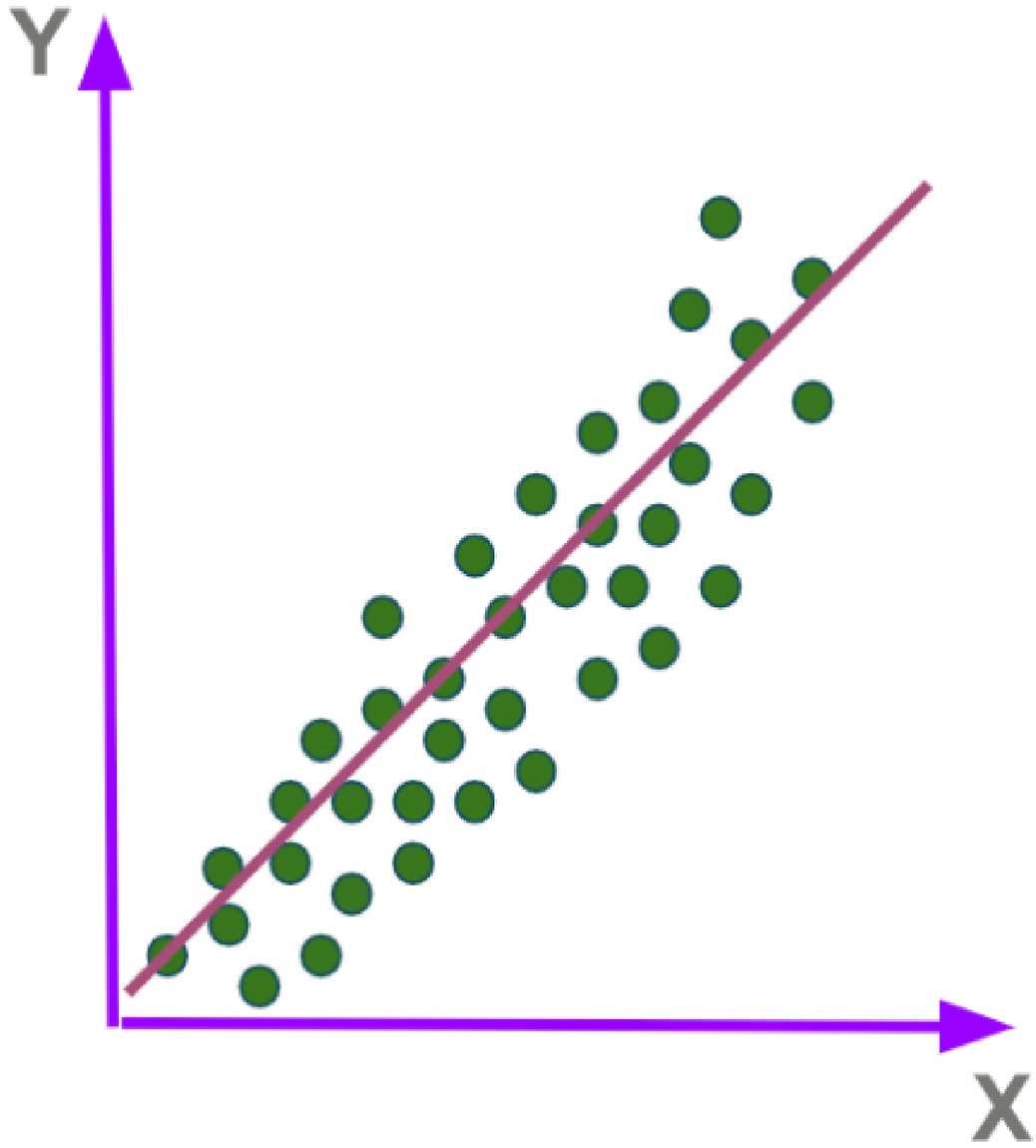
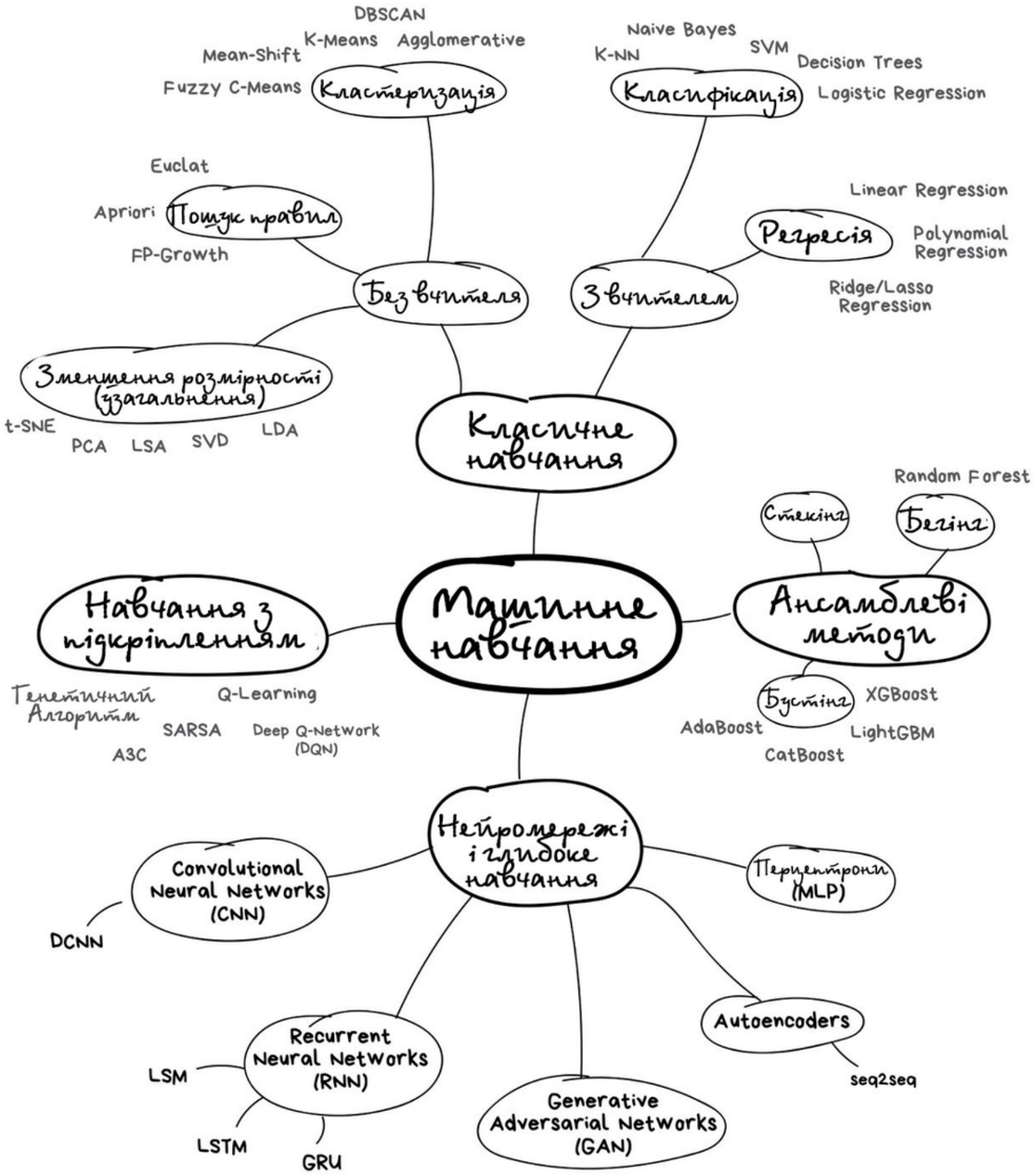
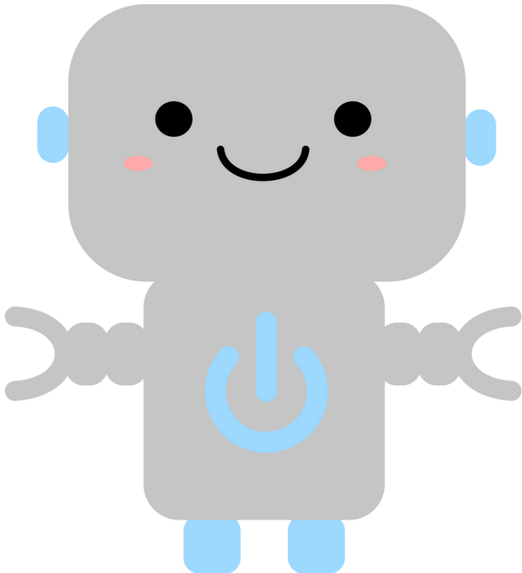
Факультет прикладної математики та інформатики 2024

