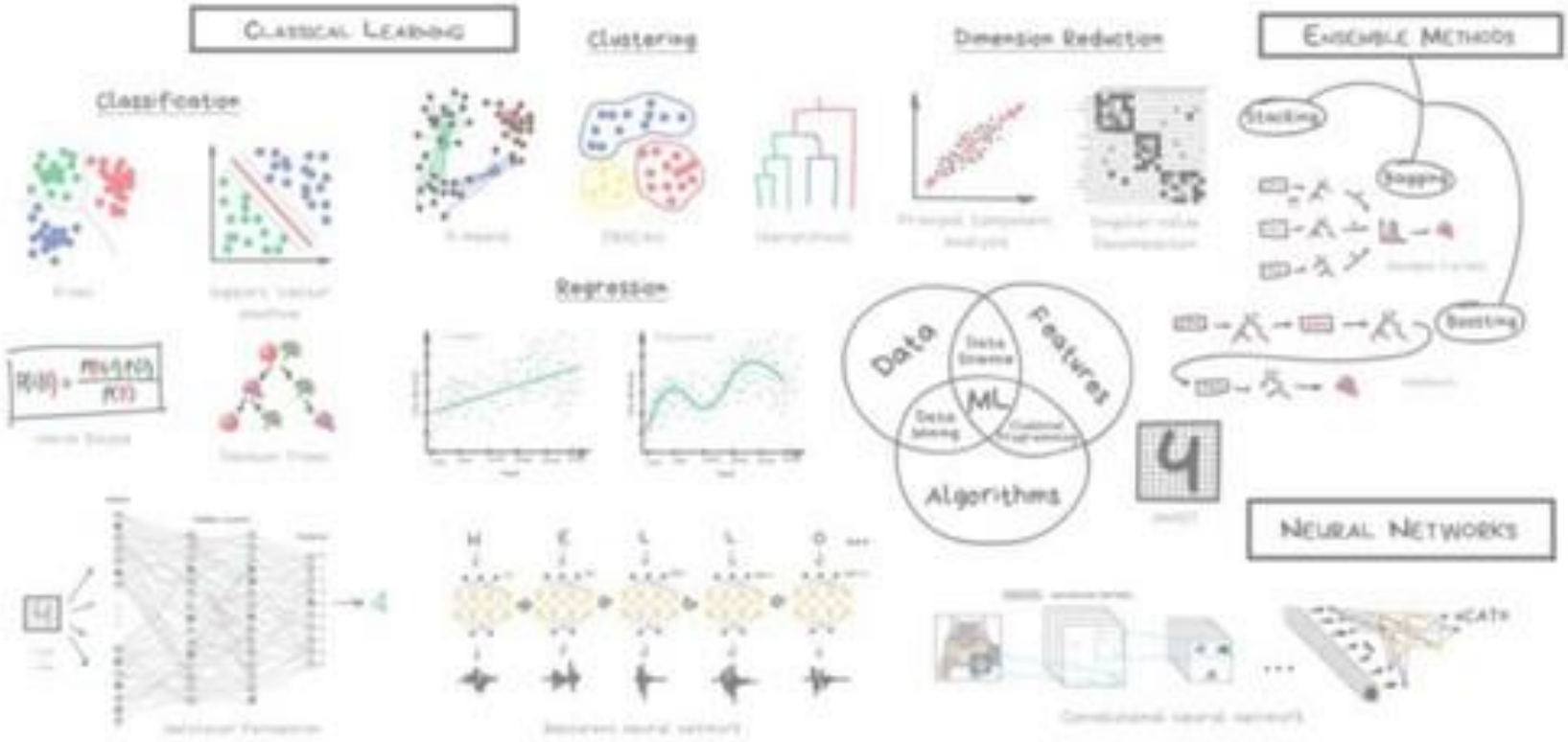
Машинне навчання



Лекція № 3

Три складові навчання

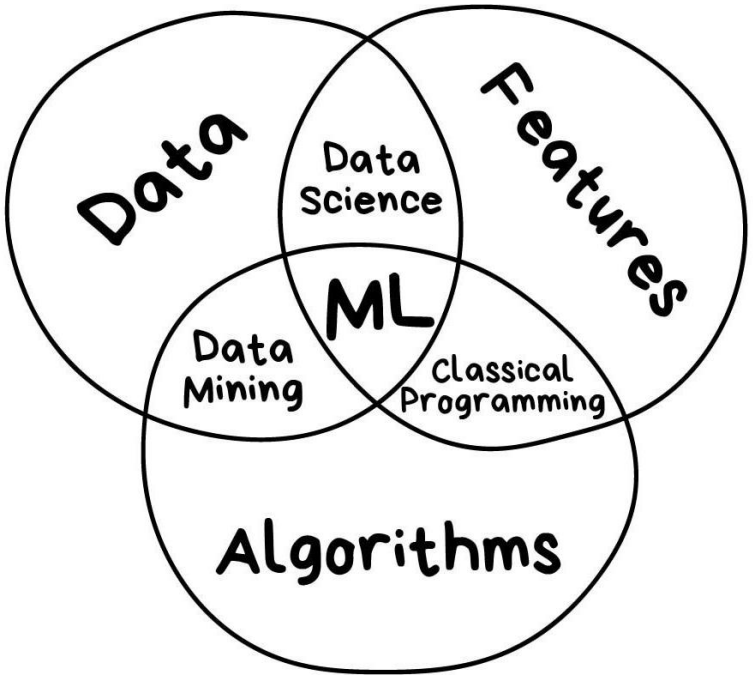
- Дані

Хочемо визначати спам — потрібні приклади спам-листів, передбачати курс акцій — потрібна історія цін, дізнатися про інтереси користувача — потрібні його лайки чи пости.

Даних потрібно якнайбільше !!!

- Ознаки (features) — властивості, характеристики,

Ознаки — ними можуть бути: пробіг автомобіля, стать користувача, ціна акцій, лічильник частоти появи слова в тексті, ...



- Алгоритм

Одну задачу можна вирішити різними методами приблизно завжди. Від вибору методу залежить точність, швидкість роботи та розмір готової моделі. Проте є один нюанс: якщо дані «погані», тонавіть найкращий алгоритм не допоможе у розв'язанні задачі.

!!! Не зациклюватися на відсотках, краще зібрати більше даних.

