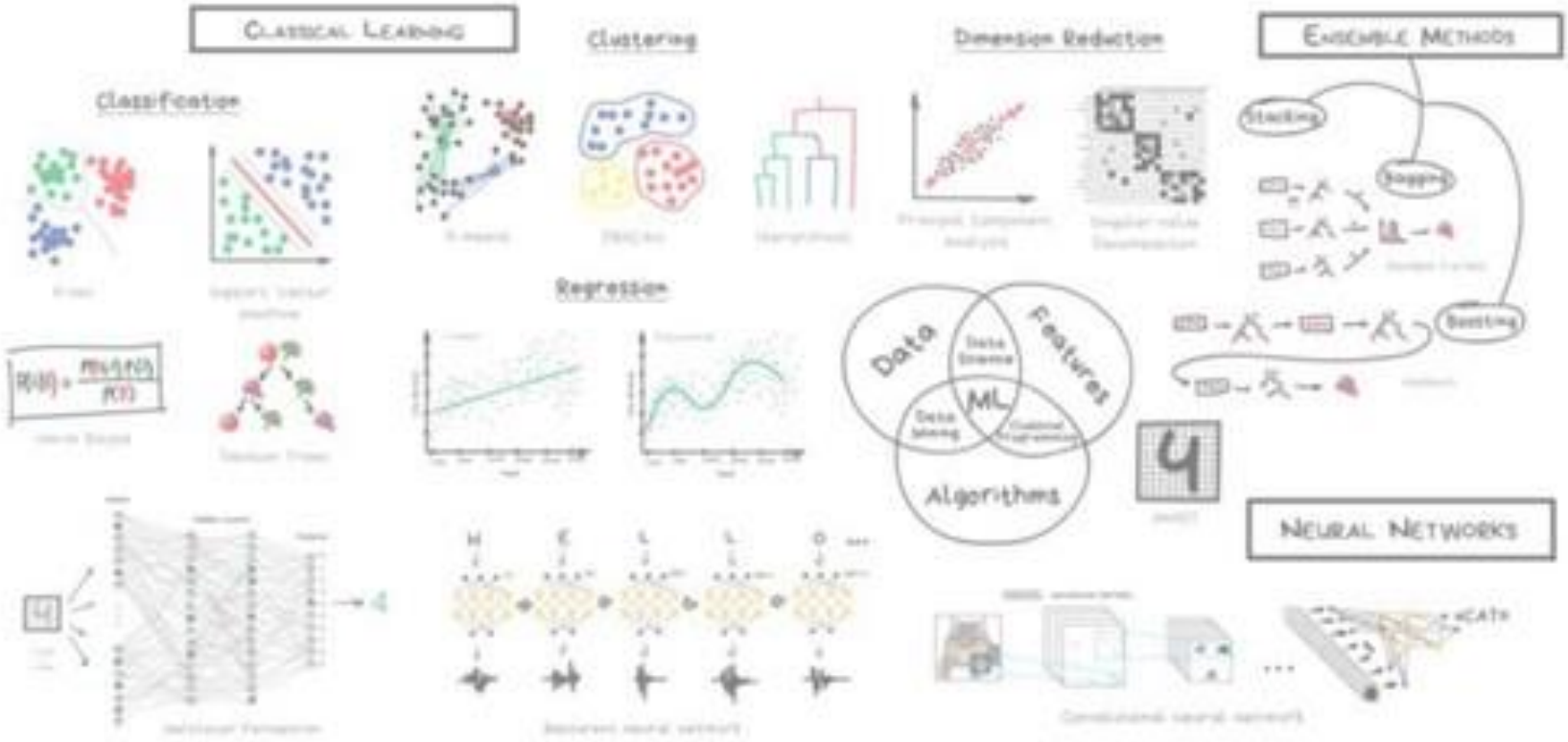
Машинне навчання

Метод дерева прийняття рішень (decision tree)



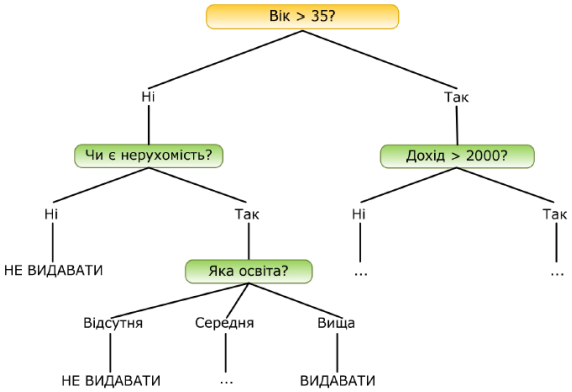
Лабораторна робота №1

Метод дерева прийняття рішень (decision tree)

