13, 18, 22, 25]  # колко пари са вложени за покупки в магазина
8. })

9. # Разделяне на данните на X (продължителност на престоя) и y (суми вложени за покупки в магазина)
10.X = shopping_data.iloc[:, :-1].values
11.y = shopping_data.iloc[:, 1].values

12.# Изобразяване на данните в графика
13.plt.scatter(X, y, color='blue')
14.plt.title('Престой на клиентите в магазина')
15.plt.xlabel('Продължителност на престоя (минути)')
16.plt.ylabel('Харчения в магазина (лева)')
17.plt.show()

18.# Създаване на модел за линейна регресия
19.model = LinearRegression()
20.# Обучение на модела с данните
21.model.fit(X, y)

22.# Извеждане параметрите на линейната регресия - наклон и свободен член
23.print('Наклон на линията:', model.coef_)
24.print('Свободен член:', model.intercept_)

25.# Изчисляване на коефициента на детерминация R^2
26.r2 = model.score(X, y)
27.print(f'R^2 = {r2:.2f}')

28.# Построяване на правата на регресия
29.line_y = model.predict(X)
30.plt.scatter(X, y, color='blue')
31.plt.plot(X, line_y, color='red')
32.plt.title('Престой на клиентите в магазина')
33.plt.xlabel('Продължителност на престоя (минути)')
34.plt.ylabel('Харчения в магазина (лева)')
35.plt.show()
```

```
36.# Предсказване на харченията за престой от 12 минути
37.new_X = [[12]]
38.new_y = model.predict(new_X)
39.print(f'За престой от 12 минути се очаква харченето да е {new_y[0]:.2f} лева.')
```

Отговор:

Наклон на линията: [0.86703297]
Свободен член: 4.150549450549457
R^2 = 0.99
За престой от 12 минути се очаква харченето да е 14.55 лева.



Задача 4: В таблицата са дадени данни за цената на имоти в евро в зависимост от площта в квадратни метри.

Area	45	60	65	70	80	100
Price	50000	80000	92000	99000	110000	160000

Да се построи линеен регресионен модел в Python. Да се изчисли коефициента на детерминация. Да се изведат параметрите на линейната регресия - наклон и свободен член. Да се построи права на регресията. Да се прогнозира каква ще е цената на имота при площ от 75 квадратни метра.

а) Данните да се задават в началото на файла (като редове от 4 до 8 в задача 3) (1 точка)

б) Данните да се заредят от файл с разширение .csv (1 точка).

Пояснения към подточка б).

- Таблицата се записва в Excel и се избира тип CSV (разделен със запетаи) (*.csv)
- редове от 4 до 8 от задача 3 се заменят примерно с:
  imoti_data = pd.read_csv('houses_prices.csv', sep=';')
  или

imoti_data = pd.read_csv('houses_prices.csv', sep=',')

- редове 10 и 11 от задача 3 са заменят примерно с:
  X = imoti_data['Area'].values.reshape(-1, 1)
  y = imoti_data['Price'].values