ЖИТОМИРСЬКА ПОЛІТЕХНІКА
ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ЛІНІЙНА РЕГРЕСІЯ

Лекція 6

ПЛАН

1. Машинне навчання
2. Що таке лінійна регресія?
3. Математичні основи та формули
4. Алгоритми навчання (МНК, градієнтний спуск)
5. Схема проведення лінійної регресії
6. Види регресій
7. Метрики оцінки якості
8. Практичне застосування

МАШИННЕ НАВЧАННЯ

Машинне навчання (МН, Machine Learning, ML) - великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися.

Машинне навчання - це «*це сфера штучного інтелекту, зосереджена на розробці алгоритмів, які самостійно вдосконалюються шляхом аналізу накопичених даних*».

Машинне навчання (МН) – це підрозділ штучного інтелекту, який розглядає побудову алгоритмів, які можуть навчатися на наявних даних.

Задача МН виглядає наступним чином: уявімо собі, що в нас є певний набір об'єктів прикладів і певний набір міток, тобто, реакцій, відповідей. Між прикладами/спостереженнями і відповідями є певна прихована залежність.

Задача МН – знайти цю приховану залежність для прогнозування відповідей на основі нових даних.

У найзагальнішому випадку розрізняють два типи машинного навчання: навчання по прецедентах, або **індуктивне навчання, і дедуктивне навчання**.

Індуктивне навчання знайоме кожному, адже воно полягає у спостереженні за світом та побудові певних моделей, які пояснюють причини тих чи інших явищ. Потім такі моделі неодноразово перевіряються, певні з них «виживають» і використовуються, покращуються. А деякі моделі згодом цілком відкидаються.

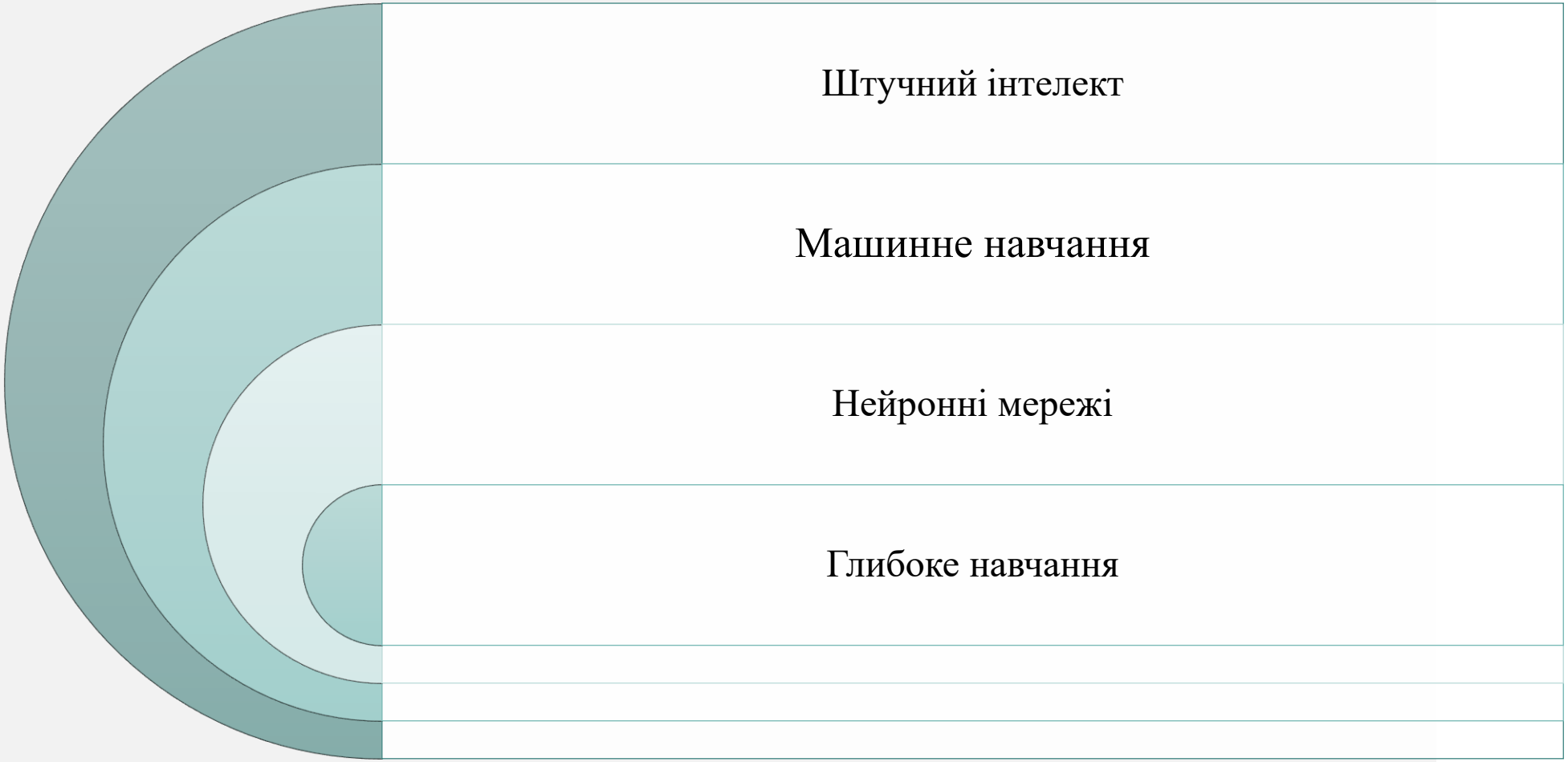
Дедуктивне навчання подібне до математики в школі, коли учню дають готові формули і розказують, як застосовувати їх на практиці

Навчання по прецедентах (індуктивне), в свою чергу, поділяють на три основних типи:

контрольоване навчання, або навчання з учителем (supervised learning),

неконтрольоване навчання (unsupervised learning), або навчання без учителя,

навчання з підкріпленням (reinforcement learning).


