









Вступ до інтелектуального аналізу даних





Надія Іванівна Недашківська

Інститут прикладного системного аналізу Національного технічного
університету України "Київський політехнічний інститут ім. Ігоря
Сікорського"

Київ-2021

-  Конспект лекцій.
-  Методичні вказівки до виконання лабораторних робіт.
-  Гудфеллоу Я., Бенджіо И., Курвилль А.. *Глубокое обучение*. пер. с англ. А. А. Слинкина. 2-е изд., испр. М.: ДМК Пресс, 2018, 652 с.
-  Себастьян Рашка, Вахид Мирджалили. *Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2*, 3-е изд.: Пер. с англ.-СПб.: Диалектика 2020.-848 с.
-  Николенко С., Кадури́н А., Архангельская Е. *Глубокое обучение*. СПб.: Питер, 2018, 480 с.

-  Дж. Вандер Плас. *Python для сложных задач. Наука о данных и машинное обучение*. СПб.:Питер,2018. 576с.
-  Уэс Маккинли. *Python и анализ данных* М.: ДМК Пресс, 2015. – 482 с.
-  Scikit-Learn Documentation.
<https://scikit-learn.org/>

-  Бринк Х., Ричардс Дж., Феверолф М. *Машинное обучение*. СПб.: Питер, 2017. 336 с.
-  Мюллер А., Гидо С. *Введение в машинное обучение с помощью Python*. М.: O'Reilly Media, 2017. 392 с.
-  Силен Д., Мейсман А., Али М. *Основы Data Science и Big Data. Python и наука о данных*. СПб.: Питер, 2017. 336 с.
-  Матеріали щодо машинного навчання.
<http://www.machinelearning.ru>. 2019.

Інтелектуальний аналіз даних — виявлення знань у великих об'ємах даних :

- раніше невідомих знань,
- нетривіальних знань,
- практично корисних знань,
- доступних для інтерпретації людиною.

(Григорій П'ятецький-Шапіро)

Алгоритм машинного навчання - це алгоритм, здатний навчатися на даних.

«Комп'ютерна програма навчається на досвіді E відносно деякого класу задач T і міри якості P , якщо якість на задачах з T , виміряна за допомогою P , зростає із зростанням досвіду E »
(Mitchell, 1997)

Класифікація - віднесення прикладу до одного з k визначених класів.

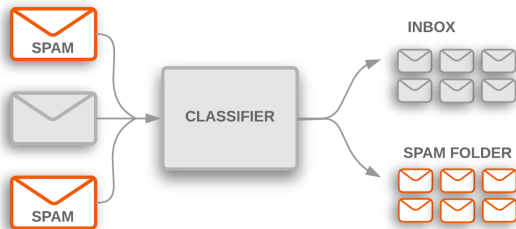
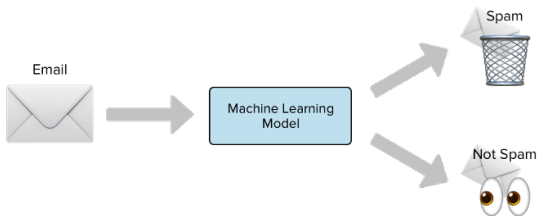
Алгоритм навчання (класифікатор) породжує функцію

$$f : R^n \rightarrow \{1, \dots, k\}$$

Приклади:

- Розпізнавання зображень, представлених матрицею значень яскравості пікселів.
- Фільтрація електронної пошти.
- Надання кредиту.

Приклад: фільтрація електронної пошти



Використання МН для задачі кредитного скорингу

X - вектор ознак позичальників

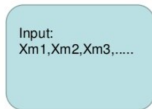
Y - цільова змінна -
чи надавати кредит

$Y = 1$ - так

$Y = 0$ - ні

$X_{11}, X_{12}, X_{13}, \dots, Y_1$
 $X_{21}, X_{22}, X_{23}, \dots, Y_2$
 \dots
 $X_{n1}, X_{n2}, X_{n3}, \dots, Y_n$

Історія минулих кредитів



Ознаки нового позичальника



Чи надавати кредит новому позичальнику?

