hillel
IT school

Lesson 04
KNN, K-means, Linear Regression

In the Previous Lessons

- Difference between AI, ML and DL
- Supervised, unsupervised, reinforcement, semi-supervised, self-supervised learnings
- ML project stages
- Data splits
- Datasets
- Python ML tools basic review

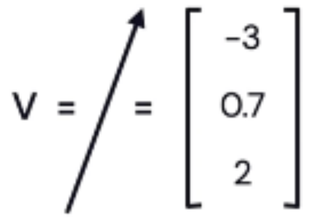
Math in ML: Data Representations

Як ми можемо представити дані (зображення, текст, налаштування користувача тощо) у спосіб, зрозумілий комп'ютерам?

Ідея: організувати інформацію у вектор

A **vector** is a 1-dimensional array of numbers.
It has both a *magnitude* (length) and a *direction*.

The totality of all vectors with *n* entries is an
n-dimensional vector space.



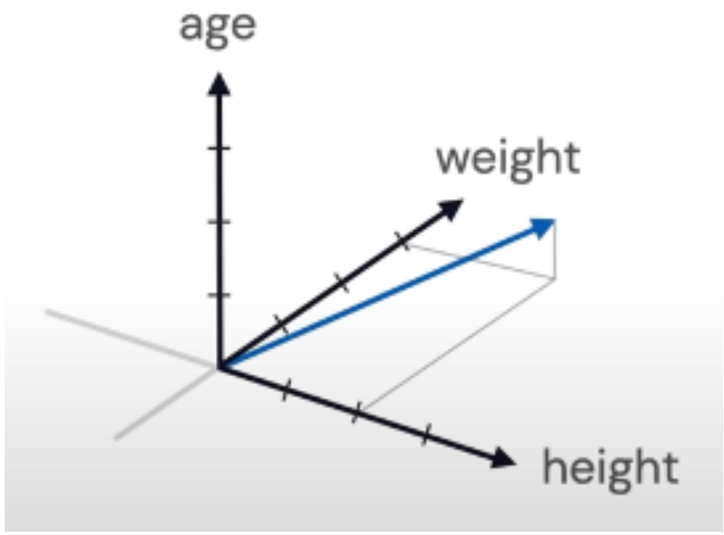
Math in ML: Data Representations

У контексті ML:

Вектор ознак — це вектор, записи якого представляють «особливості» деякого об’єкта.

$$p = \begin{bmatrix} 64 \\ 131 \\ 23 \end{bmatrix} \begin{matrix} \text{height} \\ \text{weight} \\ \text{age} \end{matrix}$$

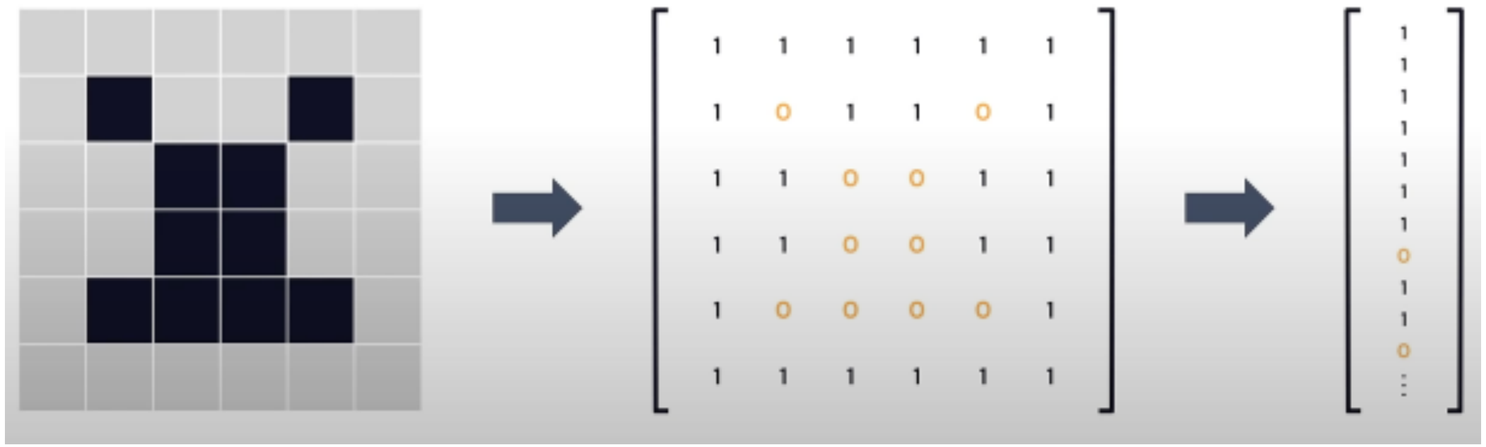
“p” for “patient”



Векторний простір, що їх містить, називається простором ознак.

