10-11 классы

Большие данные

Презентация занятия

Решающие деревья

15 занятие

инжиниум®
МГТУ им. Н.Э. Баумана

2022

Тема: Решающие деревья

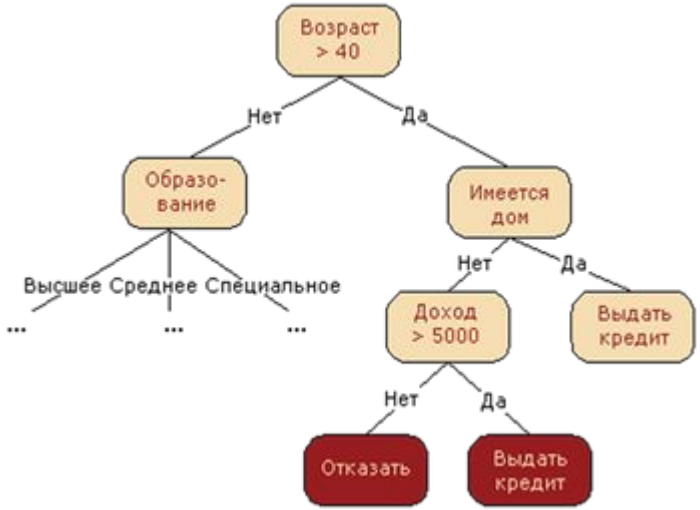
Мы приступаем к изучению машинного обучения. Начнем знакомство с решающими деревьями, разберемся что это за подход и почему он работает, как он обучается.

Познакомимся с библиотекой scikit-learn. На примере решающих деревьев затронем понятия – обучить, переобучить, недоучить модель. Разберем различные метрики

Тема: Решающие деревья

Что такое решающие деревья? Интуитивно простой и понятный метод машинного обучения

Предположим, что перед нами стоит задача – выдавать человеку кредит или нет? Рассмотрим рисунок.



Тема: Решающие деревья

Давайте разберем искусственный пример. Какое дерево решений можно построить? Интересует зависимость переменной Y, от переменных X1 и X2

Нарисуйте это дерево

```
In [70]: data
```

Out[70]:

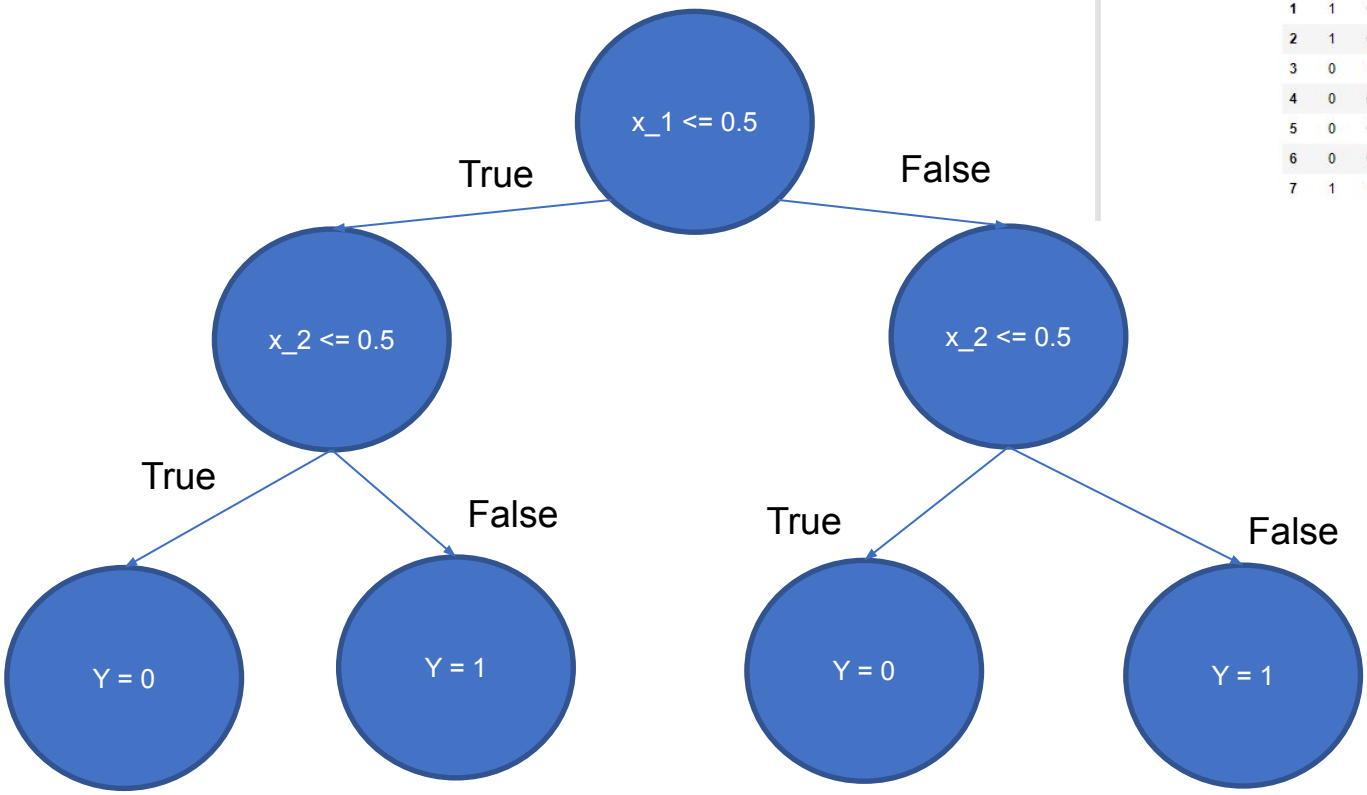
	X_1	X_2	Y
0	1	0	1
1	1	0	1
2	1	0	1
3	0	1	1
4	0	0	0
5	0	0	0
6	0	0	0
7	1	1	0