Лінійна регресія



Кравець Ольга
Кравець Назар
група ПМО-41


Машинне навчання
Класичне навчання
Без вчителя
Кластеризація
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Пошук правил
Euclat
Apriori
FP-Growth
Зменшення розмірності (узагальнення)
t-SNE
PCA
LSA
SVD
LDA
З вчителем
Класифікація
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Регресія
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Навчання з підкріпленням
Генетичний Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Ансамблеві методи
Стекінг
Random Forest
Бегінг
Бустінг
XGBoost
AdaBoost
LightGBM
CatBoost
Нейромережі і глибоке навчання
Convolutional Neural Networks (CNN)
DCNN
Перцептрони (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq



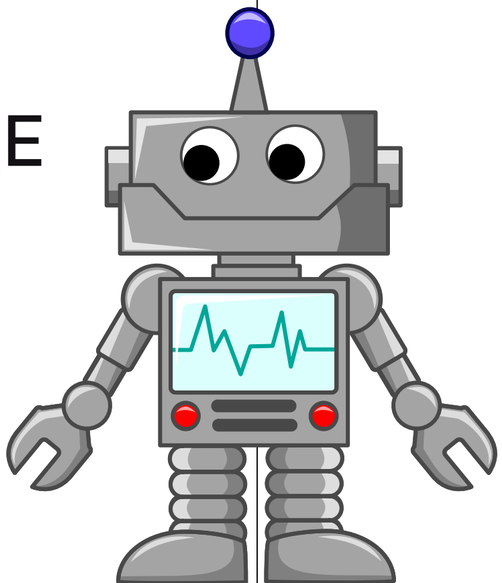
Машина може

ПЕРЕДБАЧАТИ

ЗАПАМ'ЯТОВУВАТИ

ВІДТВОРЮВАТИ

ВИБИРАТИ НАЙКРАЩЕ



Машина не може

СТВОРЮВАТИ НОВЕ

РІЗКО ПОРОЗУМНІШАТИ

ВИЙТИ ЗА РАМКИ ЗАВДАННЯ

ВБИТИ ВСІХ ЛЮДЕЙ


Основні типи машинного навчання
Прості дані
Зрозумілі ознаки
Класичне навчання
Даних немає,
але є середовище,
з яким можна
взаємодіяти
Навчання з
підкріпленням
Коли якість
проблема
Ансамблі
Складні дані
Незрозуміло
де ознаки
Є надія на чудо
Нейромережі
і глибоке
навчання
вічні конкуренти

Класичне навчання

Дані наперед категоризовані або числові

Дані без жодних міток

З вчителем

Передбачити категорію

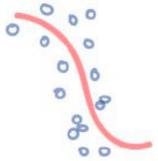
Передбачити значення

Класифікація

«Розклади речі за кольором»

Регресія

«Розклади речі за довжиною»



Без вчителя

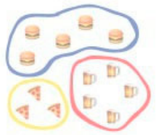
Поділити за подібністю

Виявити послідовності

Знайти залежності

Кластеризація

«Розклади схожі речі в купки»



Асоціація

«Знайди які речі я часто ношу разом»