Якщо імовірність $Y >$ порогу, то давати кредит

Якщо імовірність $Y \leq$ порогу, то не давати кредит

Регресія - прогнозування числового значення за вхідними даними.

Алгоритм навчання породжує функцію

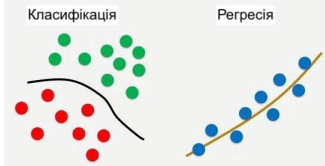
$$f : R^n \rightarrow R$$

Приклади:

- Прогнозування розміру страхової премії.
- Прогнозування майбутньої вартості цінних паперів.
- Пошук на фотографії координат прямокутника, в якому знаходиться обличчя людини.

Класифікація VS Регресія

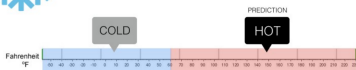
Класифікація VS Регресія



Регресія:
Якою буде температура завтра?



Класифікація:
Завтра буде холодно чи спекотно?



Структурний вивід - на виході породжується вектор (або інша структура, яка містить кілька значень), між елементами якого існують **важливі зв'язки**.

Приклади:

- транскрипція,
- машинний переклад,
- граматичний розбір,
- піксельна сегментація зображення,
- анотування доріг на аерофотознімках,
- підписування зображень.

Транскрипція - аналіз неструктурованого представлення даних і перетворення його в дискретну текстову форму.

Приклади:

- Програмі розпізнавання тексту пред'являється фотографія тексту, а вона повинна повернути текст у вигляді послідовності символів. Google Street View виконує обробку табличок з адресами будинків.
- Розпізнавання мови, коли програмі пред'являється аудіосигнал, а вона виводить послідовність символів (в компаніях Microsoft, IBM і Google).

Виявлення аномалій - нетипових подій або об'єктів.

- Приклад: виявлення шахрайства з кредитними картами на основі моделювання купівельних звичок.

Шумозаглушення:

- на вхід подається зашумлений приклад $\hat{x} \in R^n$,
- алгоритм має відновити початковий приклад x на основі зашумленого \hat{x} або повернути умовний розподіл ймовірності $p(x|\hat{x})$.

Задача синтезу і вибірки - алгоритм генерує нові приклади, схожі на навчальні дані.

Приклади:

- мультимедія,
- відеоігри.

Алгоритм генерує певний вихід за заданим входом.

Приклад:

- синтез мови, вхід - написана фраза, вихід - звук.

Задачі МН: оцінка функції ймовірності та функції щільності ймовірності

Задачі оцінки функції ймовірності та функції щільності ймовірності

Алгоритм оцінює функцію $p_{model} : R^n \rightarrow R$, де

- $p_{model}(X)$ - функція ймовірності, якщо X - дискретна випадкова величина або
- $p_{model}(X)$ - функція щільності ймовірності, якщо X - неперервна випадкова величина,

в просторі, з якого були взяті приклади.

Пошук асоціативних правил - знаходження **частих** залежностей, асоціацій у вигляді правил "Якщо - То" між об'єктами або подіями.

Приклади:

- Аналіз ринкових кошиків (Basket Analysis).
- Аналіз симптомів і хвороб, що спостерігаються у пацієнтів.
- Сиквенційний аналіз.

Приклад: Телекомунікаційні компанії. Встановлено послідовність збоїв $\{x_5, x_2, x_7, x_{13}, x_6\}$.

Факт появи збою x_2 - швидка поява збою x_7 .

Алгоритми машинного навчання можна умовно розділити на три великі класи:

- з вчителем (supervised learning),
- без вчителя (unsupervised learning),
- з частковим залученням вчителя, напівконтрольоване (semi-supervised) навчання,

залежно від того, на якому досвіді, наборі даних вони можуть навчатися.