Постановка задачі

Розглянемо задачу класифікації, в якій об'єктам з $X = \mathbb{R}^n$ (об'єкти описуються n числовими признаками): $\boldsymbol{X} = \{\boldsymbol{x_1} \dots \boldsymbol{x_n}\}$ відповідає один з кількох класів $\boldsymbol{Y} = \{\boldsymbol{0}, \dots \boldsymbol{M}\}$.
Нехай задана вибірка пар "об'єкт-відповідь": $(x_i, yi), i = 1 \dots N$.
Необхідно побудувати алгоритм класифікації $a(x)$: $X \to Y$.

Дерево прийняття рішень



Етапи розв'язання

1. Розділити всю вибірку на навчальну та тестову
2. Побудувати алгоритм навчання на навчальній вибірці
3. Перевірити точність роботи алгоритму на тестовій вибірці
4. Порівняти результати з Decision tree з sklearn
5. Оформити результати у вигляді звіту.

Метод дерева прийняття рішень (decision tree)

Етапи розв'язання

1. Додати до таблиці ознак навчальної вибірки колонку з цільовим вектором
2. Порахувати ентропію (показник Gini) для вихідної таблиці

$$S = -\sum_{i=1}^{M} p_i \log_2 p_i \quad \text{або} \quad S = \sum_{i=1}^{M} p_i(1 - p_i)$$

p_i – імовірність реалізації i-того класу

3. Відсортувати таблицю за першою ознакою.
4. Знайти значення ознаки для першого поділу таблиці, коли значення цільового вектору змінюється вперше.
5. Порахувати приріст інформації при такому діленні

$$Q = Q_0 - \sum_{i=1}^{M} \frac{N_i}{N} S_i$$

6. Повторити процедуру для всіх можливих ділень за першою ознакою та за всіма ознаками і знайти оптимальне перше ділення таблиці для якого приріст інформації є максимальним.
7. Рекурсивно для кожного ділення повторити процедуру ділення.
8. Вихід з рекурсії здійснити за умови
   - перевищення фіксованої глибини дерева
   - мінімальної кількості об'єктів у вузлі
   - мінімального значення ентропії вузла

Метод дерева прийняття рішень (decision tree)
Приклад результатів

У якості результатів роботи програми отримуємо словник з правилами рухів по дереву:

Мітка:

“” – корінь дерева
“l” – гілка вліво
“r” – гілка вправо

Елементи словника:

[0] – номер ознаки для ділення
[1] – номер рядка для ділення
[2] – значення ознаки для ділення
[3]: none – не завершена гілка
int(i) – клас до якого належить об’єкт

{'': [3, 70, 1.75, None], 'l': [3, 69, 1.65, None], 'lr': [None, None, None, 2], 'll': [3, 58, 1.45, None], 'lll': [3, 32, 0.8, None], 'llll': [None, None, None, 0], 'lllr': [2, 24, 5.2, None], 'lllrl': [None, None, None, 1], 'lllrr': [None, None, None, 2], 'llr': [2, 7, 4.95, None], 'llrl': [None, None, None, 1], 'llrr': [3, 1, 1.55, None], 'llrrl': [None, None, None, 2], 'llrrr': [None, None, None, 1], 'r': [0, 3, 5.85, None], 'rl': [None, None, None, 2], 'rr': [1, 20, 3.15, None], 'rrl': [None, None, None, 2], 'rrr': [3, 0, 1.9, None], 'rrrl': [None, None, None, 1], 'rrrr': [None, None, None, 2]}

Accuracy in train = 1.0
Accuracy in test = 0.9333333333333333

Метод дерева прийняття рішень (decision tree)

Постановка задачі на **60** балів

1. Розділити всю вибірку на навчальну та тестову
2. Побудувати алгоритм навчання на навчальній вибірці з використанням вбудованого алгоритму DecisionTreeClassifier() з бібліотеки sklearn
3. Перевірити результати роботи на тестовій вибірці.
4. Вивести на екран дерево.
5. Оформити результати у вигляді звіту.