ОСНОВИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ, НЕЙРОННИХ МЕРЕЖ та ГЛИБОКОГО НАВЧАННЯ

Модуль 1. ШТУЧНИЙ ІНТЕЛЕКТ. МАШИННЕ НАВЧАННЯ

Лекція 1.3. Датасети

НАБОРИ ДАНИХ
DATASETS

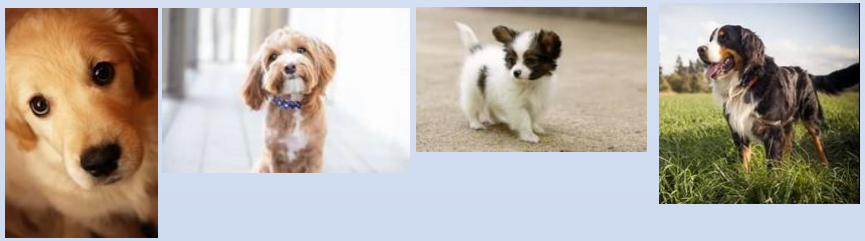
Набір даних

Формально: **Dataset → Набір даних** → колекція спеціальним чином організованих даних, що застосовується в задачах машинної обробки даних.

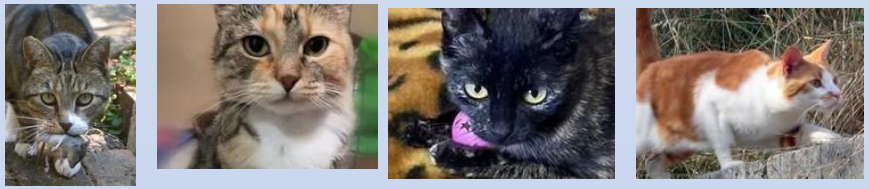
ДАННІ

DATASET

Dogs



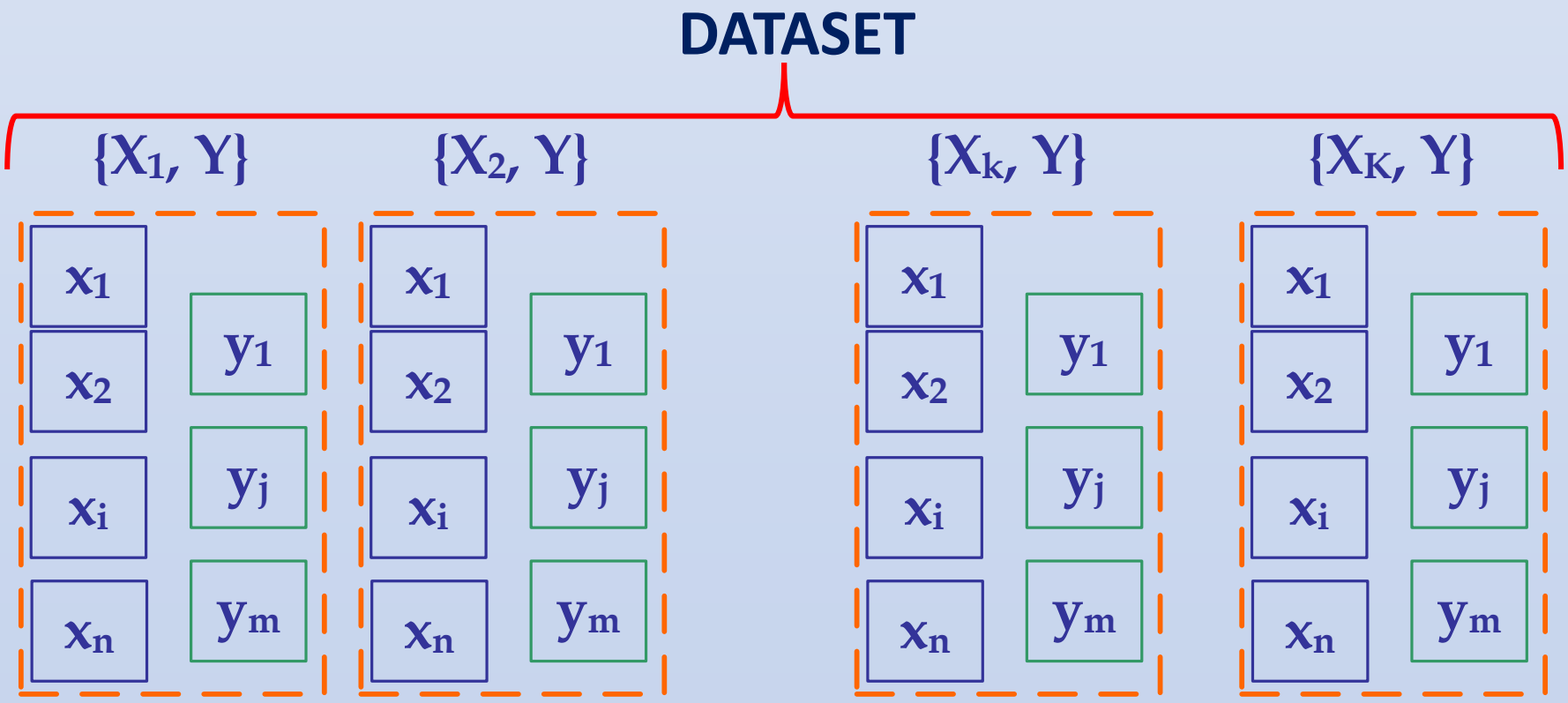
Cats



Birds

Набір даних