Штучний інтелект
Машинне навчання
Нейронні мережі
Глибоке навчання

Вираховувати шахрайство з банківськими картками

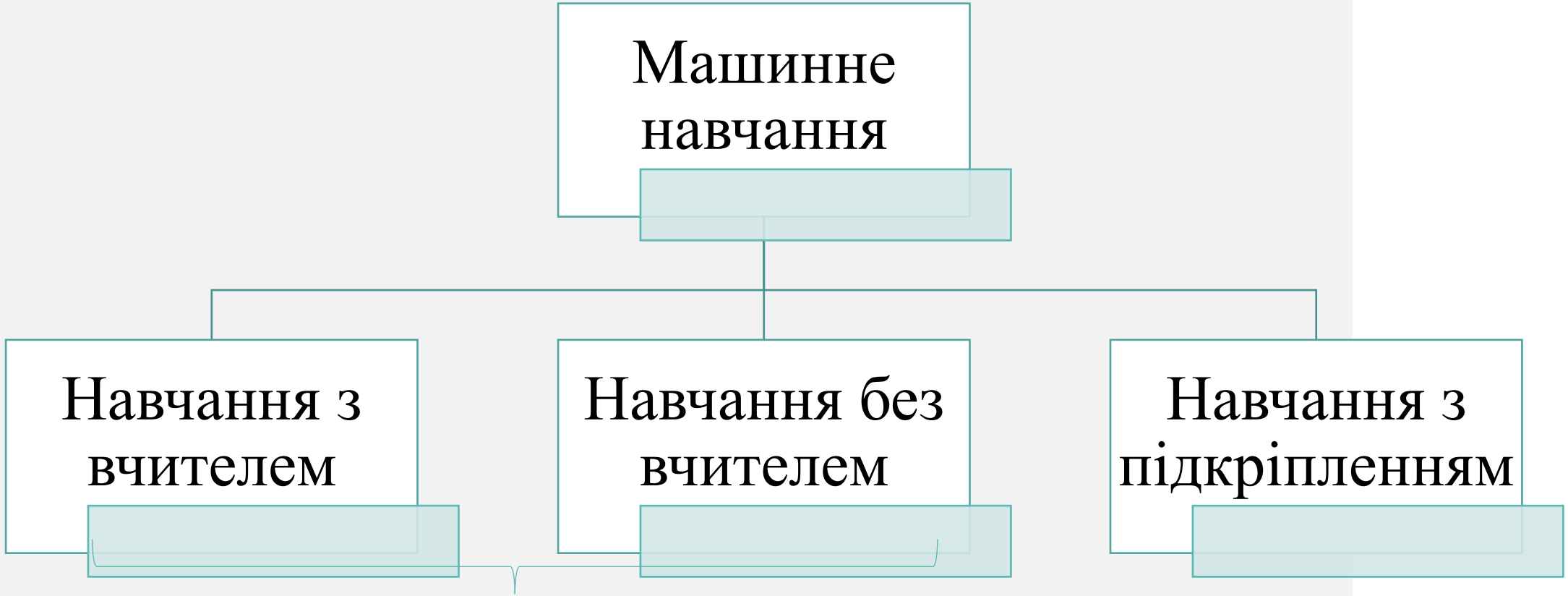
Моделювати ризики для інвестицій або кредитування

Робити фінансові прогнози

Сегментувати клієнтів

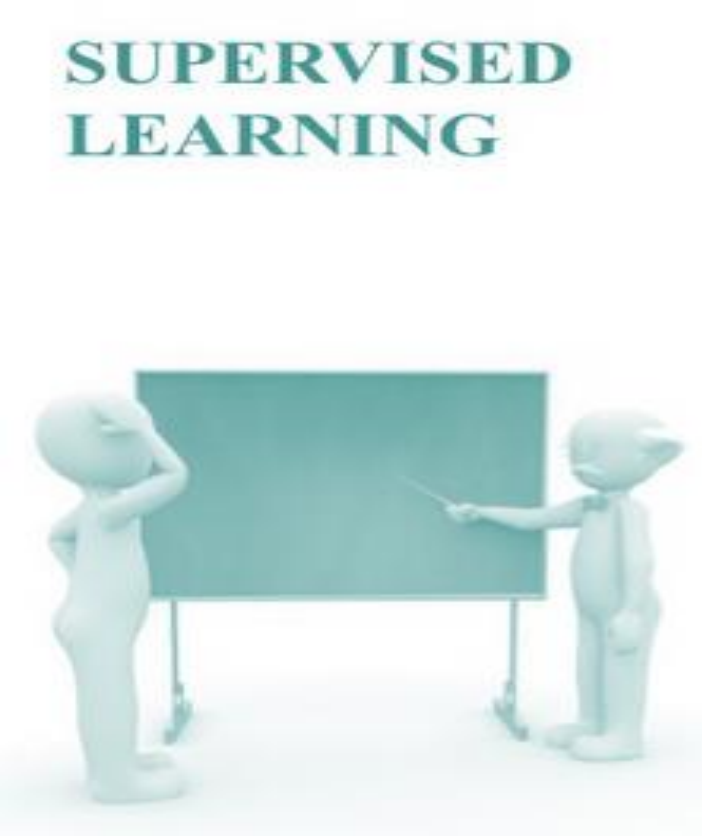
Створювати системи рекомендацій

Основні типи машинного навчання



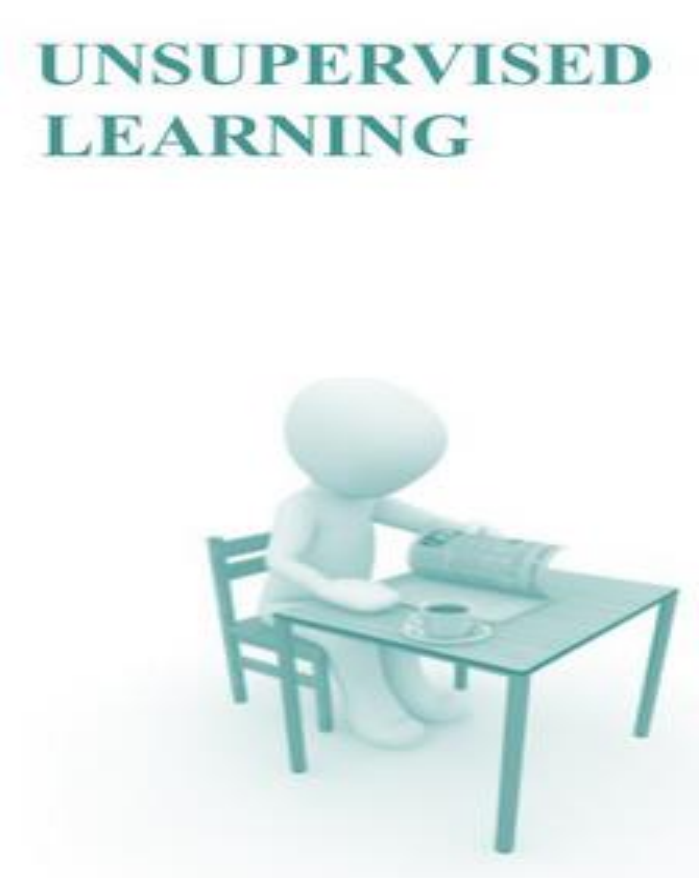
Навчання з вчителем

Є набір прикладів, до кожного прикладу є правильна відповідь. Задача системи – навчитися по прикладах надавати правильну відповідь, задану вчителем. Вчителями є ми.



Навчання без вчителя

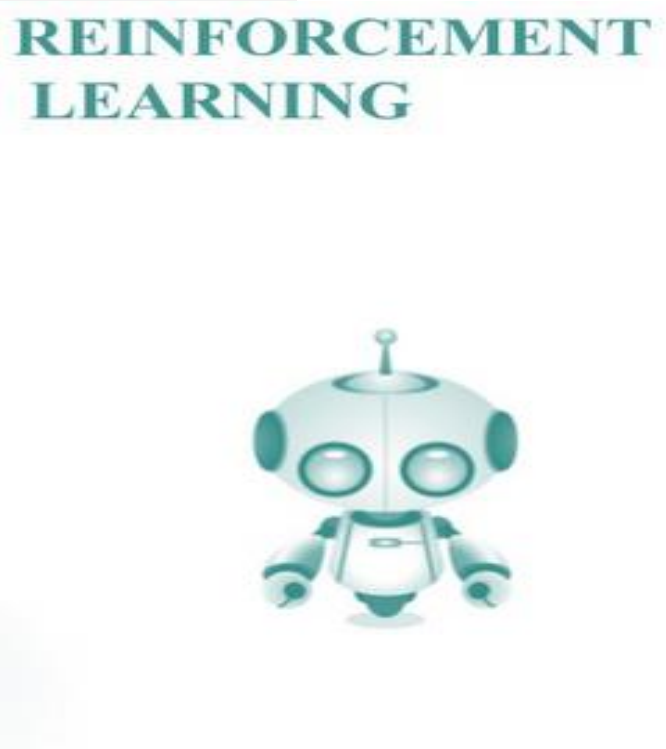
Є великий набір даних. В цих даних є приховані закономірності. Задача системи – знайти закономірності, наприклад, розбивши дані на певні групи чи кластери.