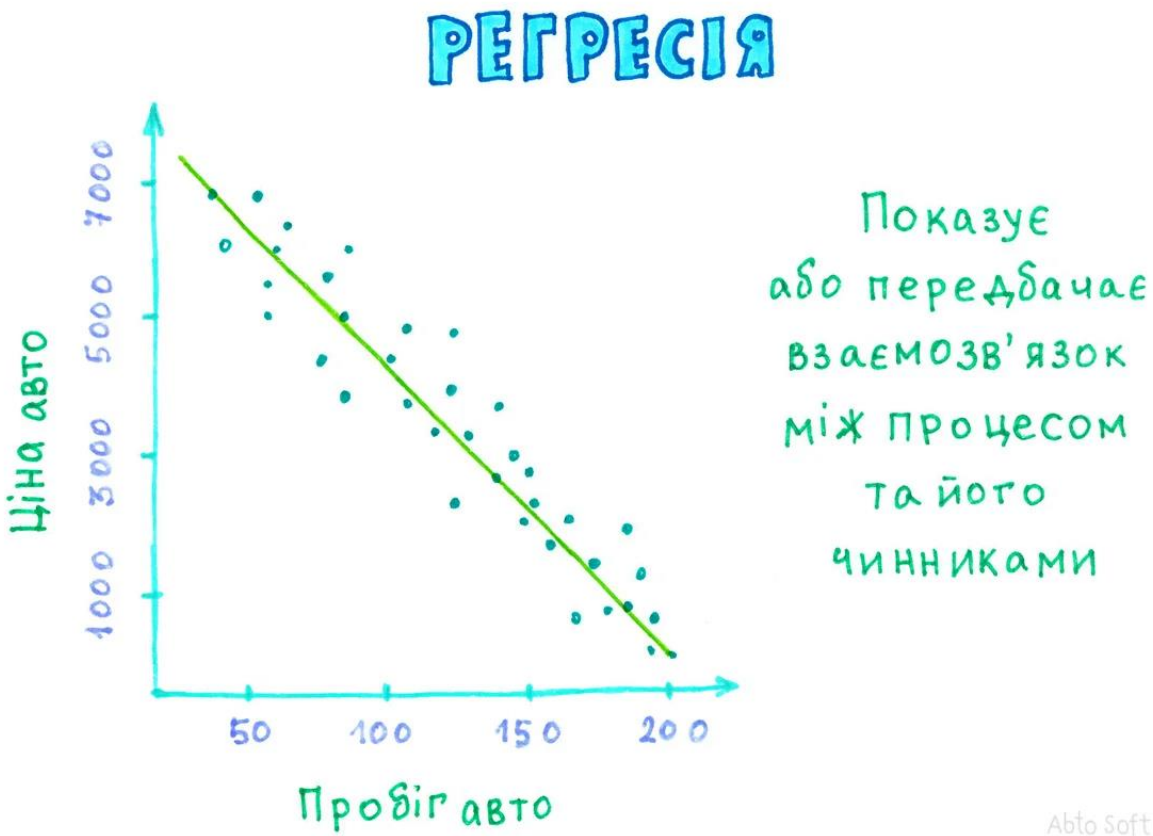

Машинное обучение
Классическое обучение
Без учителя
С учителем
Кластеризация
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Классификация
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Поиск правил
Euclat
Apriori
FP-Growth
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
Обучение с подкреплением
Генетический Алгоритм
A3C
SARSA
Q-Learning
Deep Q-Network (DQN)
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
AdaBoost
XGBoost
LightGBM
CatBoost
Нейросети и глубокое обучение
Convolutional Neural Networks (CNN)
DCNN
Перцептроны (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

Типи задач машинного навчання:

1. Регресія
2. Класифікація
3. Кластеризація
4. Прогнозування
5. Зменшення розмірності
6. Виявлення аномалій
7. Пошук правил

Регресія

Регресію дуже полюбляють аналітики та фінансисти, тому що вона може показати залежності між різними факторами процесу. Наприклад, як на ціну будинку впливає той чи інший район розташування?



Або що більше впливає на вартість авто: рік випуску чи пробіг? Машина намагається побудувати криву на графіку, яка відображає залежність. Але, на відміну від людини з крейдою та дошкою, робить вона це з застосуванням математики.