Набором даних називається сукупність великого числа прикладів.

Приклад - це вектор $x \in R^n$,
кожен елемент якого - **ознака**, отримана в результаті кількісного виміру деякого об'єкту чи події.

Приклади також називаються **вимірами, точками**.

Ознаки називаються **атрибутами**.

Навчання з вчителем. Дано: навчальна вибірка даних X , де кожен приклад включає **мітку або цільовий клас** y - це число або y загальному вигляді вектор.

Потрібно: спрогнозувати y на основі X , розрахувати оцінку

$$\hat{p}(y|X) - ?$$

Припущення: навчальні дані мають бути схожими на дані, на яких потім буде застосовуватися побудована модель.

- *Класифікація.*
- *Регресія.*
- *Навчання ранжуванню (learning to rank).*

Навчання ранжуванню (learning to rank) - впорядкувати наявні об'єкти в порядку спадання цільової функції.

- Наприклад, на основі текстів документів і минулої поведінки користувача.
- Використовується в пошукових і рекомендаційних системах.
- Цільова функція називається **релевантністю** і є мірою відповідності даного документу зробленому запиту.

Навчання без вчителя - виявити корисні властивості заданого набору даних.

- **Кластеризація:** розділити дані на **наперед невідомі** класи, використовуючи деяку міру схожості.
Приклади:
 - персоналізація користувачів веб-сайта,
 - сегментація медичного знімку для виявлення захворювання.
- **Задача оцінки щільності:** оцінити розподіл, з якого отримано вхідні дані, знаючи апіорні імовірності їх появи.
- **Задачі синтезу або очищення від шуму.**

Зниження розмірності

- **Дано:** вхідні дані, які мають велику розмірність.
- **Потрібно:** отримати представлення цих даних в просторі меншої розмірності, яке буде досить повно відображати вхідні дані. Цілі:
 - Зменшення обчислювальних витрат.
 - Сжимання даних для більш ефективного збереження інформації. Потрібне також зворотнє перетворення.
 - Отримання нових ознак (feature extraction).
 - Уникнення перенавчання.
 - Візуалізація даних.

Навчання з частковим залученням вчителя, напівконтрольоване (semi-supervised) навчання

Навчання з частковим залученням вчителя, напівконтрольоване навчання

Дано: багато нерозмічених даних.

Ідея: модель спочатку навчається на нерозмічених даних, а потім, використовуючи це наближення, до-навчається на розмічених.

Приклад: **Навчання з підкріпленням (reinforcement learning)**

Розпишемо спільний розподіл для вектора $x \in R^n$:

$$p(x) = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1}).$$

Тобто, оцінювання $p(x)$ (задачу без вчителя) можна представити як n задач МН з вчителем.

З іншого боку,

$$p(y|x) = \frac{p(x, y)}{\sum_z p(x, z)}.$$

Висновок. Одні й ті ж технології МН можуть застосовуватися до розв'язання як задач з вчителем, так і без вчителя.

Описові (descriptive) задачі

Приклади:

- кластеризація,
- пошук асоціативних правил.

Прогнозні (predictive) задачі

Приклади:

- класифікація,
- регресія,
- пошук асоціативних правил

Етап 1: за навчальною вибіркою будується модель

- інформація про клієнтів, яким раніше видавалися кредити на різні суми, і інформація про їхнє погашення,
- повідомлення, класифіковані вручну як спам або як лист,
- розпізнані раніше матриці зображень.

Етап 2: модель застосовується для прогнозу на нових наборах даних - об'єктах з невизначеним значенням залежної змінної

Кластеризація в порівнянні з класифікацією

- Не потрібно мати окрему залежну змінну, кластеризація - задача навчання без учителя.
- Кластеризація використовується на початкових етапах дослідження, це описова задача.

Алгоритм машинного навчання - це алгоритм, здатний навчатися на даних.