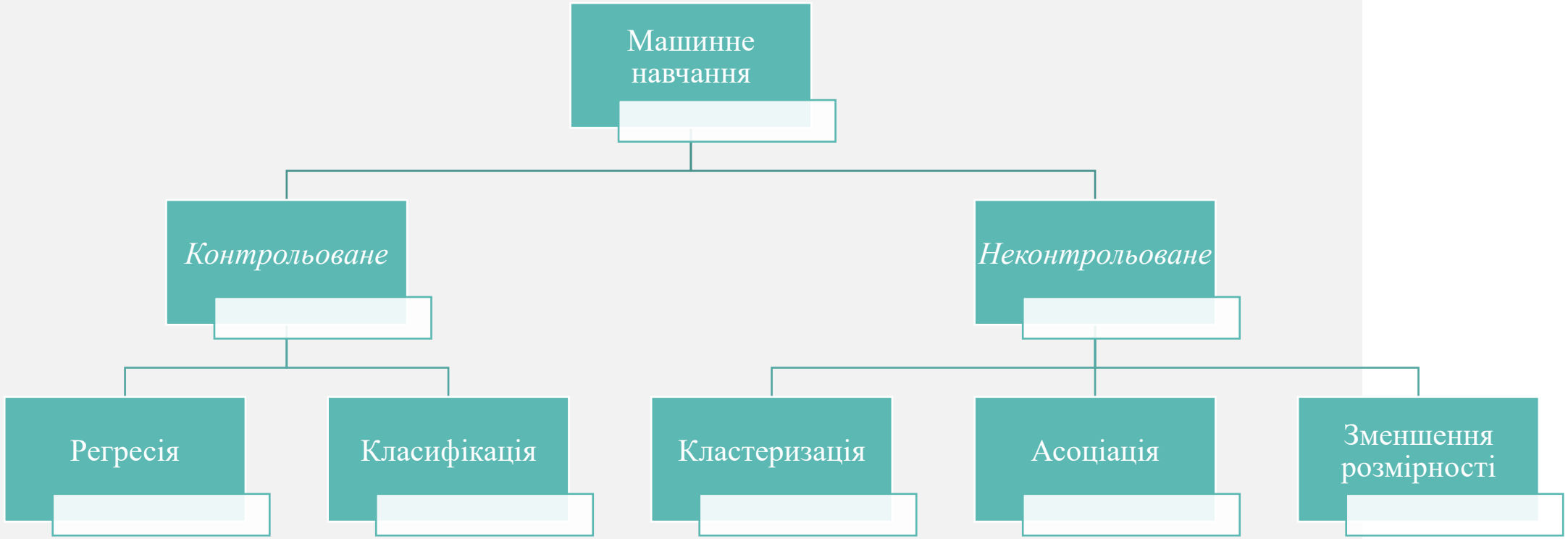
Навчання з підкріпленням

Є певне середовище, в якому є певний агент, що контролюється комп’ютером. Агент може вчиняти певні дії. Певні дії приводять до позитивних відкликів чи негативних відкликів. Задача – максимізувати позитивні і мінімізувати негативні відклики.

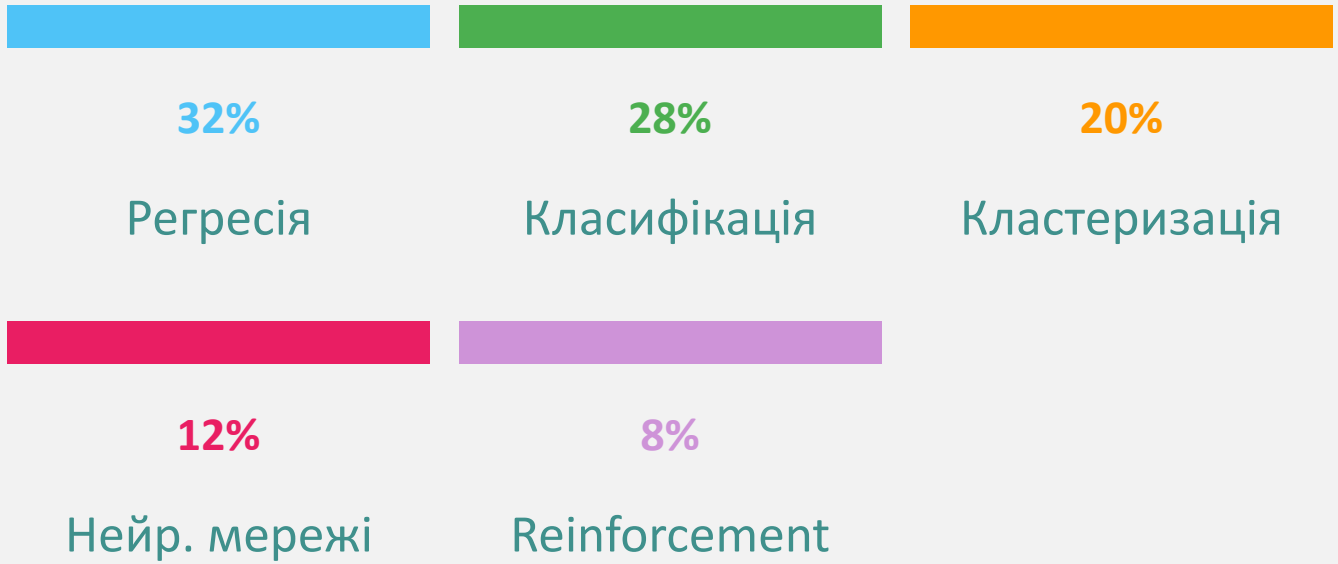
Приклад: гра, в якій треба максимізувати набрані бали, або виграти усю гру



Задачі МН класифікуються ще на декілька типів по виду вирішуваної проблеми:



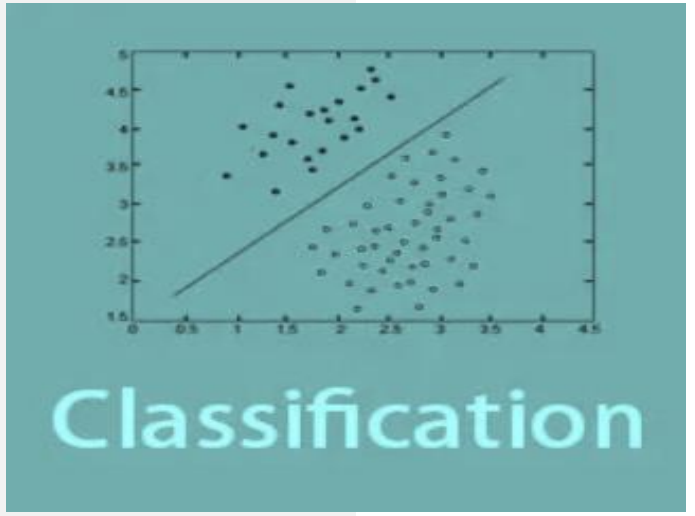
Класифікація задач та місце регресії



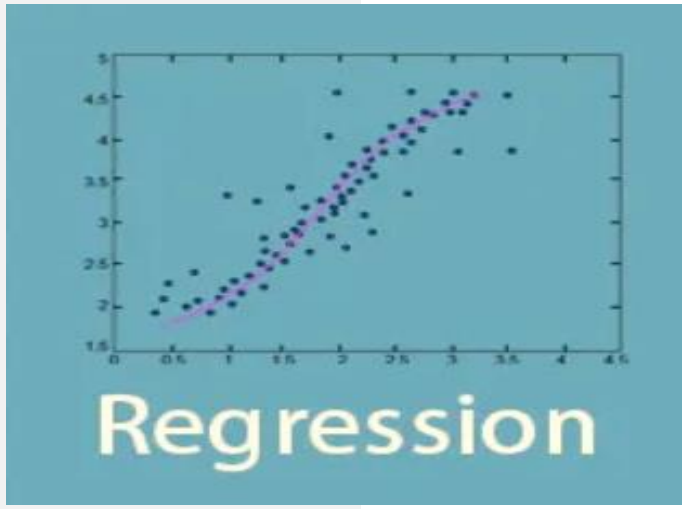
Регресія займає ~32% усіх задач у машинному навчанні - найбільша категорія серед supervised learning методів.

Задачі

Класифікація - передбачення категорії об'єкта

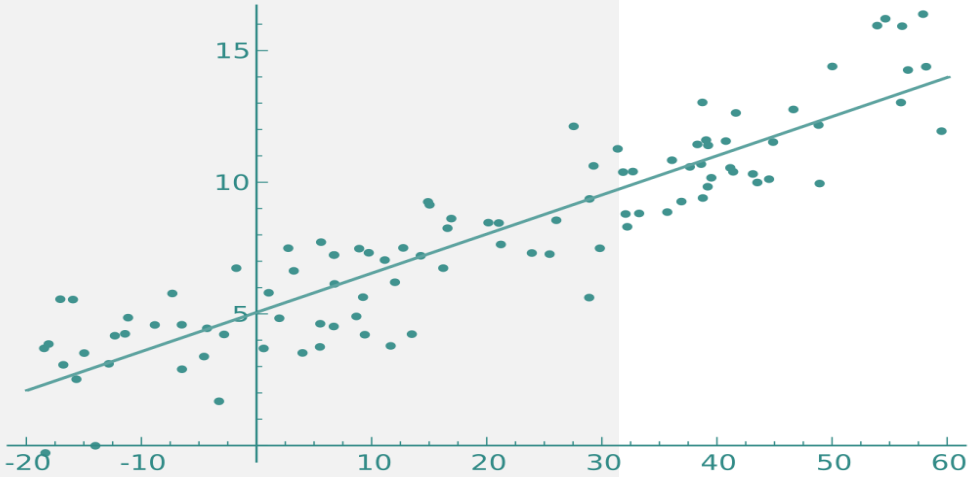


Регресія - передбачення місця на числовій прямій.



РЕГРЕСІЯ

Регресія служить для визначення виду зв'язку між змінними і дає можливість для прогнозування значення однієї (залежної) змінної, відштовхуючись від значень інших (незалежних) змінних.



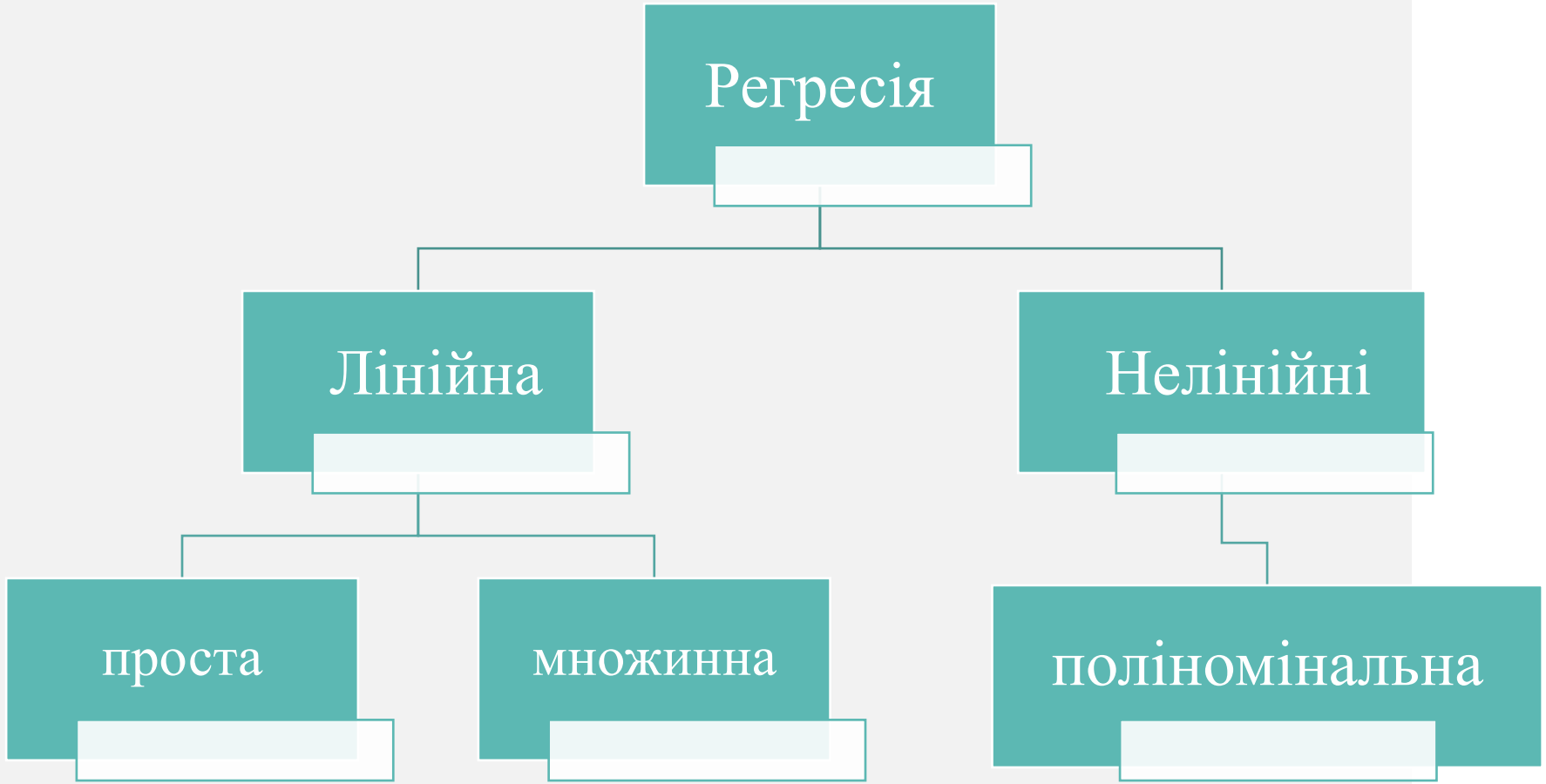
Регресію використовують для:

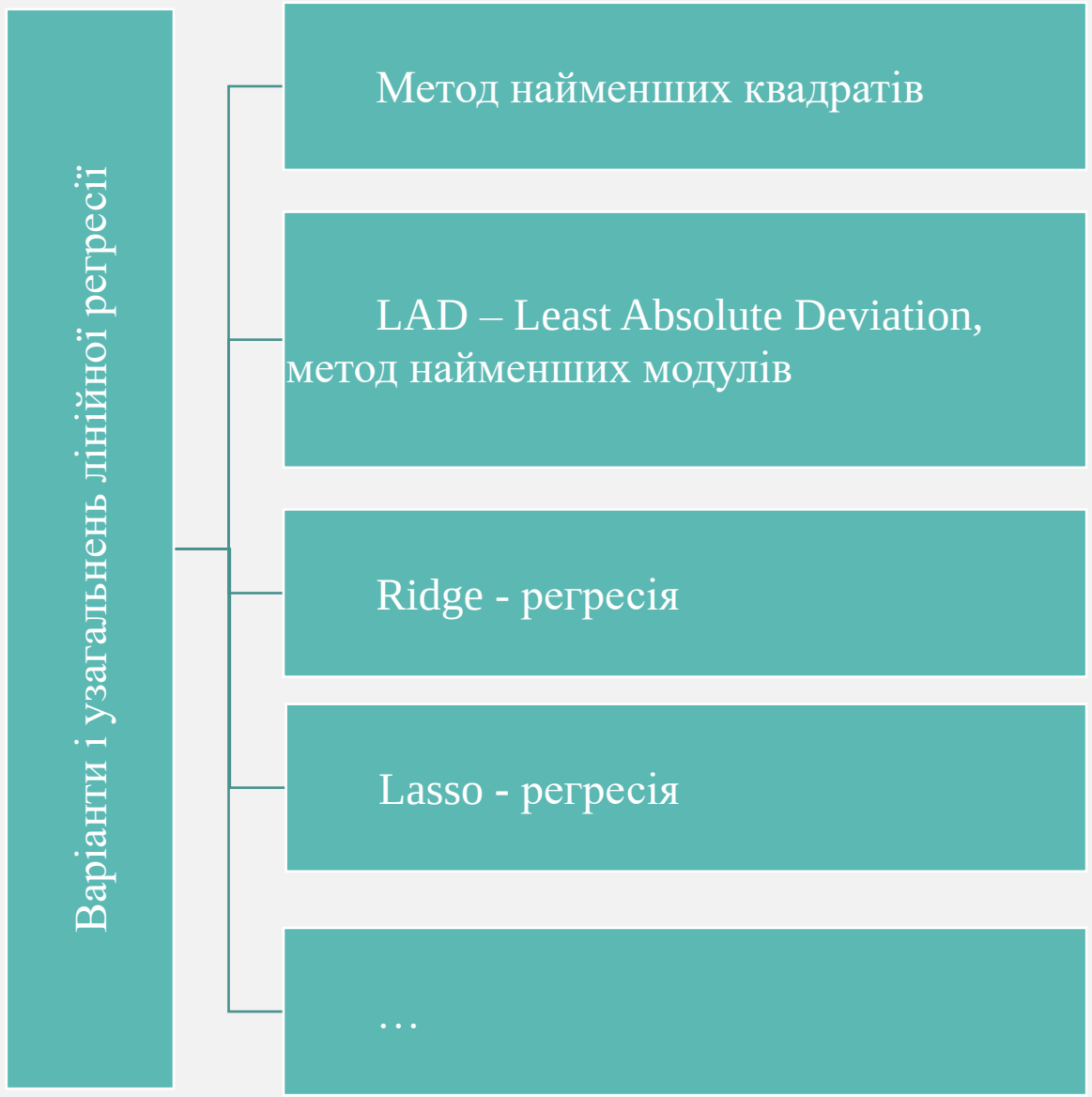
- Прогнозу вартості цінних паперів
- Аналізу попиту, обсягу продажів
- Встановлення медичних діагнозів
- Встановлення будь-якої залежності числа від часу

Регресію використовують у:

- Економіці
- Менеджменті
- Психології

Види


Регресія
Лінійна
Нелінійні
проста
множинна
поліноміальна


Варіанти і узагальнень лінійної регресії
Метод найменших квадратів
LAD – Least Absolute Deviation, метод найменших модулів
Ridge - регресія
Lasso - регресія
…

Види

Проста лінійна

$$\hat{y} = wx + b$$

Застосування: 1 ознака → 1 вихід

Приклад: Ціна vs площа

Множинна лінійна

$$\hat{y} = w_1x_1 + ... + w_nx_n + b$$

Застосування: n ознак → 1 вихід

Приклад: Ціна vs площа, к-сть кімнат, поверх...

Поліноміальна

$$\hat{y} = w_0 + w_1x + w_2x^2 + w_3x^3$$

Застосування: Нелінійна залежність

Приклад: Траєкторія руху, зріст дитини

Ridge (L2)

$$L = MSE + \lambda\|w\|^2$$

Застосування: Регуляризація (мультикол.)

Приклад: Генетичні дані, тексти

Lasso (L1)

$$L = MSE + \lambda\|w\|_1$$

Застосування: Feature selection

Приклад: Відбір важливих ознак

Elastic Net

$$L = MSE + \lambda_1\|w\|_1 + \lambda_2\|w\|^2$$

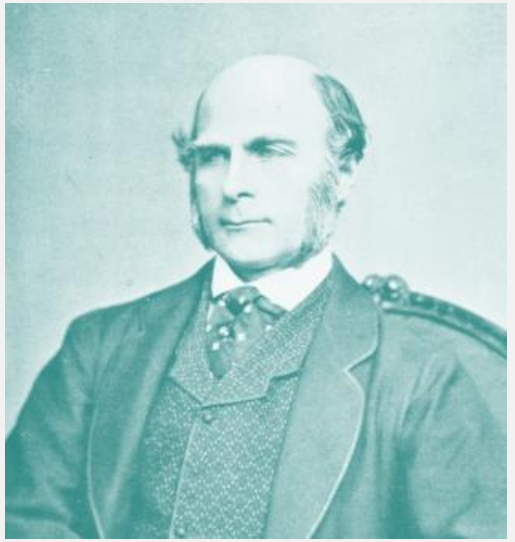
Застосування: Комбінація L1+L2

Приклад: Фінансові дані

Лінійна регресія - «найстаріший» тип, що з'явився два з половиною століття тому.

Вперше метод найменших квадратів опублікував **Адрієн Марі Лежандр** в 1805 році, хоча Гаусс прийшов до нього раніше і успішно використовував для передбачення орбіти «комети» Церери.