Набір даних часто складається з пар векторів (або скалярів) входу – ознак (**features**) та відповідних векторів (або скалярів) виходу, зазвичай позначують як ціль (**target**), або мітка (**label**).



Ознаки

Ознака - окрема властивість або характеристика спостережуваного об'єкту (явища), яку можливо виміряти. Обрання інформативних, розрізнювальних і незалежних ознак є ключовим кроком алгоритмів AI, ML.

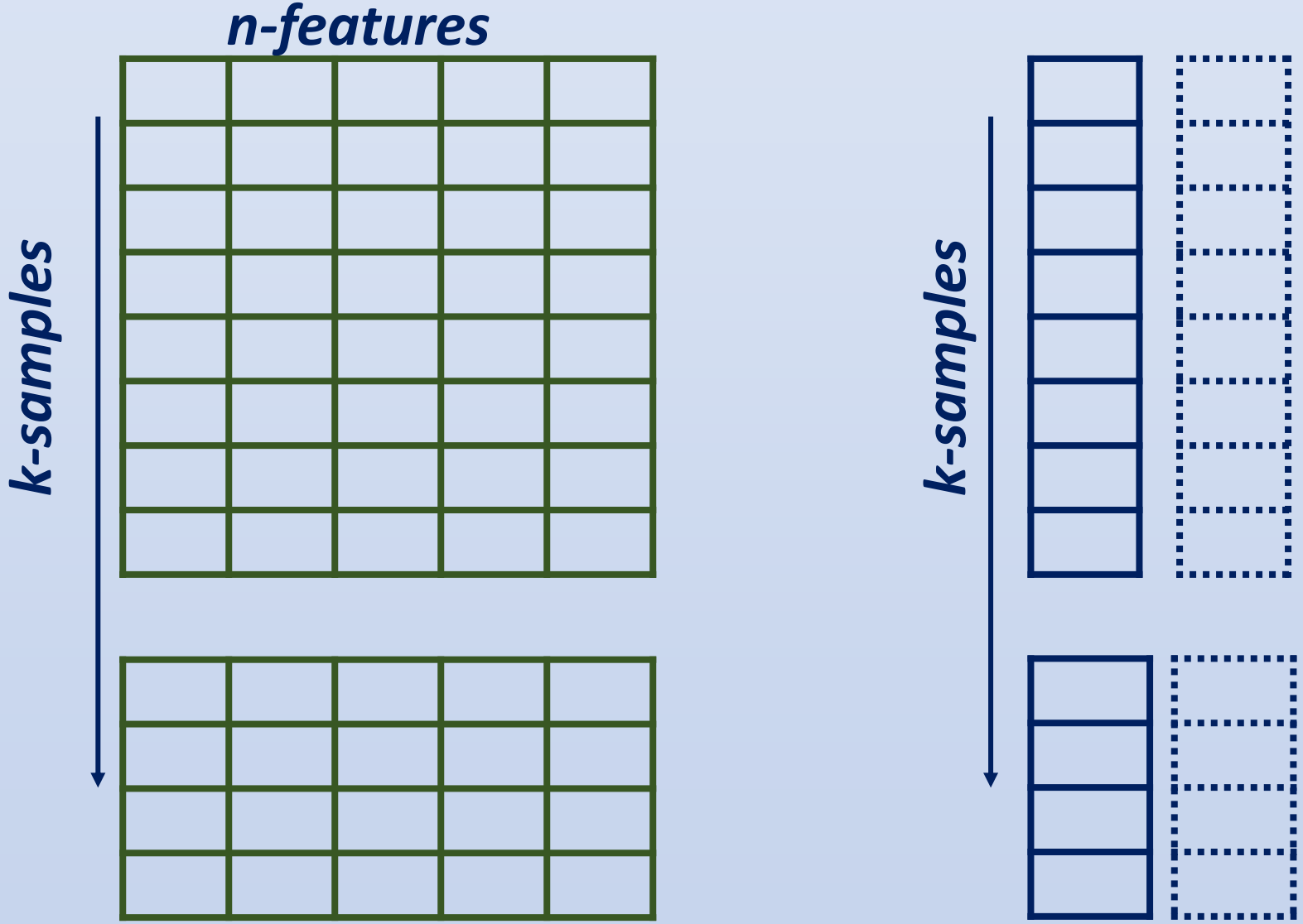
Вектор ознак (**feature vector**) — ***n***-вимірний вектор числових ознак, що представляють певний об'єкт. Векторний простір, пов’язаний з цими векторами – простір ознак (**feature space**).