«Комп'ютерна програма навчається на досвіді E відносно деякого класу задач T і міри якості P , якщо якість на задачах з T , виміряна за допомогою P , зростає із зростанням досвіду E » (Mitchell, 1997)

Як вибрати міру якості - ? Розглянемо приклад задачі лінійної регресії.

Вхід: вектор $x \in R^n$. Вихід: $y \in R$.

Результат моделі лінійної регресії \hat{y} - лінійна функція вхідних даних x :

$$\hat{y} = w^T x,$$

де $w \in R^n$ - вектор параметрів - ваги.

Потрібно: покращити w так, щоб функція помилки (втрат) зменшувалася по мірі того, як алгоритм отримує новий приклад з навчального набору (X^{train}, y^{train}) :

$$X^{train} = \{x_1, x_2, \dots, x_m\}, x_i \in R^n, y^{train} \in R^m.$$

Ілюстрація задачі лінійної регресії

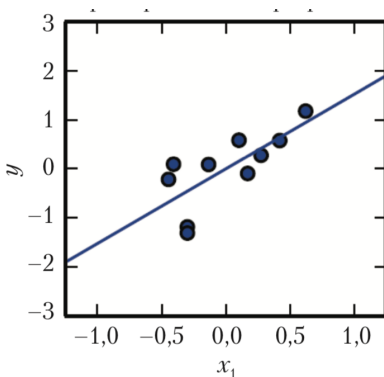


Рис.: Приклад лінійної регресії

Задача мінімізації функції помилки на навчальному наборі

Функція помилки:

$$MSE^{train} = \frac{1}{m} \sum_i (\hat{y}_i^{train} - y_i^{train})^2 = \frac{1}{m} \|\hat{y}^{train} - y^{train}\|^2 \rightarrow \min$$

Прирівняємо градієнт MSE^{train} до нуля:

$$\frac{\partial}{\partial w} \left(\frac{1}{m} \|\hat{y}^{train} - y^{train}\|^2 \right) = 0$$

$$\frac{\partial}{\partial w} \left((X^{train} w - y^{train})^T (X^{train} w - y^{train}) \right) = 0!!!$$

$$\frac{\partial}{\partial w} \left(w^T X^{trainT} X^{train} w - 2w^T X^{trainT} y^{train} + y^{trainT} y^{train} \right) = 0$$

$$2X^{trainT} X^{train} w - 2X^{trainT} y^{train} = 0$$

$$w = (X^{trainT} X^{train})^{-1} X^{trainT} y^{train}$$

Ілюстрація задачі лінійної регресії

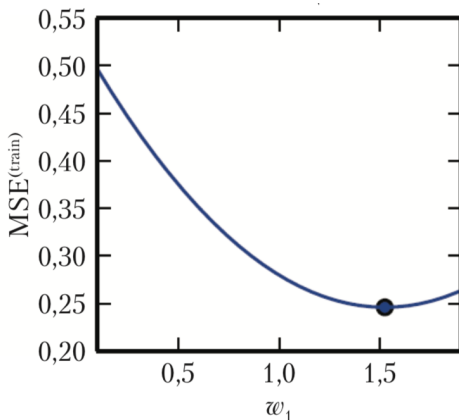


Рис.: Мінімальне значення MSE^{train} в прикладі лінійної регресії

Міра якості в задачі лінійної регресії

Вхід:

$X^{test} = \{(x_1, x_2, \dots, x_p) | x_i \in R^n\}$ - тестовий набір даних для оцінки якості роботи моделі,

$y^{test} = (y_1, y_2, \dots, y_p)^T$ - вектор міток, який містить правильні значення y для кожного з прикладів.

Функція помилки:

$$MSE = \frac{1}{p} \sum_i (\hat{y}_i^{test} - y_i^{test})^2 = \frac{1}{p} \|\hat{y}^{test} - y^{test}\|^2,$$

де \hat{y}^{test} - вектор значень прогнозу моделі на тестовому наборі, $\|\cdot\|$ - евклідова відстань.

Дякую за увагу!