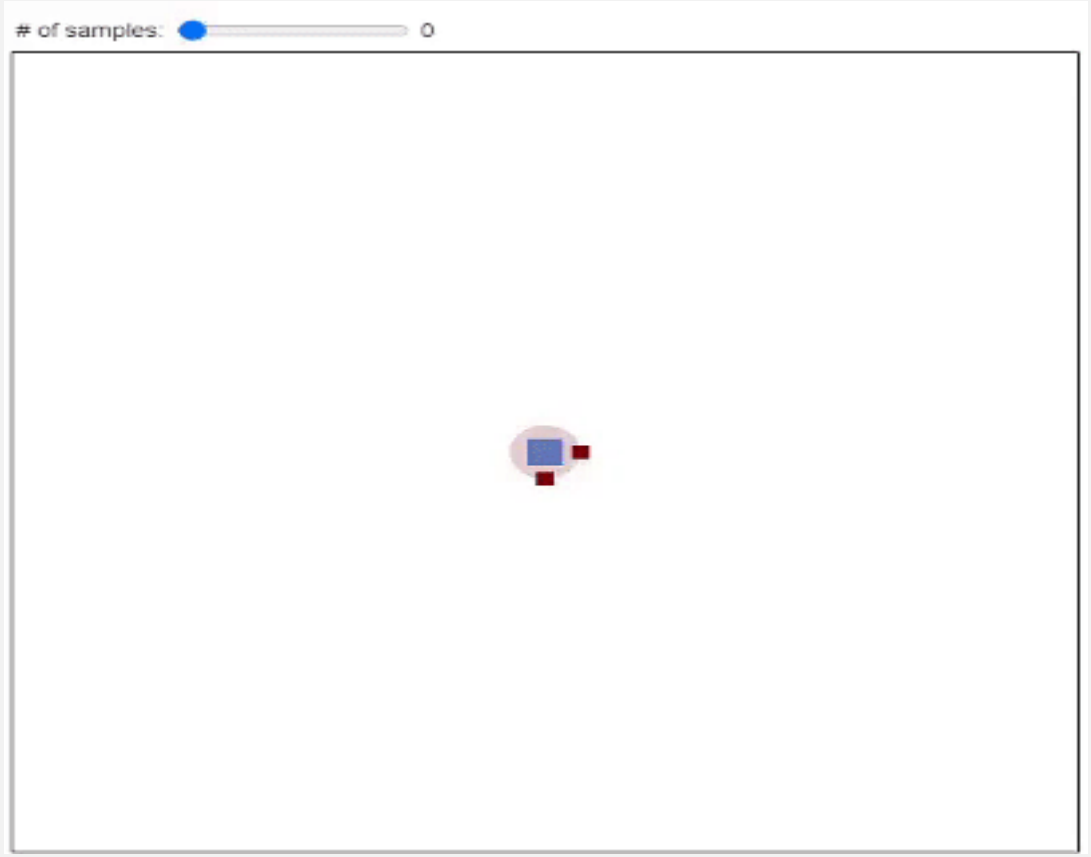
Карл Фрідріх Гаус (1809 рік)



Френсіс Гальтон (1886 рік)

Проста (Парна) лінійна регресія - це модель, що дозволяє моделювати взаємозв'язок у двовимірному просторі вибірки, утвореному однією незалежною змінною та однією залежною змінно (зазвичай x і y — координати у декартовій системі координат).

Модель призначена для знаходження лінійної функції залежності, яка якомога точніше прогнозує значення залежної змінної як функції незалежної змінної.



GoogleColab - https://colab.research.google.com/github/fbeilstein/machine_learning/blob/master/workbook_09_linear_regression.ipynb#scrollTo=26_N4hKfD-Fe

ЛІНІЙНА РЕГРЕСІЯ

Якщо коефіцієнт кореляції дає розуміння чи є лінійна залежність між двома змінними, то лінійна регресія дає модель для оцінки як зміниться одна змінна при зміні іншої.

ЩО ТАКЕ ЛІНІЙНА РЕГРЕСІЯ?

Лінійна регресія — метод supervised learning, що моделює лінійну залежність між:

Вхідні змінні (X)

Незалежні ознаки або предиктори

Вихідна змінна (y)

Цільова неперервна величина

Параметри моделі

Ваги w та зміщення b

Проста

1 предиктор

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

Множинна

k предикторів

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$

Ціль

Мінімізація похибки (MSE, MAE)

Результат

Неперервне числове значення

Проста лінійна регресія

Лінійною регресією назвемо регресію виду:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

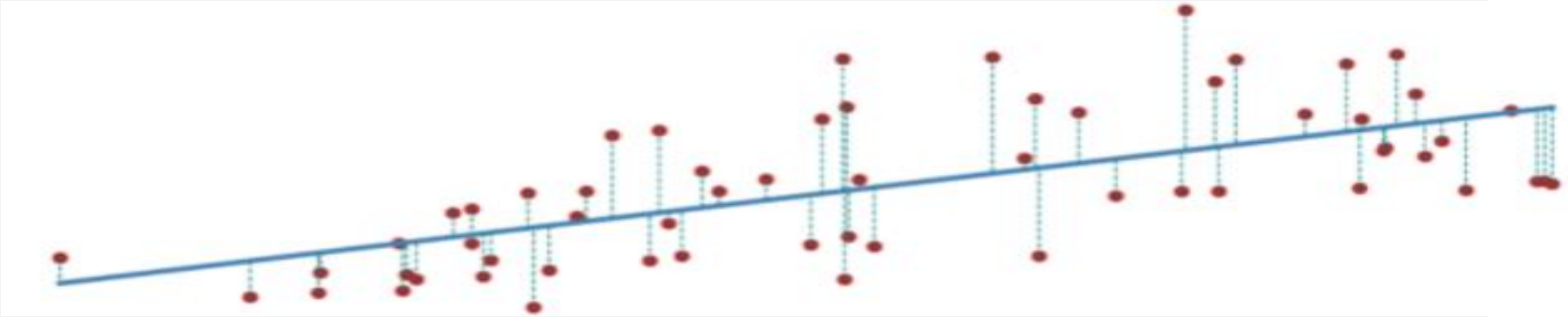
У випадку парної (простої) регресії вираз для лінійної регресії набуває вигляду:

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

Де β_0 зсув по осі ординат, $\beta 1$ - регресійні коефіцієнти, ε - випадкова помилка змінної Y в і-м спостереженні.

β_0 і β_1 називаються модальними коефіцієнтами.

Щоб створити модель, необхідно дізнатися значення цих коефіцієнтів. І як тільки ці коефіцієнти знайдені, можна використовувати модель для прогнозування.



АЛГОРИТМИ НАВЧАННЯ

МНК — Метод найменших квадратів

1. Записати матрицю ознак X (n×p)
2. Обчислити X^TX та X^Ty
3. Розв'язати систему: $w = (X^TX)^{-1}X^Ty$
4. Ваги w — оптимальні коефіцієнти
5. Передбачити: $\hat{y} = Xw$

⊕ Точний результат за одну ітерацію
⊖ Складність $O(p^3)$ — повільно при p>>1000

Градієнтний спуск (Gradient Descent)

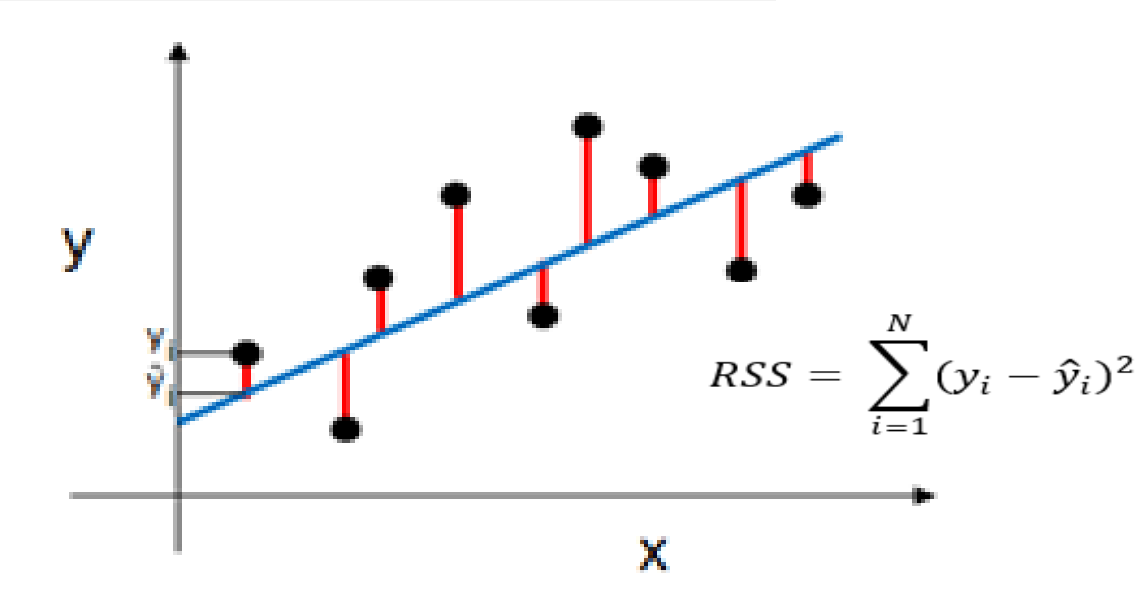
1. Ініціалізувати w = 0 (або рандом)
2. Обчислити градієнт $\partial L/\partial w$
3. Оновити: $w \leftarrow w - \alpha \cdot \partial L/\partial w$
4. Повторити кроки 2-3
5. Зупинитись при збіжності