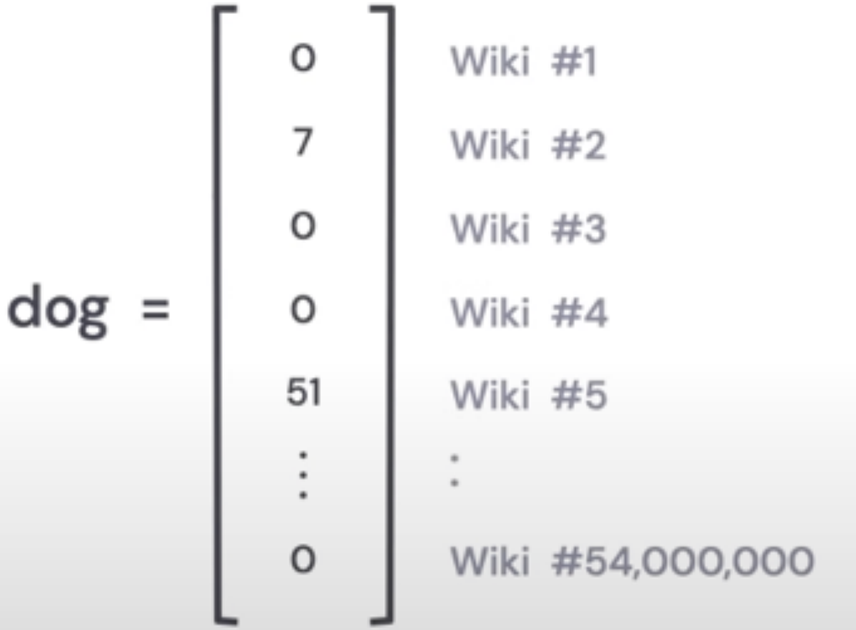
Example of Data Representations: Images

У чорно-білих зображеннях чорні та білі пікселі відповідають 0 і 1. Пікселі відтінків сірого – це числа від 0 до 255.
Обидва збираються в одновимірний масив чисел.



Example of Data Representations: Words & Documents

Дано набір документів, призначте кожному слову вектор, і-й запис якого є кількістю разів, коли слово з'являється в і-му документі.



Example of Data Representations: One-Hot Encodings

Призначте кожному слову вектор з однією 1 і 0 в іншому місці. Це називається one-hot кодуванням (або «стандартним базовим вектором»).
Наприклад, припустимо, що наша мова має лише чотири слова:

KNN

K-найближчих сусідів (K-nearest neighbors або KNN) - це алгоритм машинного навчання, який використовується як для задач класифікації, так і для задач регресії. Основна ідея полягає в тому, що об'єкти, які схожі за значеннями ознак, мають схожі класи або значення цільової змінної.

K-means

K-means є алгоритмом кластеризації, який використовується для групування об'єктів в кластери на основі їхньої схожості. Головна ідея полягає в тому, щоб розділити набір даних на K кластерів так, щоб об'єкти всередині одного кластеру були подібні між собою, а об'єкти з різних кластерів були відмінні.

KNN vs K-means

KNN (K-Nearest Neighbors) та K-Means - це два різні алгоритми машинного навчання, які використовуються для різних завдань. Основні відмінності між ними:

1. Тип завдання:
- KNN: Використовується для задач класифікації та регресії. Визначає клас об'єкта або прогнозує значення на основі його найближчих сусідів.
- K-Means: Використовується для задач кластеризації. Групує об'єкти в кластери на основі їхньої схожості.
2. Тип алгоритму:
- KNN: Алгоритм, що навчається на основі тренувальних даних і намагається визначити клас (або значення) нового об'єкта, порівнюючи його з тренувальними прикладами.
- K-Means: Кластерний алгоритм, оскільки він групує об'єкти в кластери з певною кількістю центроїдів.
3. Призначення:
- KNN: Застосовується для індивідуальних прогнозів для нових об'єктів на основі його найближчих сусідів.
- K-Means: Використовується для групування об'єктів в кластери, здебільшого для виявлення структури та закономірностей в даних.
4. Визначення параметрів:
- KNN: Вимагає визначення параметра K (кількість найближчих сусідів), який визначається перед використанням алгоритму.
- K-Means: Вимагає визначення кількості кластерів K перед використанням алгоритму.
5. Тренування:
- KNN: Немає явного тренування. Модель вивчається в момент прийняття рішення для нових об'єктів.
- K-Means: Вимагає ітеративного тренування, де центроїди перераховуються та призначення кластерів оновлюються на кожній ітерації.
6. Вимір відстані:
- KNN: Використовується відстань між признаками для визначення найближчих сусідів (зазвичай Евклідова відстань).
- K-Means: Використовується відстань між точками у просторі ознак для обчислення схожості між об'єктами та центроїдами.