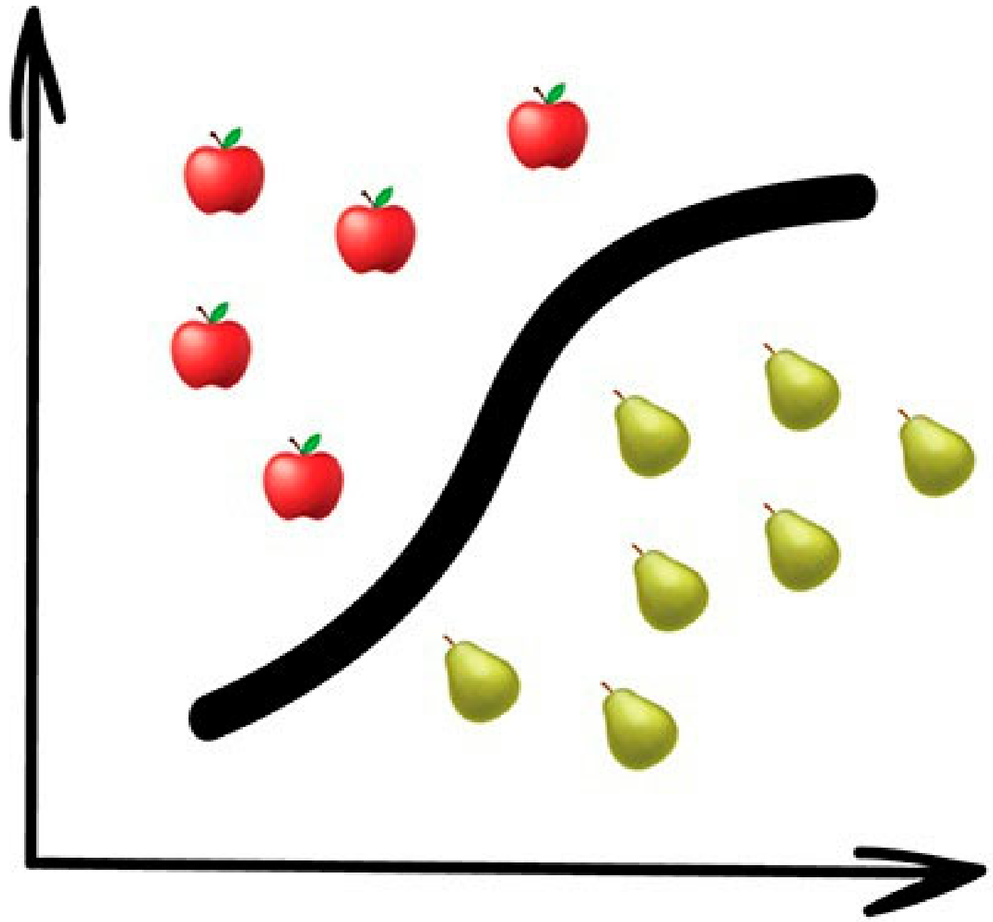
Зменшення Розмірності (узагальнення)

«Збери з речей найкращі наряди»

Класифікація

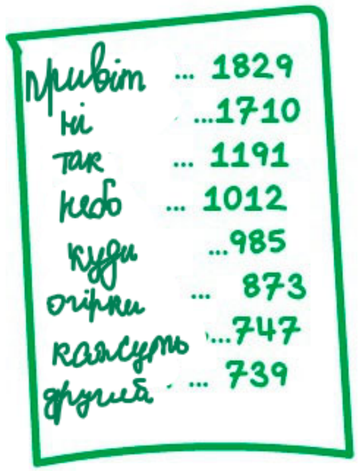
Сьогодні використовують для:

- Спам-фільтри
- Визначення мови
- Пошук схожих документів
- Аналіз тональності
- Розпізнавання рукописних букв і цифр
- Визначення підозрілих транзакцій



Простенький спам-фільтр

(використовували десь до 2010 року)



нормальні листи

672 рази

«КОТИК»