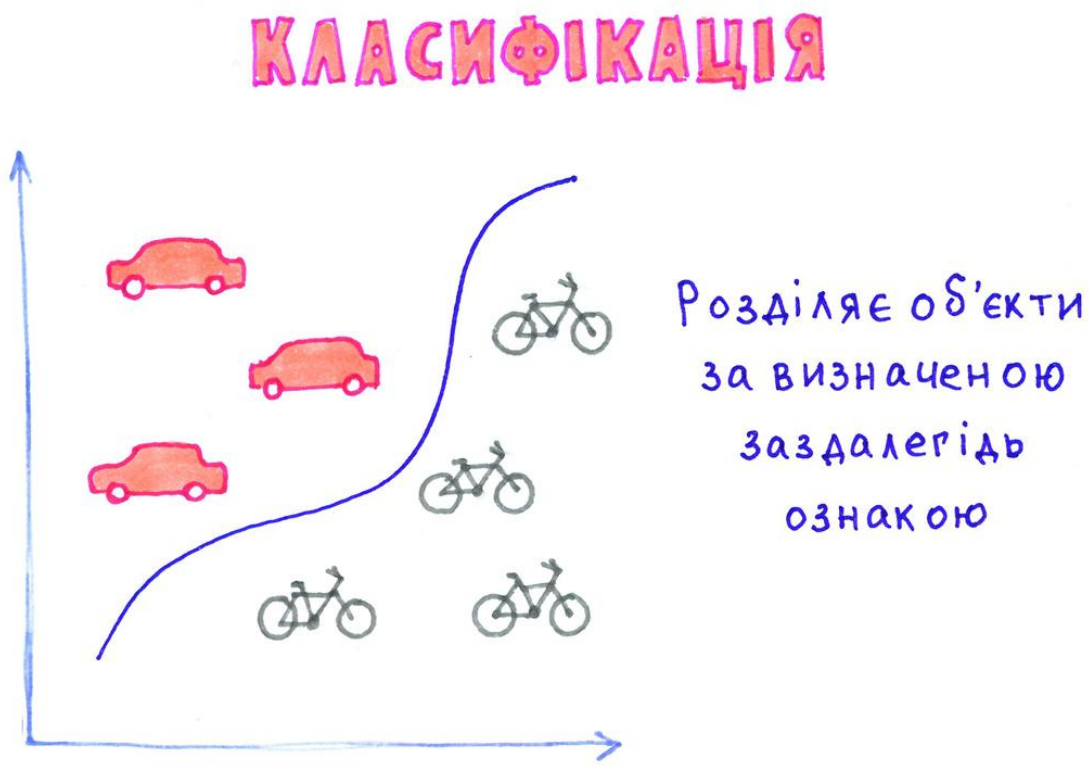
Регресія

Моделі регресійного аналізу зазвичай використовують, щоб показати або передбачити взаємозв'язок між процесом та тим, що цей процес може спровокувати. Тут варто пам'ятати, що така кореляція – не завжди причинність, тобто навіть пряма у простій лінійній регресії, яка добре відображає залежності між даними, може не дати конкретної відповіді про причинно-наслідковий зв'язок. Саме тому регресійний аналіз не використовують для **інтерпретації** причинно-наслідкових зв'язків між змінними. Однак такий аналіз може вказати, як і наскільки змінні пов'язані одна з одною, а визначення причин і наслідків – це предмет глибших досліджень за допомогою інших алгоритмів та методів.

Графік регресії може показати позитивний зв'язок, негативний зв'язок або відсутність зв'язку процесу з тими чи іншими факторами. Якщо лінія регресії є горизонтальною або вертикальною – між змінними ніякого зв'язку немає. Тобто, якщо повернутися до прикладу з авто та пробігом, де на одній осі буде ціна авто, а на іншій – його пробіг, горизонтальна або вертикальна пряма регресії буде означати, що пробіг ніяк не впливає на вартість авто. Якщо зі зростанням **x** зростає **y** (нижня частина графіка перетинає вісь, а верхня прагне в поле графіка) – залежність позитивна, тобто ціна буде рости зі збільшенням пробігу, якщо навпаки – негативна, тобто що більший пробіг, то менша ціна.

Класифікація

Алгоритми класифікації дозволяють розділити об'єкти відповідно до зазначених заздалегідь класів, наприклад розділити кішок і собак, музику за жанром, шкарпетки за кольорами.



Найпростішим із технічного погляду завданням класифікації є бінарна класифікація – коли об'єкти потрібно розподілити між двома класами. Наприклад, на вході є транспортні засоби, та ми знаємо, що це або автомобілі, або велосипеди. Машина за певними алгоритмами розподіляє кожен з транспортних засобів до одного з визначених класів (авто або велосипед).

Класифікація

У багатокласовій класифікації кількість класів може досягати декількох тисяч, і рішення стає значно складнішим. Також бувають класи, що перетинаються, – у таких випадках об'єкт може одночасно належати до декількох класів, і нечіткі класи – коли належність до того чи іншого класу визначається ступенем (зазвичай від 0 до 1).

Сьогодні алгоритми класифікації використовують для великої кількості завдань: визначення мови, спам-фільтрів, визначення шахрайства (коли зловмисник сплачує за послуги вкраденими коштами), пошуку схожих документів тощо.

Щоб класифікація спрацювала, потрібно мати розмічені дані з категоріями і ознаками, які машина буде вчитися визначати. Залежно від певних ознак алгоритм визначає, до якого з класів можна віднести об'єкт.

Кластеризація

Алгоритми кластеризації дозволяють розділити об'єкти у випадку, коли класи заздалегідь не зазначені, а кластери мають бути сформовані за схожістю елементів. Алгоритми кластеризації використовують у завданнях сегментації ринку, для аналізу нових даних, стиснення зображень. Машина визначає схожі ознаки у об'єктів та групує їх у кластери.

Кількість кластерів, як із класами, може бути визначена дослідником або самою машиною. Також дослідник може задати ознаки, за якими потрібно розділити вибірку, або машина визначить їх сама.