Вектор (скаляр) міток (**label vector**) — ***m***-вимірний вектор цільових значень (те, що потрібно передбачити)

Типова структура набоу даних

Матриця ознак

Вектор цілей



Використання dataset

Типова, на різних етапах створення ML моделі використовують три набори даних.

1. Модель початково налаштовують на тренувальному наборі даних (**training dataset**), який є набором прикладів, що використовують для допасовування параметрів моделі (ваг в ANN).

2. Випробувальний набір даних (**test dataset**) — набір даних, який використовують для забезпечення неупередженої оцінки допасованості (узгодженості) моделі на тренувальному наборі даних.

Використання dataset

3. Оцінка допасованості моделі виконується на затверджувальному наборі даних (**validation dataset**).

Затверджувальний набір даних забезпечує неупереджену оцінку допасованості моделі на тренувальному наборі даних при налаштовуванні гиперпараметрів моделі.

УВАГА Терміни випробувальний (test set) набір та затверджувальний (validation set) набір, часто використовують таким чином, що їхні значення **міняються місцями**. Тобто, «випробувальний набір» стає розробницьким набором (**development set**), а «затверджувальний набір» - набором, що використовують для оцінювання продуктивності повністю визначеної моделі.

Набір даних

Окремі набори даних широко використовуються в академічних колах як **стандартні (тестові) набори**, що підтверджують результати наукових досліджень.

Деякі набори даних є **відкритими** для використання, інші надаються за, звичайно символічну, плату.

Scikit-learn datasets

Іграшкові датасети - невеликі стандартні набори даних, для яких не потрібно завантажувати будь-який файл із зовнішнього веб-сайту.

Датасети **реального світу** - інструменти для завантаження великих наборів даних, завантажуючи їх у разі необхідності.

Датасети, **що генеруються** - генератори випадкових вибірок, які можна використовувати для створення штучних наборів даних контрольованого розміру та складності.

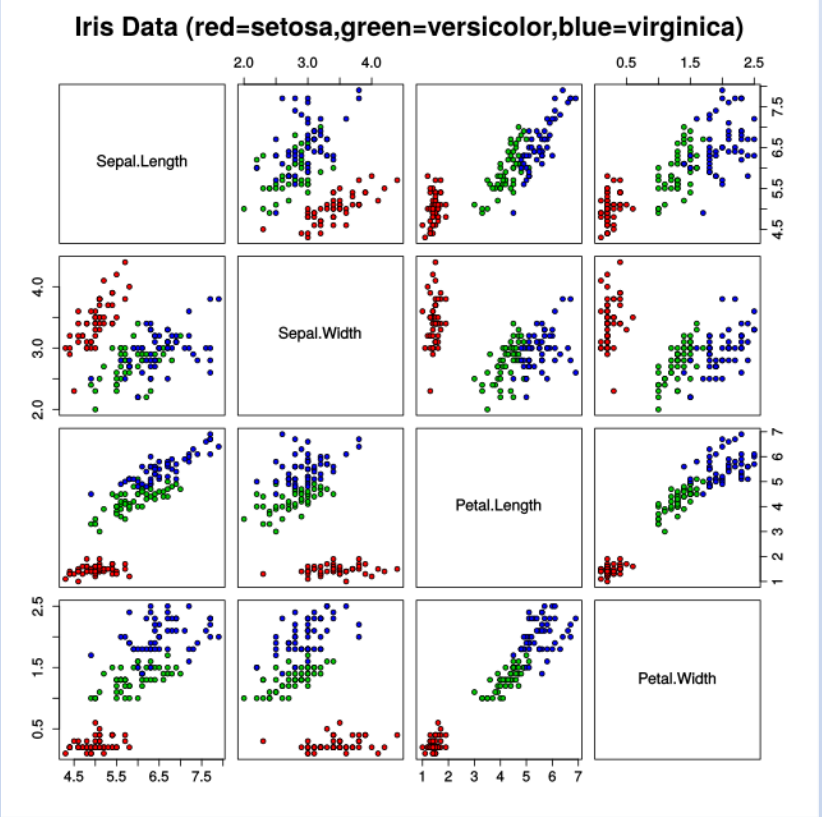
Зразки **зображень** JPEG, опублікованих їх авторами за ліцензією Creative Commons. Зображення можуть бути корисними для тестування алгоритмів і конвеєрів на 2D-даних.