13 разів

спам-листи

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

формула Баєса

не спам

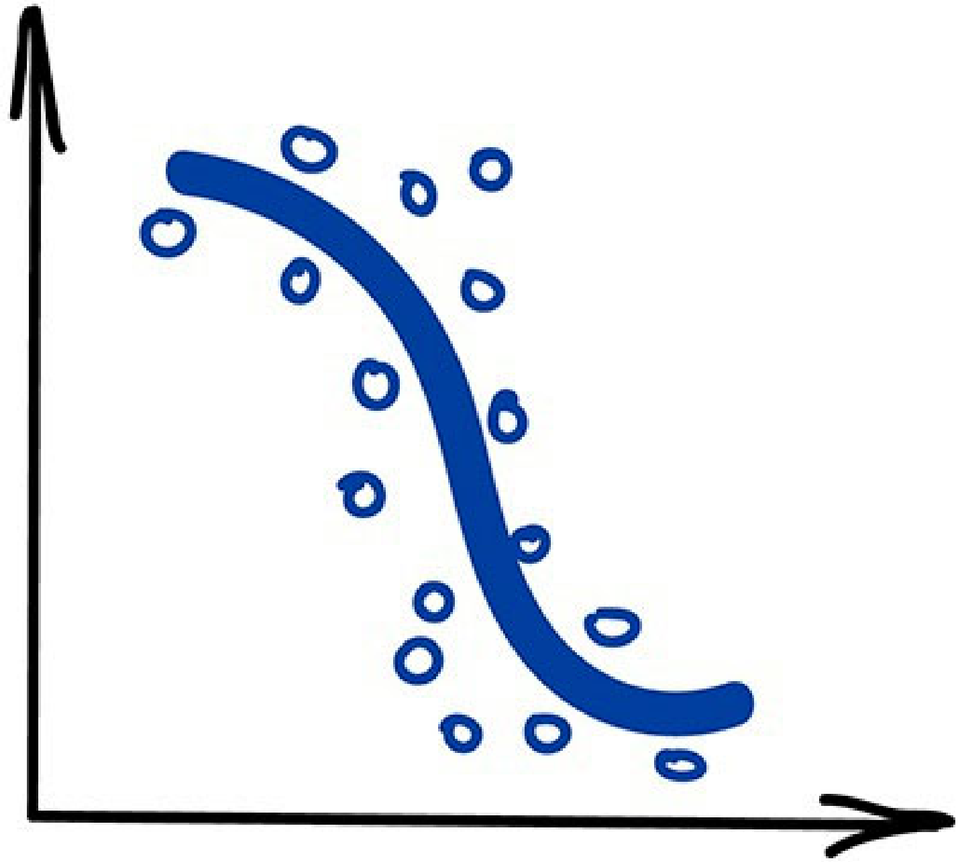
Баєс

РЕГРЕСІЯ

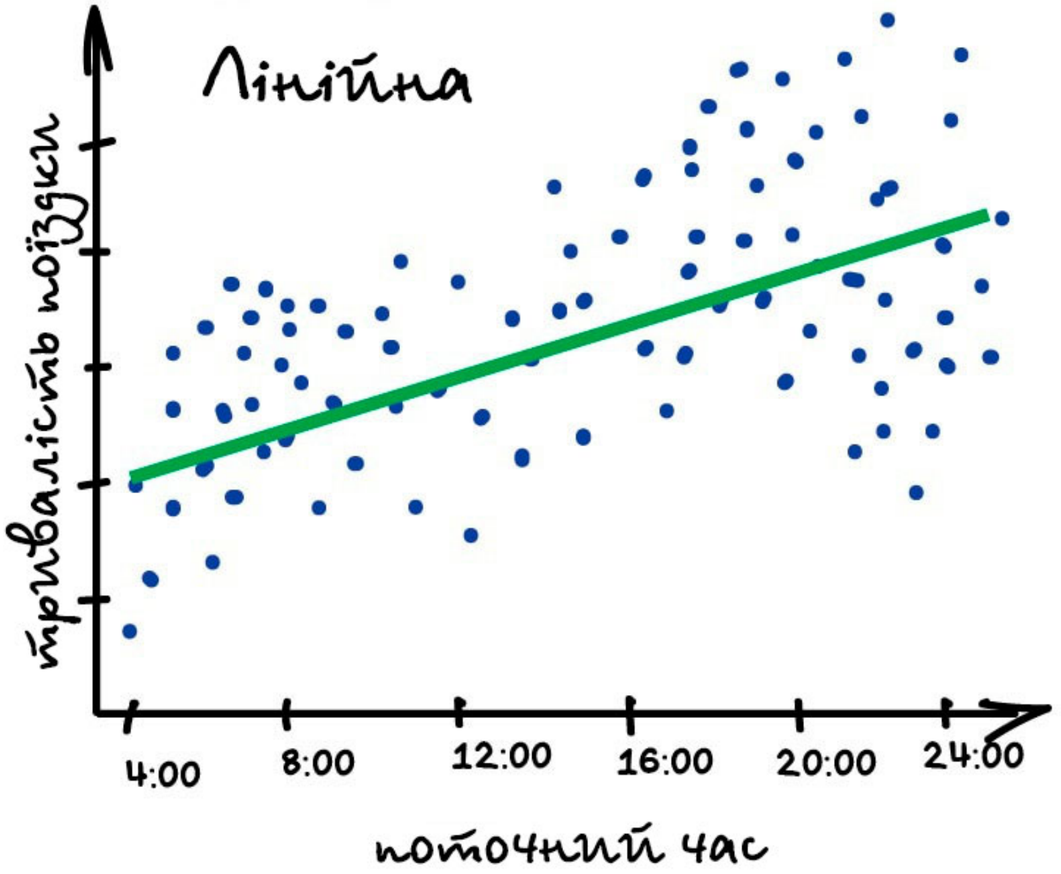
– це метод визначення зв’язку між однією змінною (y) та іншими змінними (x).

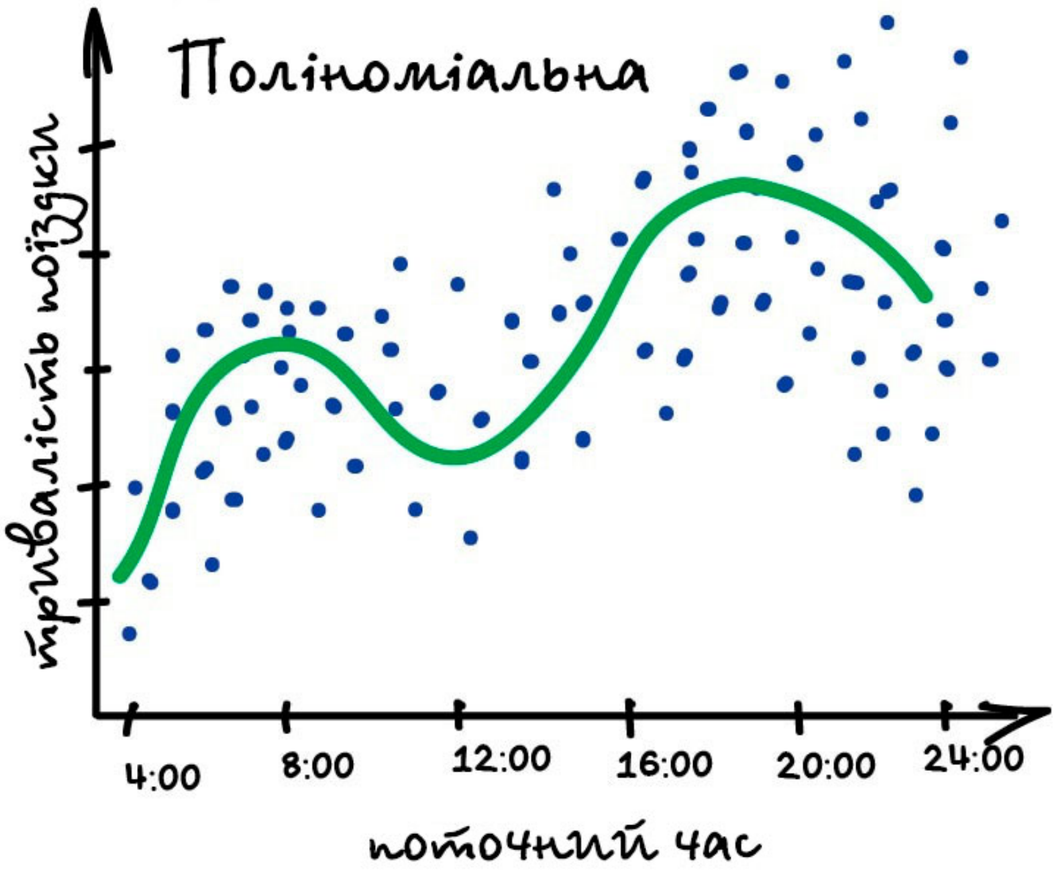
Сьогодні використовують для:

- Прогноз вартості цінних паперів
- Аналіз попиту, обсягу продажів
- Медичні діагнози
- Будь-які залежності числа від часу