Прогнозування

Сьогодні вигадування сценаріїв майбутнього більше не звучить як сюжет фантастичної книги, а впало на плечі математиків і фахівців із машинного навчання. Прогностичних моделей існує безліч, і всі вони вирішують завдання передбачення часових рядів – знаходження майбутніх значень залежно від часу.

Прогнозування використовують у медичній діагностиці, оцінці кредитоспроможності, передбаченні попиту, під час прийняття рішень на фінансових ринках.

Прогнозування

Алгоритм при побудові прогнозу включає в себе наступні елементи:

а) Аналіз часової послідовності на предмет наявності пропущених значень і значень, що випадають, та корекція цих значень.

Для знаходження пропущених значень і значень, що випадають, у свою чергу, існують окремі алгоритми. Під час збору реальних даних за деякий час вони можуть зникнути або бути введеними неправильно. Тому їх потрібно визначити експериментально.

б) Визначення наявності тренду і його типу. Визначення періодичності в послідовності.

в) Перевірка послідовності на стаціонарність, тобто що процес у майбутньому буде розвиватися так само, як в минулому і сьогоденні. У реальній природі таких процесів не існує, але процеси, характеристики яких змінюються дуже повільно, можна зарахувати до стаціонарних. Наприклад, кількість хлібу, яку українці з'їдають щодня, практично не залежить від погоди та пори року або інших зовнішніх факторів. І хоча ця цифра буде змінюватися, її динаміка буде незначною, тому такий процес можна сміливо називати стаціонарним.

Прогнозування

Алгоритм при побудові прогнозу включає в себе наступні елементи:

г) Аналіз послідовності на предмет необхідності попередньої обробки.
Усі дані повинні мати однакові характеристики й не відрізнятися один від одного.

ґ) Вибір моделі залежно від даних (стаціонарні чи ні) та бажаного результату (короткотерміновий чи довготерміновий прогноз тощо).

д) Визначення параметрів моделі. Прогноз на підставі обраної моделі.

е) Оцінка точності прогнозування моделі. Для цього прогноз зазвичай будують для вже відомих даних і порівнюють їх. Наприклад, у нас є дані про ціни на пальне з 2000 по 2018 роки. Для оцінки точності моделі ми будуємо прогноз цін на 2018 рік на підставі даних з 2000 по 2017 роки й порівнюємо отримані дані з реальними.

є) Аналіз похибки обраної моделі.

ж) Визначення адекватності обраної моделі і, якщо результат виявиться незадовільним (недостовірним чи не достатньо точним), її заміна і повернення до попередніх пунктів.

Зменшення розмірності

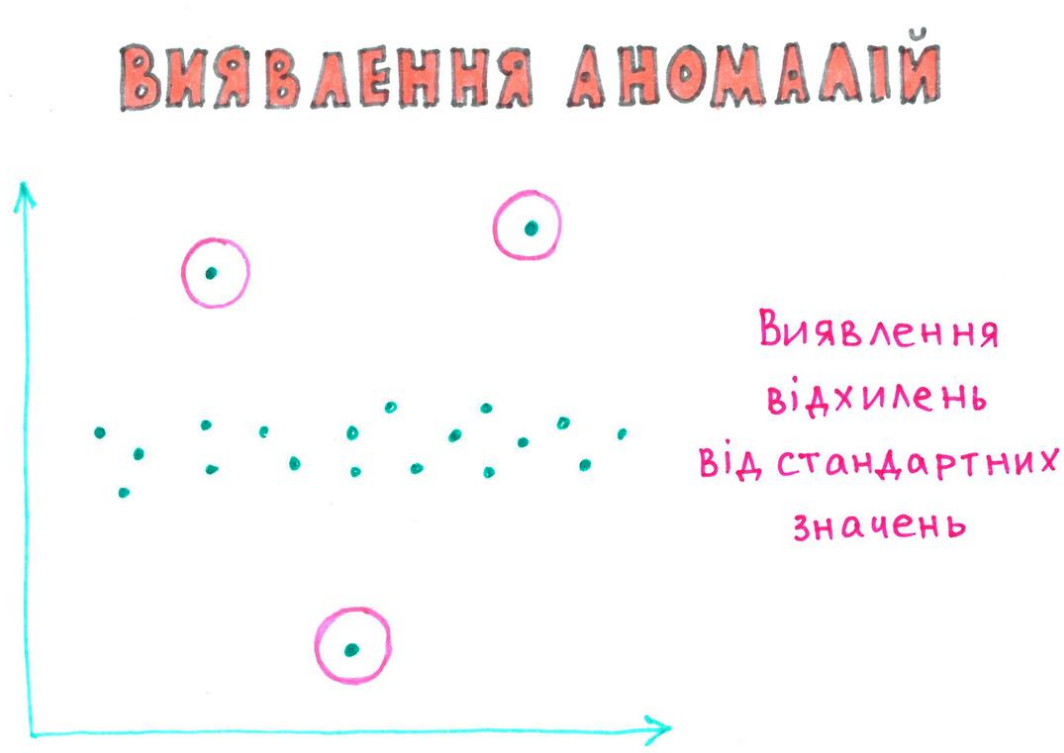
Мета – зведення більшого числа ознак до меншого для зручності їх подальшого використання. Використовується для побудови рекомендаційних систем, пошуку схожих документів чи візуалізації.