⊕ Масштабується на великі дані
⊖ Потребує підбору learning rate α

МНК — Метод найменших квадратів

Оцінка ("навчання") модальних коефіцієнтів

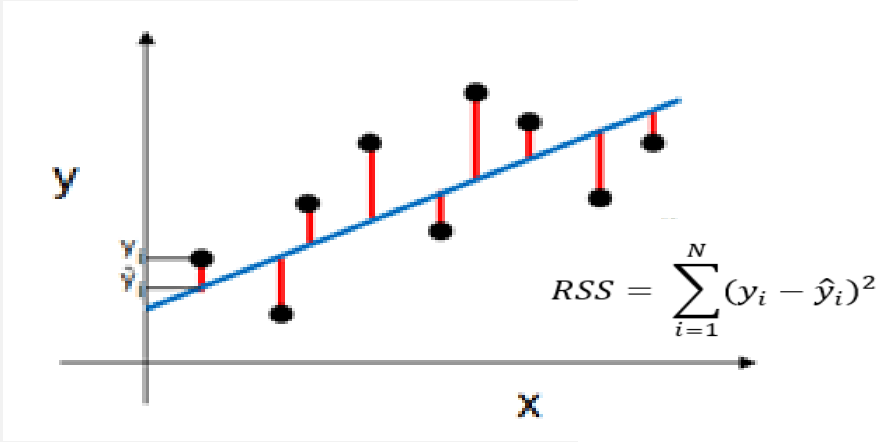
Коефіцієнти оцінюються з використанням методу найменших квадратів, що означає, що необхідно знайти лінію (математично), яка мінімізує суму квадратів помилок (англ. Residual Sum of Squares (RSS)



RSS дозволяє визначити рівень ***варіативності*** даних на відстані від лінії.

Лінія найкращої відповідності має найменше значення RSS.

Математичний метод, заснований на мінімізації суми квадратів відхилень деяких функцій від шуканих змінних, називається ***методом найменших квадратів*** (МНК, англ. Ordinary Least Squares (OLS)).



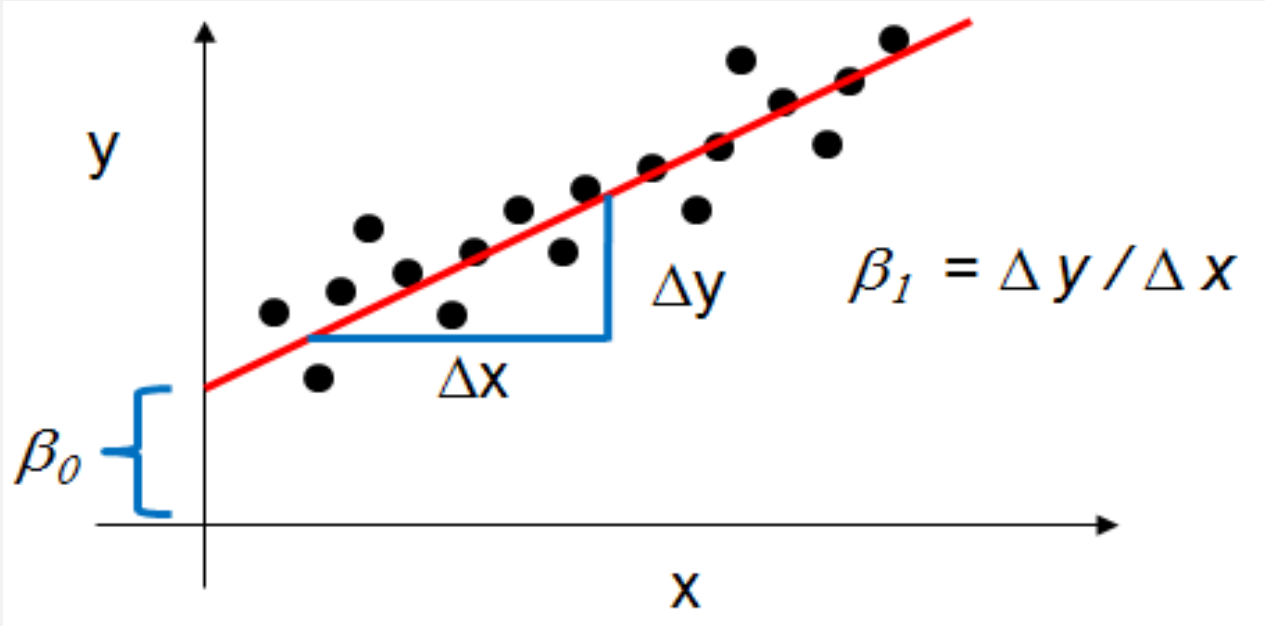
$$\sum (y_i - (\beta_0 + \beta_1 x_i))^2 \rightarrow \min$$

Як модальні коефіцієнти відносяться до лінії найменших квадратів?

β_0 *є 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

$$
\begin{aligned}
&S(\beta 0,\beta 1)= \\
&[6-(\beta 0+1\beta 1)]^2+ \\
&[5-(\beta 0+2\beta 1)]^2+ \\
&[7-(\beta 0+3\beta 1)]^2+ \\
&[10-(\beta 0+4\beta 1)]^2.
\end{aligned}
$$

β0'

$[6-(\beta0+\beta1)]^2+[5-(\beta0+2\beta1)]^2+[7-(\beta0+3\beta1)]^2+[10-(\beta0+4\beta1)]^2=$

$2[6-(\beta0+\beta1)]+2[5-(\beta0+2\beta1)]+2[7-(\beta0+3\beta1)]+2[10-(\beta0+4\beta1)]=$

$12-2(\beta0+\beta1)+10-2(\beta0+2\beta1)]+14-2(\beta0+3\beta1)+20-2(\beta0+4\beta1)=$

$56-2\beta0-2\beta1-2\beta0-4\beta1-2\beta0-6\beta1-2\beta0-8\beta1=$

$56-8\beta0-20\beta1$

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

β1‘

$[6-(\beta0+\beta1)]^2+[5-(\beta0+2\beta1)]^2+[7-(\beta0+3\beta1)]^2+[10-(\beta0+4\beta1)]^2=$

$2[6-(\beta0+\beta1)]+2*2[5-(\beta0+2\beta1)]+2*3[7-(\beta0+3\beta1)]+2*4[10-(\beta0+4\beta1)]=$

$12-2\beta0-2\beta1+20-4\beta0-8\beta1+42-6\beta0-18\beta1+80-8\beta0+32\beta1=$

$154-20\beta0-60\beta1$

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

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.