Передбачаємо корки на дорогах

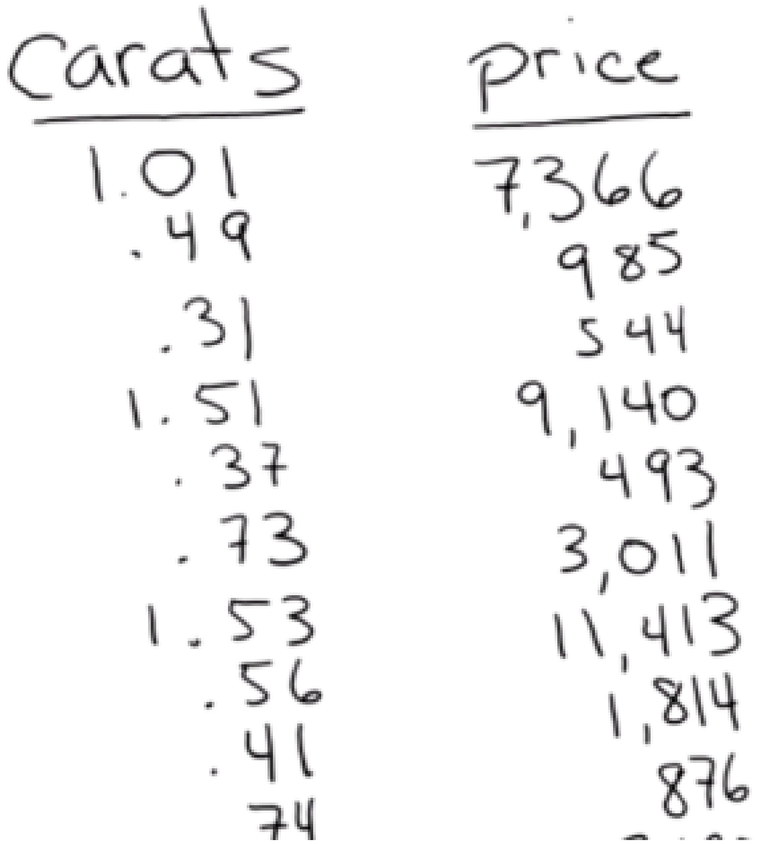




Регресія

Приклад

Припустимо, ми хочемо купити діамант. У нас є кільце, яке належало моєї бабусі. Його оправа містить діамант масою 1,35 карата і я хочу знати скільки він буде коштувати. Я беру з собою блокнот і ручку і йду в ювелірний магазин. Там я записую всі ціни на діаманти, які є у вітрині, а також їх масу в каратах. Починаючи з першого діаманта, який має масу 1,01 карата і коштує 7366$. Я йду далі і роблю те ж саме для інших діамантів в магазині.

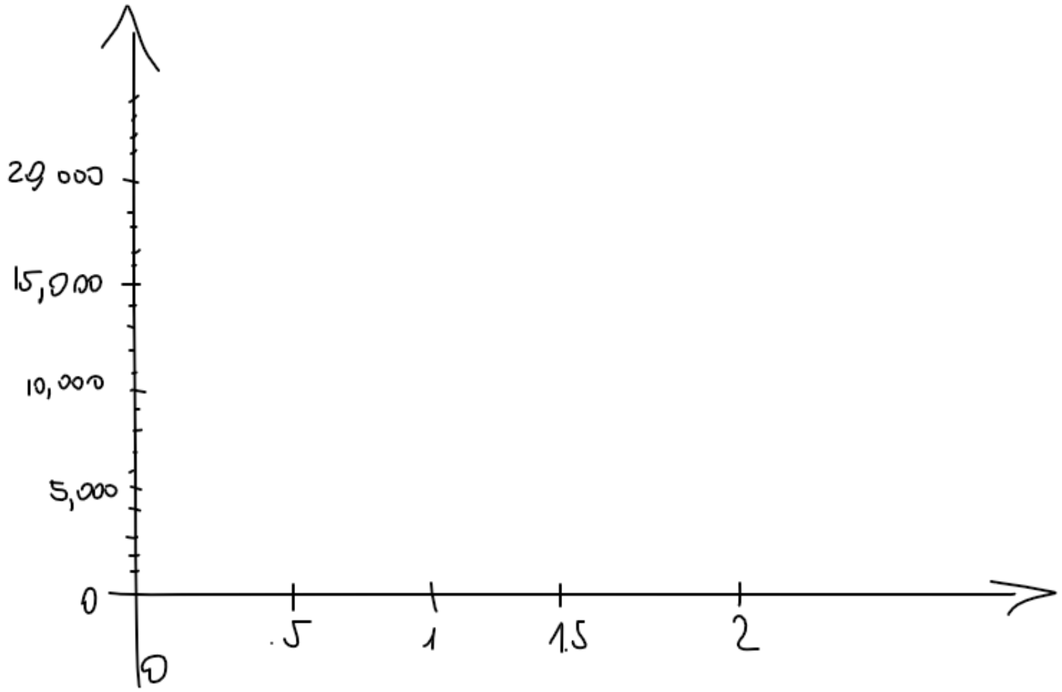


Carats	Price
1.01	7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
74	

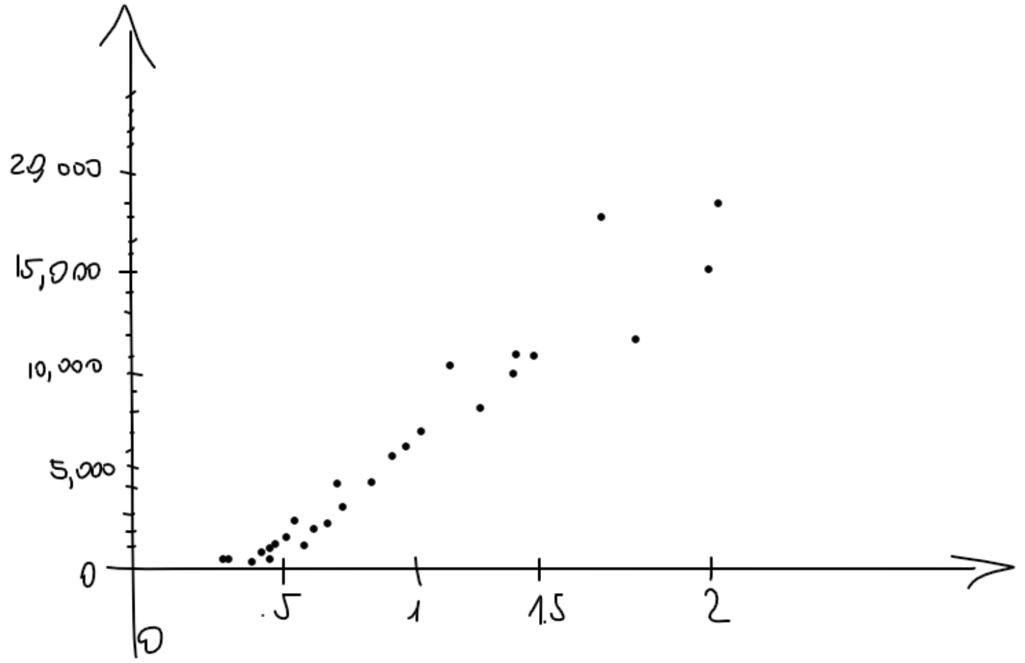
Постановка точного питання

Скільки буде коштувати діамант масою 1,35 карата?

