МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Звіт

до лабораторної

роботи з дисципліни

**"Машинне навчання та аналіз даних"**

для студентів освітньо-кваліфікаційного рівня «бакалавр»

за спеціальністю 122 «Комп’ютерні науки та інформаційні технології», освітня програма «Інформатика»

Виконав студент группи ІТІНФ-18-2

Туркін Микита

ХАРКІВ 2020

**Лабораторна робота №4. Дерева рішень. Дерева з множинними атрибутами**

**Мета:** Ознайомитись з побудовою дерев рішень та з дикретизацією атрибутів.

**Хід роботы**

**Варіант 2**

Вхідні данні:

|  |  |  |
| --- | --- | --- |
| Класс | А1 | А2 |
| + | 7 | 15 |
| – | 24 | 32 |
| + | 8 | 135 |
| – | 75 | 56 |
| – | 38 | 24 |
| + | 12 | 62 |

Потрібно провести дискретизацію атрибутів А1 і А2. Для цього побудуємо основну процедуру sample\_dataframe з допоміжным процедурами:

def sample\_dataframe(data):

    for ind in range(1, len(data.columns)):

        data = data.sort\_values(data.columns[ind], ascending=False)

        current\_series = data.iloc[:, ind].copy()

        prev\_series = data.iloc[:, ind - 1]

        levels = count\_levels(prev\_series, current\_series)

        pos\_neg = calc\_positive\_negative(prev\_series, '-', '+')

        entire\_entropy = calc\_entropy(\*pos\_neg)

        levels\_gain = []

        for level in levels:

            less\_level = data.loc[current\_series < level]

            more\_level = data.loc[current\_series > level]

            pos\_neg\_less\_level = calc\_positive\_negative(

                less\_level[prev\_series.name], '-', '+')

            pos\_neg\_more\_level = calc\_positive\_negative(

                more\_level[prev\_series.name], '-', '+')

            less\_level\_entropy = calc\_entropy(\*pos\_neg\_less\_level)

            more\_level\_entropy = calc\_entropy(\*pos\_neg\_more\_level)

            gain = entire\_entropy - (len(less\_level) / len(current\_series) \*

                                     less\_level\_entropy + len(more\_level) /

                                     len(current\_series) \* more\_level\_entropy)

            levels\_gain.append(gain)

        max\_gain\_level\_ind = np.argmax(levels\_gain)

        max\_gain\_level = levels[max\_gain\_level\_ind]

        data.iloc[:, ind].loc[current\_series <= max\_gain\_level] = '+'

        data.iloc[:, ind].loc[current\_series > max\_gain\_level] = '-'

    return data

Отримаємо такий результат:

Класс А1 А2

2 + + -

5 + + -

3 - - +

1 - - +

4 - - +

0 + + +

Також можна побачити етапи виконання:

Initial:

Класс А1 А2

0 + 7 15

1 - 24 32

2 + 8 135

3 - 75 56

4 - 38 24

5 + 12 62

Sorted:

Класс А1 А2

3 - 75 56

4 - 38 24

1 - 24 32

5 + 12 62

2 + 8 135

0 + 7 15

Levels: [18.0]

Entire entropy: 1

Level gains: [1.0]

Sampled:

Класс А1 А2

3 - - 56

4 - - 24

1 - - 32

5 + + 62

2 + + 135

0 + + 15

Initial:

Класс А1 А2

3 - - 56

4 - - 24

1 - - 32

5 + + 62

2 + + 135

0 + + 15

Sorted:

Класс А1 А2

2 + + 135

5 + + 62

3 - - 56

1 - - 32

4 - - 24

0 + + 15

Levels: [59.0, 19.5]

Entire entropy: 1

Level gains: [0.459147, 0.190874]

Sampled:

Класс А1 А2

2 + + -

5 + + -

3 - - +

1 - - +

4 - - +

0 + + +

**Висновок:** ознайомились з побудовою дерев рішень та з дикретизацією атрибутів.