DEMO

Check `knn_kmeans.ipynb`

Regression

Регресія — це статистичний метод, який використовується для аналізу та моделювання зв’язку між залежною змінною та однією або кількома незалежними змінними. Простими словами, регресія дозволяє знайти найкращу лінію або криву, яка описує зв’язок між змінними.

Регресію можна використовувати для передбачення, прогнозування та визначення факторів, які впливають на змінну результату.

Regression

Існує кілька типів регресійного аналізу, зокрема:

- Linear Regression
- Polynomial regression
- Logistic regression
- Ridge regression
- Lasso regression
- Elastic Net Regression
- Time-series regression
- Bayesian regression
- Nonlinear regression
- Poisson regression

Linear Regression

Лінійна регресія — це статистичний метод, який використовується для моделювання зв'язку між залежною змінною (зазвичай позначається «y») та однією або кількома незалежними змінними (зазвичай позначається «x»). Мета лінійної регресії — знайти лінійне рівняння, яке найкраще описує зв'язок між змінними. Це лінійне рівняння називається рівнянням регресії, і воно використовується для прогнозування значення залежної змінної для заданого значення незалежної змінної (змінних).

Вважається, що зв'язок між залежною змінною та незалежною змінною (змінними) є лінійним, що означає, що для наближення співвідношення можна використовувати пряму лінію.

Linear Regression

Лінія найкращого підходу виражається математичним рівнянням у формі:

$$y = wx + b$$

Де:

y - залежна змінна

x - незалежна змінна

w - нахил лінії, яка представляє зміну `y` для зміни `x` на одиницю

b - точка перетину прямої, яка представляє значення `y`, коли `x` дорівнює нулю

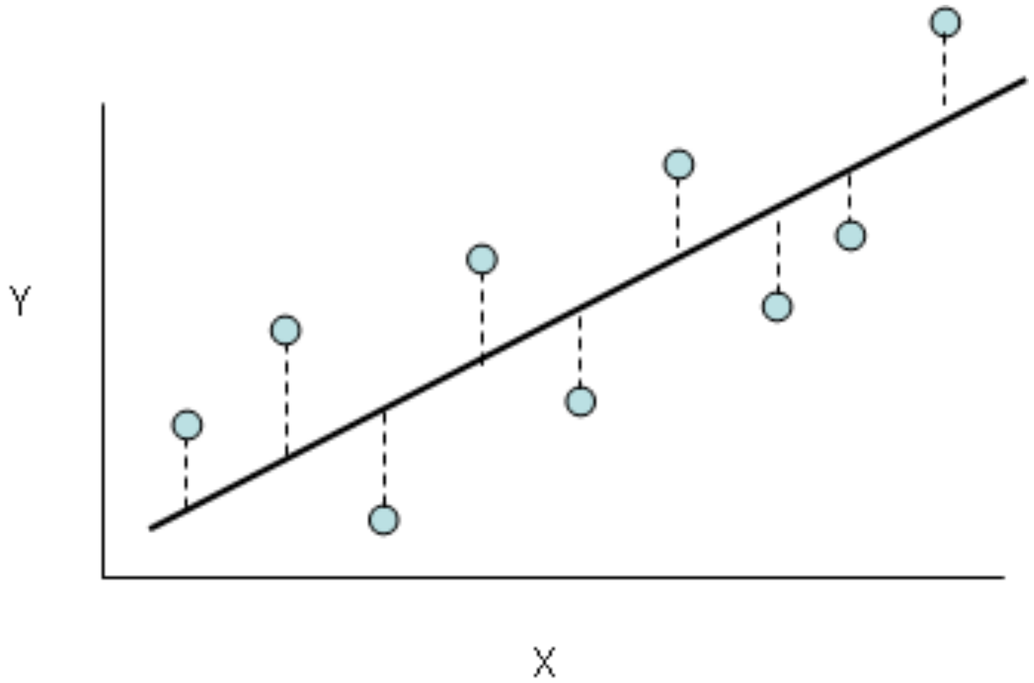
Linear Regression

Існує три основних типи лінійної регресії:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

Linear Regression

Реалізація моделі лінійної регресії має 5 основних компонентів: модель, функцію вартості (cost function), параметри, градієнт і алгоритм оптимізації (наприклад, нормальне рівняння, градієнтний спуск).