Побудова існуючих даних

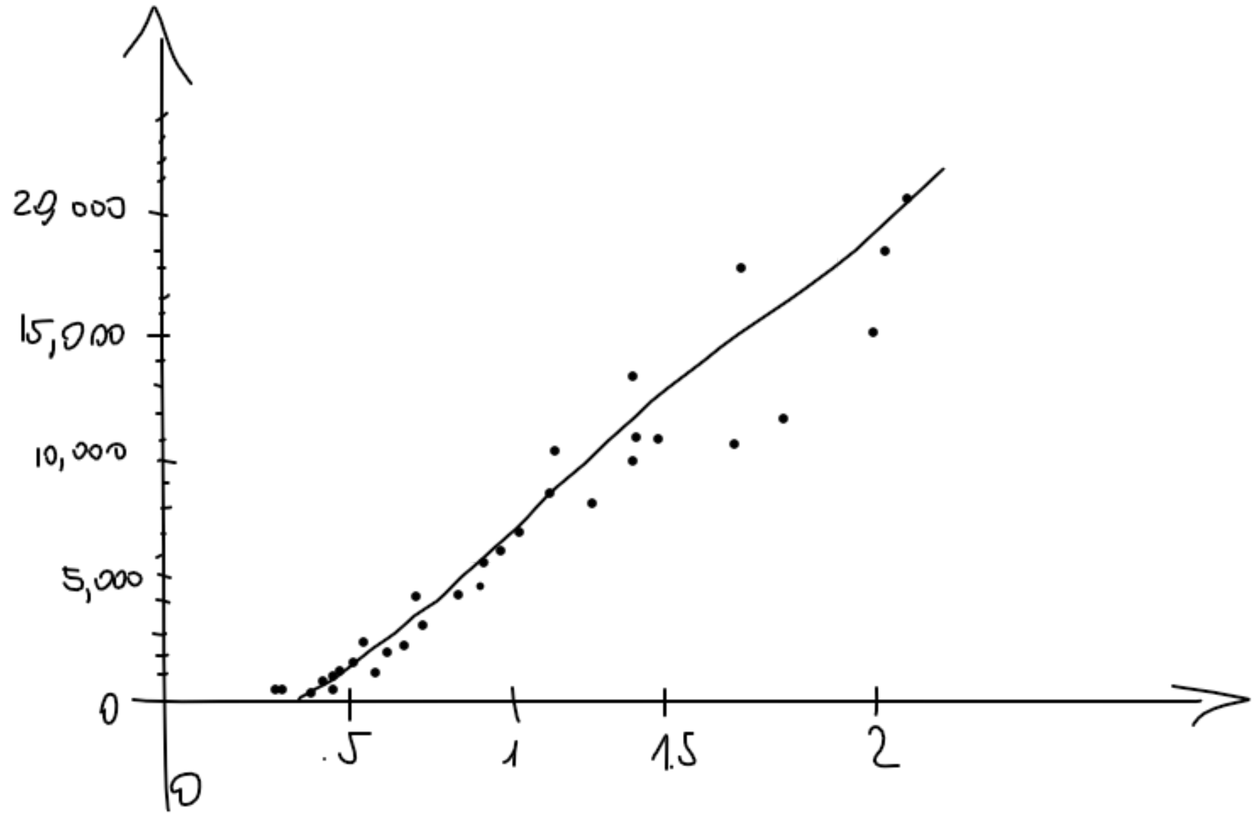


Осі маси і ціни



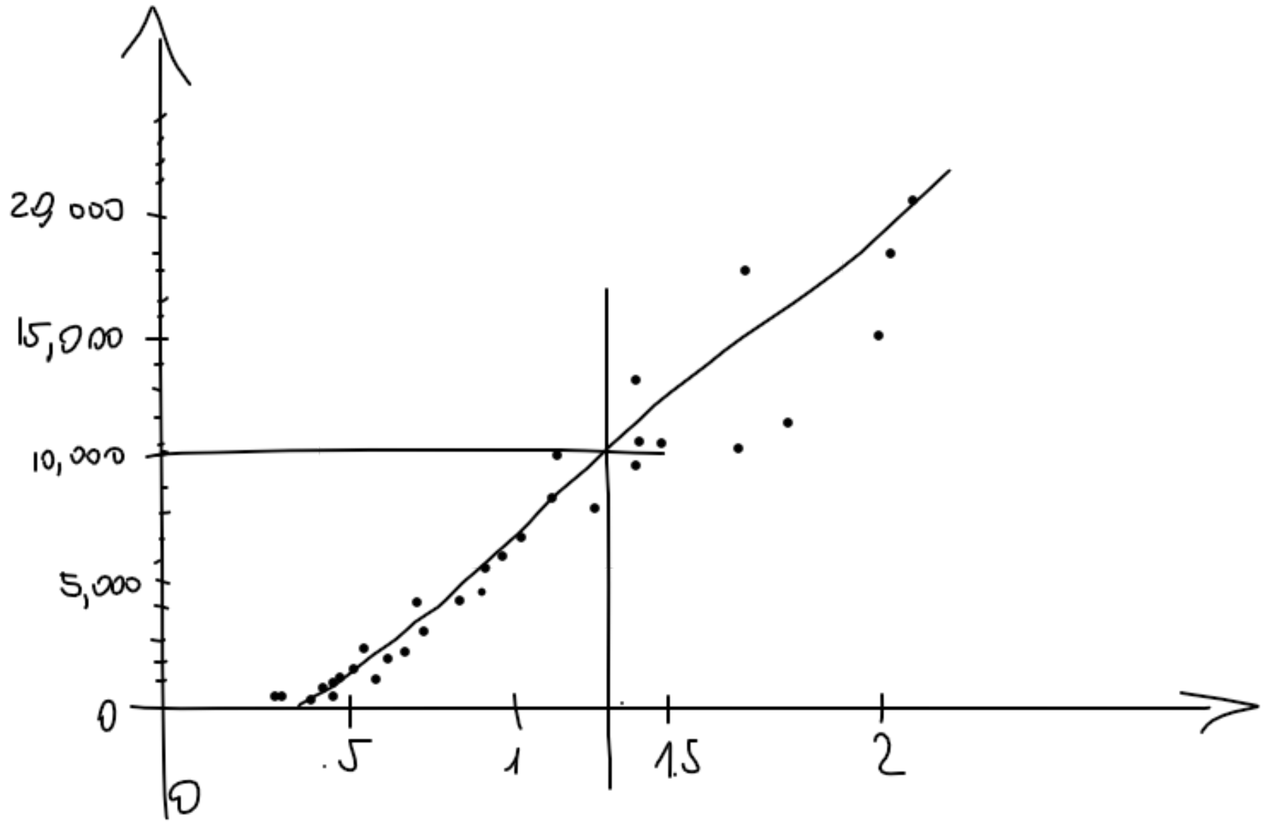
Точкова діаграма

Побудова моделі на основі точок даних



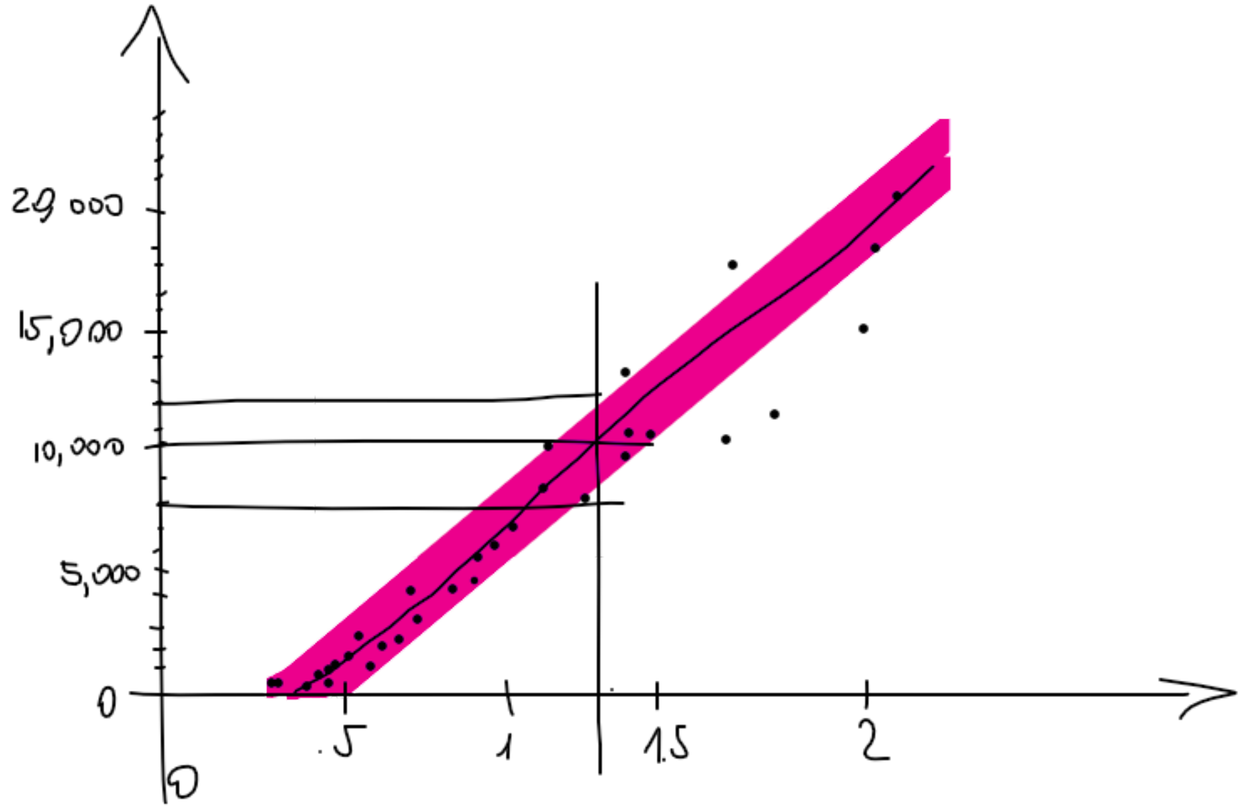
Лінія лінійної регресії

Використання моделі для пошуку відповіді



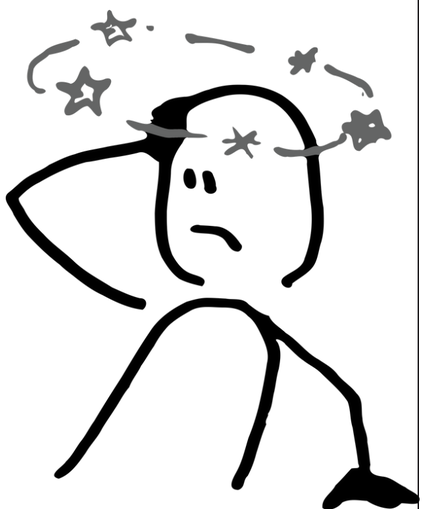
Пошук відповіді за допомогою моделі

Створення довірчого інтервалу



Довірчі інтервали

Тобто:



- МИ ПОСТАВИЛИ ЗАПИТАННЯ, НА ЯКЕ МОЖНА ВІДПОВІСТИ ЗА ДОПОМОГОЮ ДАНИХ.
- МИ СТВОРИЛИ МОДЕЛЬ, ВИКОРИСТОВУЮЧИ ЛІНІЙНУ РЕГРЕСІЮ.
- МИ ЗРОБИЛИ ПРОГНОЗ, ДОПОВНЕНИЙ ДОВІРЧИМ ІНТЕРВАЛОМ.

ПРИ ЦЬОМУ МИ НЕ КОРИСТУВАЛИСЯ МАТЕМАТИЧНИМИ ФОРМУЛАМИ АБО КОМП'ЮТЕРАМИ.

Як будується модель

Лінійна регресійна модель має наступний вигляд:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad (2.1)$$

де y– залежна змінна;

$x_1, x_2, \ldots, x_n$– незалежні змінні;

u– випадкова похибка, розподіл якої в загальному випадку залежить від незалежних змінних, але математичне очікування якої рівне нулю.

Як будується модель

Згідно з моделлю (2.1), математичне очікування залежної змінної є лінійною функцією незалежних змінних:

$$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad (2.2)$$