В результаті рішення системи з двох рівнянь отримуємо

$$\beta 0 = 3.5$$

$$\beta 1 = 1.4$$

Рівняння лінії : $y = 3.5 + 1.4x$

Мінімальна сума квадратів похибок:

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

```python
x = np.array([1, 2, 3, 4])
y = np.array([6, 5, 7, 10])

# Створюємо матрицю X, де перший стовпець - одиниці (для b0), другий - значення x (для b1)
X = np.vstack([np.ones(len(x)), x]).T

# Розв'язуємо систему методом найменших квадратів
# np.linalg.lstsq знаходить рішення системи
beta_0, beta_1 = np.linalg.lstsq(X, y, rcond=None)[0]

print(f"Коефіцієнт бета_0 : {beta_0:.2f}")
print(f"Коефіцієнт бета_1 : {beta_1:.2f}")
print(f"Рівняння прямої: y = {beta_0:.2f} + {beta_1:.2f}x")

plt.figure(figsize=(8, 5))
plt.scatter(x, y, color='red', label='Вхідні дані')
plt.plot(x, beta_0 + beta_1*x, color='#0D9488', label='Лінія регресії')

plt.title('Лінійна регресія методом найменших квадратів')
plt.xlabel('x')
plt.ylabel('y')
plt.legend()
plt.grid(True, linestyle='--', alpha=0.7)
plt.show()
```

```
Коефіцієнт бета0: 3.50
Коефіцієнт бета1: 1.40
Рівняння прямої: y = 3.50 + 1.40x
```

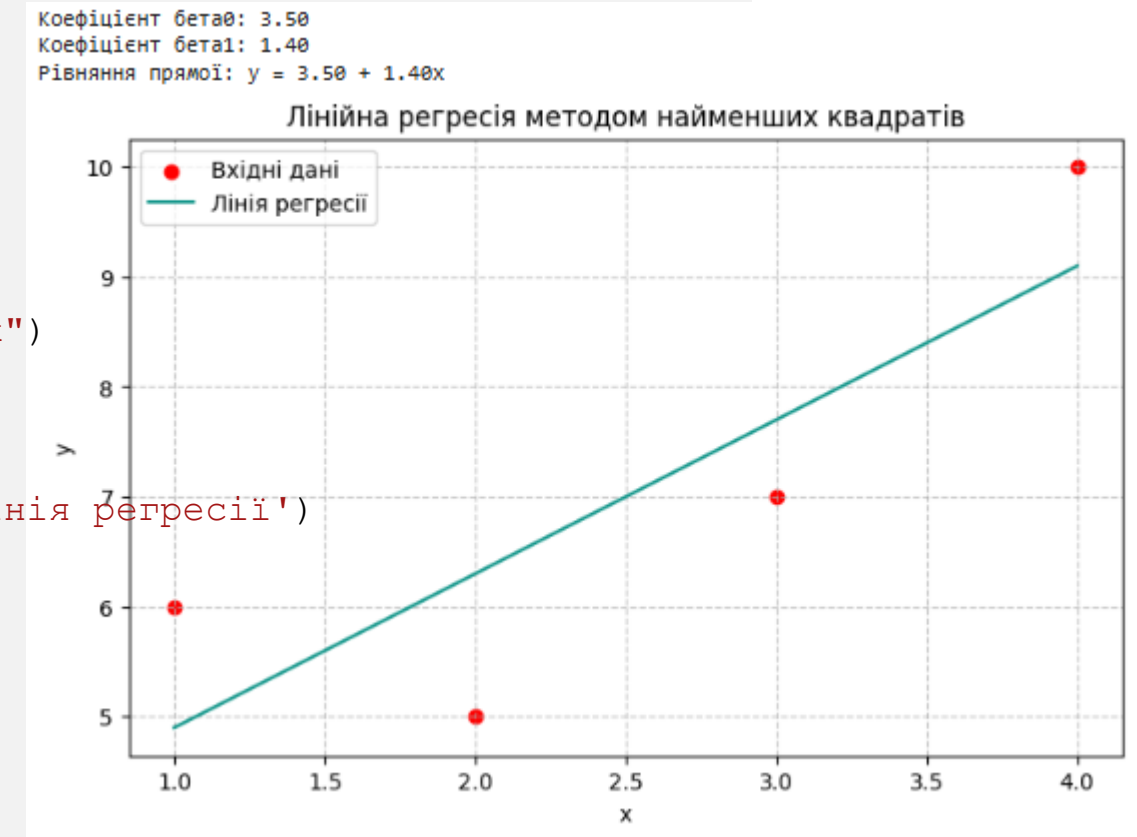
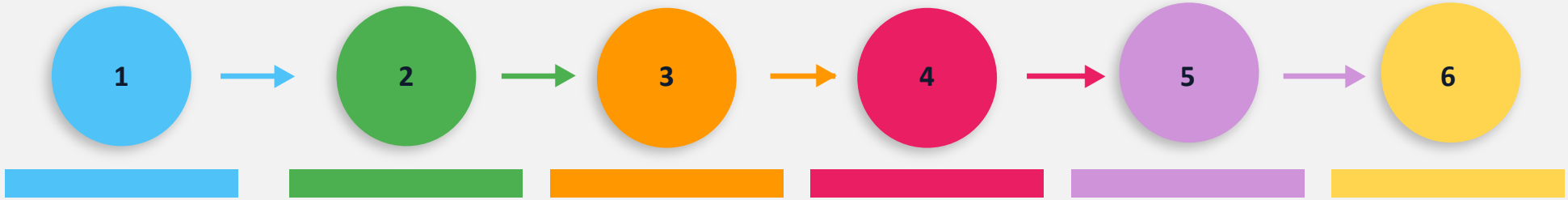


СХЕМА ПРОВЕДЕННЯ ЛІНІЙНОЇ РЕГРЕСІЇ



Дані	EDA	Обробка	Модель	Оцінка	Deploy
Збір та завантаження	Аналіз та візуалізація	Очищення, нормалізація	Навчання МНК/ГС	R^2, MSE, MAE перевірка	Впровадження моделі

ДЕТАЛІ КОЖНОГО КРОКУ

Дані:	EDA:	Обробка:	Модель:	Оцінка:	Deploy:
CSV, SQL, API, xlsx	Кореляція, розподіли, scatter plots	Стандартизація $z=(x-\mu)/\sigma$, видалення NaN	Train/test split 80/20, fit()	Cross-validation, residuals plot	pickle, joblib, REST API

ПІДСУМКИ ЛЕКЦІЇ

1. Лінійна регресія - базовий і потужний алгоритм ML для прогнозування неперервних значень
2. Займає ~32% від усіх задач машинного навчання - найбільша категорія
3. Основний алгоритм навчання - МНК (аналітично) або Градієнтний спуск (ітеративно)
4. Основні види: проста, множинна, поліноміальна, Ridge, Lasso, Elastic Net
5. Оцінка якості: MSE, RMSE, MAE, R^2, MAPE — кожна метрика для різних задач
6. Широке застосування: фінанси, медицина, нерухомість, енергетика, виробництво


Як бачимо, до кінця
наступного місяця
в тебе буде більше 5
десятків чоловіків
краще зроби
замовлення тортів
оптом
кількість
чоловіків
1
0
учора
сьогодні

Самостійна робота

Що потрібно знати о регресії –

1. Для чого застосовується регресія?
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?