Тема: Решающие деревья

Примерно такое решающее дерево у вас должно получиться

Out[70]:

	X_1	X_2	Y
0	1	0	1
1	1	0	1
2	1	0	1
3	0	1	1
4	0	0	0
5	0	0	0
6	0	0	0
7	1	1	0



Тема: Решающие деревья

Теперь давайте смоделируем это дерево

Для начала – подключим необходимые модули. Установка sklearn -> pip install scikit-learn и создадим датайрейм (запишите его в переменную data)

```
from sklearn import tree
import pandas as pd
import matplotlib.pyplot as plt
```

```
In [70]: data
```

Out[70]:

	X_1	X_2	Y
0	1	0	1
1	1	0	1
2	1	0	1
3	0	1	1
4	0	0	0
5	0	0	0
6	0	0	0
7	1	1	0

Тема: Решающие деревья

Теперь давайте создадим наше дерево. Выберем основным критерием энтропию. О том, что такое энтропия мы поговорим позднее

```
]: clf = tree.DecisionTreeClassifier(criterion='entropy')
```

Сохраним все, что может влиять на итог в переменную X (PandasDataFrame), а результат у переменную y (PandasSeries)

```
[74]: X
```

Out[74]:

	X_1	X_2
0	1	0
1	1	0
2	1	0
3	0	1
4	0	0
5	0	0
6	0	0
7	1	1

```
In [89]: y
```

```
Out[89]: 0    1
         1    1
         2    1
         3    1
         4    0
         5    0
         6    0
         7    0
         Name: Y, dtype: int64
```

Тема: Решающие деревья

Обучим наше первое дерево. Используем метод fit()

```
In [90]: clf.fit(X, Y)
```

Out[90]:

DecisionTreeClassifier

DecisionTreeClassifier(criterion='entropy')

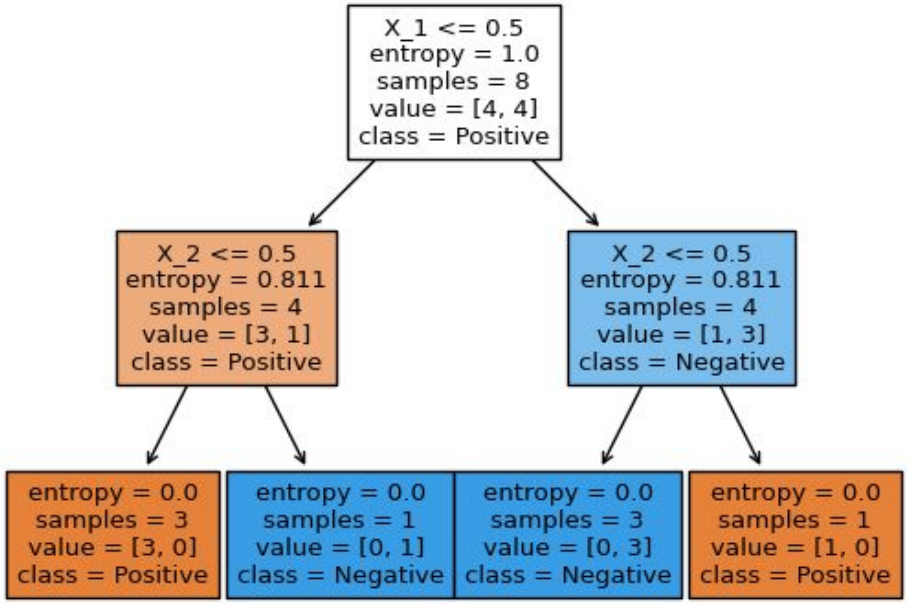
И Визуализируем наше дерево используя метод plot_tree

```
In [91]: tree.plot_tree(
             clf.fit(X, Y),
             filled=True,
             feature_names=list(X),
             class_names=['Positive', 'Negative']
         )
```

Тема: Решающие деревья

Что видим в результате?

```
In [91]:
tree.plot_tree(
    clf.fit(X, Y),
    filled=True,
    feature_names=list(X),
    class_names=['Positive', 'Negative']
)
```



Тема: Решающие деревья

Задание на программирование.

Используя датасет задачи про титаник постройте решающее дерево, которое позволит определить.

1) Выживет ли пассажир, в зависимости от его класса
2) Выживет ли пассажир в зависимости от его возраста
3) Выживет ли пассажир в зависимости от стоимости его билета