Вектор параметрів $\beta_0, \beta_1, \ldots, \beta_k$ є невідомим і задача лінійної регресії полягає в оцінці цих параметрів на основі деяких експериментальних значень y і $x_1, x_2, \ldots, x_n$. Тобто, для деяких n експериментів є відомі значення $\{y_i, y_{i1}, \ldots, y_{ip}\}_{i=1}^{n}$ незалежних змінних і відповідне їм значення залежної змінної.

Згідно з визначенням моделі для кожного експериментального випадку залежність між змінними визначається формулою:

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad (2.3)$$

Як будується модель

На основі цих даних потрібно оцінити значення параметрів $(\beta_0, \beta_1, ..., \beta_k)$, а також розподіл випадкової величини u. Зважаючи на характеристики досліджуваних змінних, можуть додаватися різні додаткові специфікації моделі і застосовуватися різні методи оцінки параметрів. Серед найпоширеніших специфікацій лінійних моделей є класична модель лінійної регресії та узагальнена модель лінійної регресії.

Як будується модель

Згідно з класичною моделлю лінійної регресії додатково вводяться такі вимоги щодо специфікації моделі та відомих експериментальних даних:

$\forall i \neq jE(u_i u_j | x_i) = 0$ (відсутність кореляції залишків);

$\forall iE(u_j^2 | x_i) = \sigma^2$ (гомоскедастичність).