Завдання – задача, у якій машина з величезної кількості даних виявляє ті, що мають будь-які закономірності. Практична користь методів зменшення розмірності в тому, що можна отримати абстракцію, об'єднавши кілька ознак в одну: автомобілі вище 2 м із двома великими задніми колесами – це трактори.

Ще одне застосування завдань зменшення розмірності – рекомендаційні системи, алгоритми, що намагаються передбачити, які об'єкти будуть цікаві користувачеві, маючи певні дані про його профіль. Такі системи не враховують оцінку всіх користувачів, а спираються лише на переваги самого користувача або користувачів зі схожим портретом, і пропонують вам, приміром, подивитися певний фільм

Виявлення аномалій

Мета – виявити аномальні відхилення від стандартних випадків. На перший погляд, завдання дуже схоже на завдання класифікації, проте воно має істотну відмінність: аномалії – явище рідкісне, тому вибірок, на яких можна навчити машинну модель або дуже мало, або немає зовсім. Саме тому методи класифікації тут не працюють.



Наприклад, клієнт банку багато років знімає зі свого рахунку приблизно однакові суми з приблизно однаковою періодичністю, перебуваючи в Україні. Та одного разу приходить запит на підтвердження транзакції з цього рахунку на велику суму, проте людина, яка робить запит, перебуває в Індії. Така транзакція, ймовірно, буде вважатися аномалією й потребуватиме додаткової перевірки. На практиці аномалії допомагають виявити шахрайство в банку, медичні проблеми або помилки в тексті.

Пошук правил

Завдання пошуку правил шукає закономірності в потоці даних. Наприклад, якщо в супермаркеті між полицею з пивом і касою поставити стійку з горішками, з високою ймовірністю кількість проданих горішків зросте. Але тут все просто – горішки часто купують із пивом, а коли товарів багато, виявлення таких правил може істотно збільшити прибуток.

Ще більш істотним є передбачення поведінки користувача на інтернет-ресурсах. За яким товаром він повернеться? Які товари можна продати «навздогін» до вже заданого? На які розділи сайту направити користувача, щоб він залишив у вас більше грошей?

Постановка задачі навчання

- Розділити дані на дві частини:
  - навчальна вибірка (більша частина)
  - тестувальна вибірка
- Навчити машину за існуючою базою даних (навчальною вибіркою) приймати необхідне рішення – побудувати алгоритм прийняття рішень
- Перевірити побудований алгоритм на тестувальній вибірці

Приклади прикладних задач

Задачі класифікації

Задачі медичної діагностики

Об'єкт – пацієнт у певний момент часу

Можливі класи: діагноз, спосіб лікування, результат захворювання

Приклади можливих ознак

- **Бінарні**: стать, головний біль, слабкість, нудота, та ін.
- **Кількісні**: зріст, вага, тиск, пульс, кількість гемоглобіну в крові, доза препарату, і т.д.