Gradient Descent

Градієнтний спуск - це оптимізаційний алгоритм, який використовується для пошуку мінімуму функції. Його основна ідея полягає в тому, щоб крок за кроком рухатися в напрямку найшвидшого зменшення значення функції (градієнту), з метою знаходження локального мінімуму.

DEMO

Check `linear_regression.ipynb`

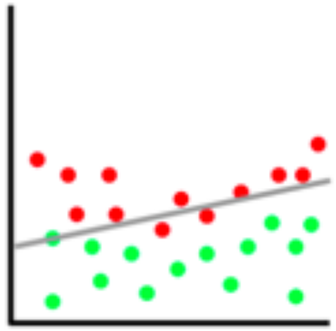
Overfitting & Underfitting

- Що таке overfitting?
- Що таке underfitting?

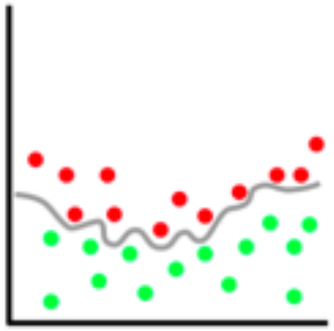
Overfitting & Underfitting

- Overfitting – хороша продуктивність на навчальних даних, погане узагальнення на інші дані
- Underfitting – низька продуктивність даних навчання та погане узагальнення на інші дані

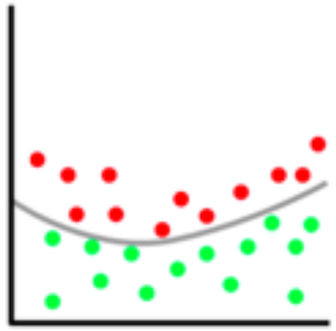
Overfitting & Underfitting



Underfitting

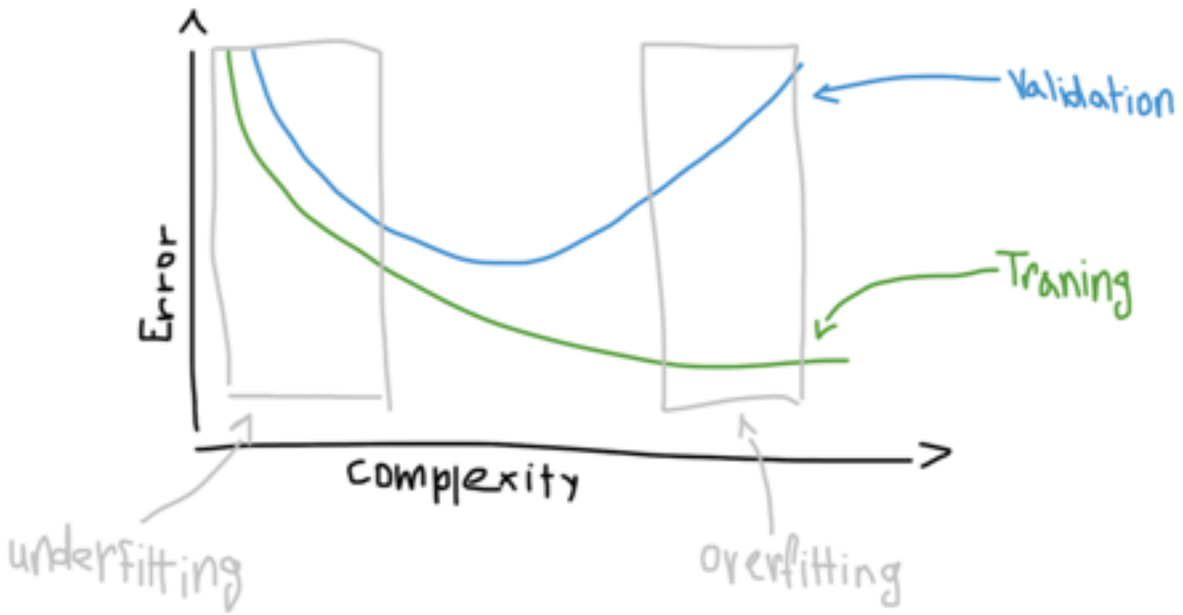


Overfitting



Balanced

Overfitting & Underfitting



hillel
IT school

Thanks for your attention