Scikit-learn IRIS

IRIS dataset складається з даних про 150 вимірювань ірисів трьох видів — Iris setosa, Iris virginica і Iris versicolor, по 50 вимірювань на кожен вид.

Для кожного екземпляра наведено чотири ознаки :

- Довжина зовнішньої частки чашолистика (sepal length);
- Ширина зовнішньої частки чашолистика (sepal width);
- Довжина внутрішньої частки пелюстки (petal length);
- Ширина внутрішньої частки пелюстки (petal width).



Scikit-learn Diabet

Diabet dataset створено на даних Національного інституту діабету, захворювань органів травлення та нирок. Мета полягає в тому, щоб передбачити на основі діагностичних вимірювань, чи є у пацієнта діабет.

Включено 442 екземпляра (пацієнти). Для кожного визначено десять ознак

- вік,
- стать,
- індекс маси тіла,
- середній артеріальний тиск,
- шість вимірювань сироватки крові.

Мітка

- 1 → tested positive for diabetes.
- 0

Scikit-learn BLOBS

Генератори виробляють матрицю ознак і відповідні дискретні цілі.

make_blobs
make_classification
make_gaussian_quantiles
make_multilabel_classification

Приклади наведені в

2024_AI_TF_lec_08_Exmpl_1_IRIS.pdf

2024_AI_TF_lec_08_Exmpl_2_Digits.pdf

2024_AI_TF_lec_08_Exmpl_3_Blobs.pdf

Tensorflow datasets

TF надає множину готових наборів даних для використання різними платформами машинного навчання.

TF дозволяє створювати складні конвеєри вхідних даних з простих, повторно використовуваних частин.

Наприклад, контейнер для текстової моделі може включати в себе вилучення символів із необроблених текстових даних, перетворення їх у вбудовані ідентифікатори за допомогою таблиць пошуку та об'єднання в пакети послідовностей різної довжини.

API tf.data дозволяє обробляти більші обсяги даних, обчислювати дані з різних форматів і виконувати складні перетворення.

Tensorflow MNIST

MNIST - база даних зразків рукописного написання цифр, марковані вручну

60000 напівтонових зображень для навчання та 10000 зображень для тестування

Всі зображення пройшли згладжування, приведені до розміру 28 X 28 пікселів

EMNIST - база даних 800000 рукописних символів (28 X 28 пікселів)

Tensorflow datasets

Деякі приклади наведені в

2024_AI_TF_lec_08_Exmpl_4_TF_Comm.pdf

2024_AI_TF_lec_08_Exmpl_5_TF_MNIST.pdf

2024_AI_TF_lec_08_Exmpl_6_TF_CIFAR.pdf

Початкові datasets

Контрольні запитання

- **Пояснить сутність побудови датасету для машинного навчання**
- **Надайте класи вбудованих датасетів в бібліотеці Scikit-learn**
- **Надайте класи вбудованих датасетів в бібліотеці Tensorflow**

Корисні та цікави посилання

- **Tensorflow datasets**

https://www.tensorflow.org/datasets/catalog/overview#all_datasets

- **Scikit-learn datasets**

https://scikit-learn.org/stable/datasets.html

- **Список наборів даних**

https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Рекомендована ЛІТЕРАТУРА

- **Глибинне навчання:** Навчальний посібник / Уклад.: В.В. Литвин, Р.М. Пелещак, В.А. Висоцька В.А. – Львів: Видавництво Львівської політехніки, 2021. – 264 с.
- Тимощук П. В., Лобур М. В. **Principles of Artificial Neural Networks and Their Applications: Принципи штучних нейронних мереж та їх застосування:** Навчальний посібник. – Львів : Видавництво Львівської політехніки, 2020. – 292 с.
- Morales M. **Grokking Deep Reinforcement Learning. –** Manning, 2020. – 907 с.
- Trask Andrew W. **Grokking Deep Learning.** – Manning, 2019. – 336 с.

The END
Модуль 1. Лекція 1.3.