Часто додається також умова нормальності випадкових відхилень, яка дозволяє провести значно ширший аналіз оцінок параметрів та їх значимості, хоча і не є обов'язковою для можливості використання наприклад методу найменших квадратів $(u_j | x_i) \sim N(0, \sigma^2)$.

Як будується модель

Передбачимо, що незалежна змінна набула значень $x_1, x_2 ..., x_n$, внаслідок чого залежна змінна набула значень $y_1, y_2 ..., y_n$. У припущенні лінійної залежності отримуємо n рівностей.

$$y_i = a_0 + a_1 x_i + \varepsilon_i, i = 1 ...n, \qquad (2.4)$$

де ε_i – незалежні і розподілені так само, як ε.

Як будується модель

Потрібно за значеннями пар (x_i, y_i) оцінити невідомі a_0, a_1. Як ми вже знаємо, кожне завдання оцінювання пов'язане з деяким критерієм якості. У теорії, що викладається нами, таким критерієм є критерій найменших квадратів:

$$\sum_{i=1}^{n} \varepsilon_i^2 - \min . \qquad (2.5)$$

Як будується модель

Знаходимо часткові похідні функції Q і прирівнюємо їх до нуля, внаслідок чого приходимо до системи лінійних рівнянь:

$$\begin{cases} \frac{\partial Q}{\partial a_0} = -2\sum_{i=1}^{n}[y_i - a_0 - a_1 x_i] = 0, \\ \frac{\partial Q}{\partial a_1} = -2\sum_{i=1}^{n}[y_i - a_0 - a_1 x_i]\, x_i = 0. \end{cases} \qquad (2.8)$$

Після очевидних перетворень отримуємо систему:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ a_0 \sum_{i=1}^{n} x_i + a_i \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad (2.9)$$

Як будується модель

Позначимо вибіркові середні:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\,y_i,\ \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2\,. \qquad (2.10)$$

У цих позначеннях після ділення кожного рівняння системи на n вона набуде вигляду:

$$\begin{cases} a_0 + a_1\bar{x} = \bar{y}, \\ a_x\bar{x} + a_1\bar{x}^2 = \overline{xv} \end{cases} \qquad (2.11)$$

а її рішення (шукані оцінки коефіцієнтів рівняння регресії) буде таким:

$$\hat{a}_0 = \frac{\overline{x^2}\cdot\bar{y} - \overline{xy}\cdot\bar{x}}{\overline{x^2} - (\bar{x})^2}$$
$$\hat{a}_1 = \frac{\overline{xy} - \bar{y}\cdot\bar{x}}{\overline{x^2} - (\bar{x})^2} \qquad (2.12)$$

Як будується модель

Якщо ввести ще позначення $S_x^2 = \overline{x^2} - \overline{(x)}^2$ і перетворити вираз $\hat{a}_0$:

$$\hat{a}_0 = \frac{\overline{x^2} \cdot \bar{y} - \overline{xy} \cdot \bar{x} \pm \bar{y} \cdot (\bar{x})^2}{S_{..}^2} = \frac{S_x^2 \cdot \bar{y} - \bar{x}(\overline{xy} - \bar{y} \cdot \bar{x})}{S_{..}^2} = \bar{y} - \hat{a}_1 \cdot \bar{x}, \quad (2.14)$$

то оцінка функції регресії набуде такого вигляду:

$$\tilde{\bar{y}}(x) = \hat{a}_0 + \hat{a}_1 x = \bar{y} - \hat{a}_1 \cdot \bar{x} + \hat{a} \cdot x = \bar{y} + \hat{a}_1 (x - \bar{x}) \quad (2.15)$$

Таким чином, отримали рівняння регресії, що є моделлю для опису даних.

ОЦІНКИ ЯКОСТІ РЕГРЕСІЇ

СЕРЕДНЯ КВАДРАТИЧНА ПОМИЛКА (АНГЛ. MEAN SQUARED ERROR, MSE)

застосовується в ситуаціях, коли нам треба підкреслити великі помилки і вибрати модель, яка дає менше помилок прогнозу

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2$$

СЕРЕДНЯ АБСОЛЮТНА ПОМИЛКА (АНГЛ. MEAN ABSOLUTE ERROR, MAE)

при застосуванні важливо враховувати, що великі відхилення об'єктів можуть суттєво вплинути на загальну помилку моделі, вказуючи на можливі помилки в обчисленні ознак або цільової величини.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - y_i|$$

КОЕФІЦІЄНТ ДЕТЕРМІНАЦІЇ

вимірює частку дисперсії, внесеною моделлю, в загальній дисперсії цільової змінної

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

СЕРЕДНЯ АБСОЛЮТНА ПРОЦЕНТНА ПОМИЛКА (АНГЛ. MEAN ABSOLUTE PERCENTAGE ERROR, MAPE)

це коефіцієнт, який не має розмірності, з дуже простою інтерпретацією

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - a(x_i)|}{|y_i|}$$

КОРІНЬ ІЗ СЕРЕДНЬОЇ КВАДРАТИЧНОЇ ПОМИЛКИ (АНГЛ. ROOT MEAN SQUARED ERROR, RMSE)

так як кожне відхилення зводиться в квадрат, будь невелике відхилення може значно вплинути на показник помилки

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y_i})^2}$$

СИМЕТРИЧНА MAPE (АНГЛ. SYMMETRIC MAPE, SMAPE)

$$SMAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \times |y_i - a(x_i)|}{|y_i| + |a(x_i)|}$$

СЕРЕДНЯ АБСОЛЮТНА МАСШТАБОВАНА ПОМИЛКА (АНГЛ. MEAN ABSOLUTE SCALED ERROR, MASE)

є дуже хорошим варіантом для розрахунку точності, так як сама помилка не залежить від масштабів даних і є симетричною

$$MASE = \frac{\sum_{i=1}^{n} |Y_i - e_i|}{\frac{n}{n-1} \sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

КРОС-ВАЛІДАЦІЯ

В цьому випадку фіксується деяка множина розбиття вихідної вибірки на дві підвибірки: навчальну і контрольну. Для кожного розбиття виконується настройка алгоритму за навчальною підвибіркою, потім оцінюється його середня помилка на об'єктах контрольної підвибірки. Оцінкою змінного контролю називається середня по всім розбиттям величина помилки на контрольних підвибірках.

(Height)

$$X$$

ДЯКУЄМО ЗА УВАГУ!