Особливості задачі

- Зазвичай багато пропусків в даних
- Необхідно виділяти синдроми – поєднання симптомів
- Необхідно оцінювати імовірність негативного результату

Приклади прикладних задач
Задачі класифікації
Задачі кредитного скорингу

Об'єкт – заявка на видачу банком кредиту

Можливі класи: 1 (так) або 0 (ні)

Приклади можливих ознак

- **Бінарні**: стать, наявність телефону, наявність автомобіля, наявність нерухомого майна, наявність кредитів в інших банках, та ін.
- **Номінальні**: місце проживання, місце роботи, посада
- **Кількісні**: зарплата, дохід родини, сума кредиту, і т.д.

Особливості задачі

- Необхідно оцінювати імовірність дефолту

Приклади прикладних задач

Задачі класифікації

Задача передбачення відтоку клієнтів

Об'єкт – абонент у певний момент часу

Можливі класи: 1(так - піде) або 0(ні – не піде)

Приклади можливих ознак

- **Бінарні**: чи корпоративний клієнт, чи користується послугами, та ін.
- **Номінальні**: тарифний план, регіон проживання
- **Кількісні**: тривалість дзвінків (вхідних та вихідних), кількість повідомлень (СМС), кількість інтернету (МБ), частота оплати послуг, і т.д.

Особливості задачі

- Дуже великі вибірки даних
- Важко виділяти ознаки за «сирими» даними
- Необхідно оцінювати імовірність відтоку клієнта

Приклади прикладних задач

Задачі регресії

Задача прогнозування вартості нерухомості

Об'єкт – квартира в Києві

Передбачити – вартість квартири

Приклади можливих ознак

- **Бінарні**: наявність метро поблизу, балкону, ліфта, охорони, паркінгу, та ін.
- **Номінальні**: район міста, тип дому (цегляний/панельний/блочний/моноліт)
- **Кількісні**: кількість кімнат, площа квартири, поверх, відстань до метро, вік будинку, і т.д.

Особливості задачі

- Вибірка неоднорідна, вартість змінюється з часом
- Різні ознаки
- Необхідно перетворення ознак у числові

Приклади прикладних задач

Задачі регресії

Задача відкриття нового ресторану

Об’єкт – місце для відкриття нового ресторану

Передбачити – прибуток ресторану через рік

Приклади можливих ознак

- Демографічні дані: середній вік потенційних клієнтів, їх достаток, та ін.
- Ціни на нерухомість поблизу
- Маркетингові дані: наявність поблизу шкіл, офісів, метро і т.д..

Особливості задачі

- Мало об’єктів, багато ознак
- Різні ознаки
- Різнорідні об’єкти (можливо краще будувати окремі моделі для великих та малих міст)

Машинне навчання на даних складної структури

- **Статистичний машинний переклад**
  Об'єкт – речення мовою оригіналу
  Відповідь – речення на іншій мові
- **Переведення мови і текст**
  Об'єкт – аудиозапис мови людини
  Відповідь – текстовий запис мови
- **Комп'ютерний зір (автопілот автомобіля)**
  Об'єкт – зображення або відеоряд
  Відповідь – рішення (об'їхати, зупинитися, ігнорувати)

Передумови успішного розв'язку задач зі складними даними

- Великі та *чисті* дані (BigData)
- Методи зменшення розмірності вхідних даних
- Методи оптимізації для задач великої розмірності
- Глибокі нейромережеві архітектури
- Зростання обчислювальних потужностей (GPU)

Особливості даних та ризики

Особливості даних

- Різнорідні (ознаки виміряні в різних шкалах)
- Неповні (виміряні не всі, є прогалини)
- Неточні (є похибки)
- Суперечливі (об'єкти однакові, а відповіді різні)
- Надлишкові (дуже великі, не вміщуються в пам'ять машини)
- Недостатні (об'єктів менше ніж ознак)
- Неструктуровані (відсутній ознаковий опис)

Ризики, пов'язані з постановкою задачі

- «брудні» дані (замовник не забезпечує якість даних)
- Незрозумілі критерії якості моделі (замовник не визначився з метою досліджень чи певними індикаторами)

Дякую за увагу