**Задание**

Построить дерево принятия решений с использованием критерия Джини

| **X1** | **X2** | **X3** | | | **Y** | |
| --- | --- | --- | --- | --- | --- | --- |
| 0,3 | 8 | -7,1 | | | 0 | |
| 0 | 99 | 78,4 | | | 0 | |
| 65 | 4,1 | 14 | | | 1 | |
| 12,6 | -40 | | 8 | 0 | |
| 12,5 | 11,3 | 1 | | | 1 | |
| 2,2 | 18,1 | 12 | | | 0 | |
| 0 | 15 | 47,1 | | | 0 | |
| 1000 | -1000 | 1 | | | 0 | |
| 45 | 46 | -7,1 | | | 1 | |

Определить результат классификации для следующих данных:

| **X1** | **X2** | **X3** |
| --- | --- | --- |
| -70 | -80 | 1,12 |
| -4 | 100 | -7,4 |
| 80 | -44,7 | 1 |
| 64,2 | 102,4 | 14 |

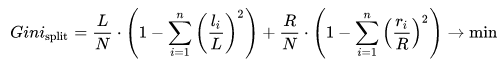
**Решение**

Следуя жадному алгоритму, мы должны выбирать локально оптимальные решения. То есть необходимо вычислить критерий Джини каждого разбиения. Он определяется:

 для разбиения на две части.



Тогда:



Эту формулу можно преобразовать к виду:

N — число примеров в узле — предке, L, R — число примеров соответственно в левом и правом потомке, li и ri — число экземпляров i-го класса в левом/правом потомке, n – количество классов.

В нашей задаче N = 9, n = 2.

Критерии разбиений:

X1>=0.3 Gini = 9 - ((2^2 + 0^2)/2 + (4^2 + 3^2)/7) = 3.4

X1>=0 Gini = 9 - ((0^2 + 0^2)/0 + (6^2 + 3^2)/9) = 4.0

X1>=65 Gini = 9 - ((5^2 + 2^2)/7 + (1^2 + 1^2)/2) = 3.9

X1>=12.6 Gini = 9 - ((4^2 + 1^2)/5 + (2^2 + 2^2)/4) = 3.6

**X1>=12.5 Gini = 9 - ((4^2 + 0^2)/4 + (2^2 + 3^2)/5) = 2.4**

X1>=2.2 Gini = 9 - ((3^2 + 0^2)/3 + (3^2 + 3^2)/6) = 3.0

X1>=0 Gini = 9 - ((0^2 + 0^2)/0 + (6^2 + 3^2)/9) = 4.0

X1>=1000 Gini = 9 - ((5^2 + 3^2)/8 + (1^2 + 0^2)/1) = 3.8

X1>=45 Gini = 9 - ((5^2 + 1^2)/6 + (1^2 + 2^2)/3) = 3.0

X2>=8 Gini = 9 - ((2^2 + 1^2)/3 + (4^2 + 2^2)/6) = 4.0

X2>=99 Gini = 9 - ((5^2 + 3^2)/8 + (1^2 + 0^2)/1) = 3.8

X2>=4.1 Gini = 9 - ((2^2 + 0^2)/2 + (4^2 + 3^2)/7) = 3.4

X2>=-40 Gini = 9 - ((1^2 + 0^2)/1 + (5^2 + 3^2)/8) = 3.8

X2>=11.3 Gini = 9 - ((3^2 + 1^2)/4 + (3^2 + 2^2)/5) = 3.9

X2>=18.1 Gini = 9 - ((4^2 + 2^2)/6 + (2^2 + 1^2)/3) = 4.0

X2>=15 Gini = 9 - ((3^2 + 2^2)/5 + (3^2 + 1^2)/4) = 3.9

X2>=-1000 Gini = 9 - ((0^2 + 0^2)/0 + (6^2 + 3^2)/9) = 4.0

X2>=46 Gini = 9 - ((5^2 + 2^2)/7 + (1^2 + 1^2)/2) = 3.9

X3>=-7.1 Gini = 9 - ((0^2 + 0^2)/0 + (6^2 + 3^2)/9) = 4.0

X3>=78.4 Gini = 9 - ((5^2 + 3^2)/8 + (1^2 + 0^2)/1) = 3.8

X3>=14 Gini = 9 - ((4^2 + 2^2)/6 + (2^2 + 1^2)/3) = 4.0

X3>=8 Gini = 9 - ((2^2 + 2^2)/4 + (4^2 + 1^2)/5) = 3.6

X3>=1 Gini = 9 - ((1^2 + 1^2)/2 + (5^2 + 2^2)/7) = 3.9

X3>=12 Gini = 9 - ((3^2 + 2^2)/5 + (3^2 + 1^2)/4) = 3.9

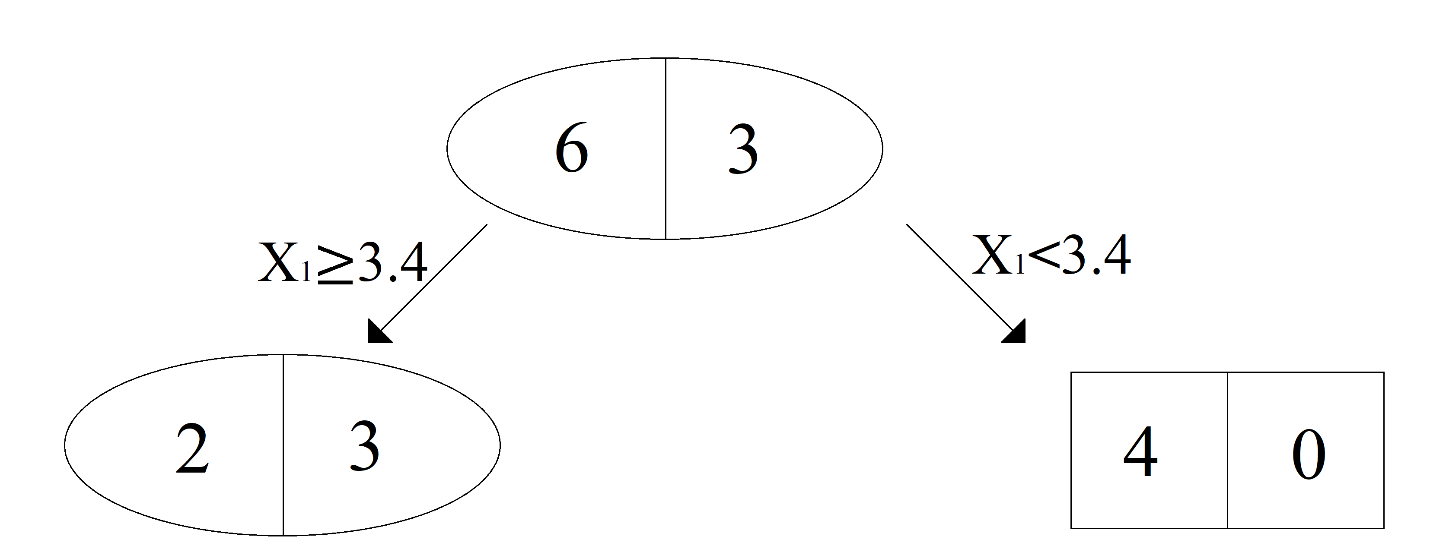
X3>=47.1 Gini = 9 - ((4^2 + 3^2)/7 + (2^2 + 0^2)/2) = 3.4

X3>=1 Gini = 9 - ((1^2 + 1^2)/2 + (5^2 + 2^2)/7) = 3.9

X3>=-7.1 Gini = 9 - ((0^2 + 0^2)/0 + (6^2 + 3^2)/9) = 4.0

Минимальное значения критерия Джини = 2.40, соответствующее разбиению X1>=12.5

При ветвлении с одной стороны был получен набор только из элементов первого класса, с другой – двух первого и трёх второго.



**X1<12.5**

**X1>=12.5**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** |  | **X1** | **X2** | **X3** | **Y** |
| 65 | 4,1 | 14 | 1 |  | 0,3 | 8 | -7,1 | 0 |
| 12,6 | -40 | 8 | 0 |  | 0 | 99 | 78,4 | 0 |
| 12,5 | 11,3 | 1 | 1 |  | 2,2 | 18,1 | 12 | 0 |
| 1000 | -1000 | 1 | 0 |  | 0 | 15 | 47,1 | 0 |
| 45 | 46 | -7,1 | 1 |  |  |  |  |  |

X1>=65 Gini = 5 - ((1^2 + 2^2)/3 + (1^2 + 1^2)/2) = 6.3

X1>=12.6 Gini = 5 - ((0^2 + 1^2)/1 + (2^2 + 2^2)/4) = 6.0

X1>=12.5 Gini = 5 - ((0^2 + 0^2)/0 + (2^2 + 3^2)/5) = 6.4

X1>=1000 Gini = 5 - ((1^2 + 3^2)/4 + (1^2 + 0^2)/1) = 5.5

X1>=45 Gini = 5 - ((1^2 + 1^2)/2 + (1^2 + 2^2)/3) = 6.3

**X2>=4.1 Gini = 5 - ((2^2 + 0^2)/2 + (0^2 + 3^2)/3) = 4.0**

X2>=-40 Gini = 5 - ((1^2 + 0^2)/1 + (1^2 + 3^2)/4) = 5.5

X2>=11.3 Gini = 5 - ((2^2 + 1^2)/3 + (0^2 + 2^2)/2) = 5.3

X2>=-1000 Gini = 5 - ((0^2 + 0^2)/0 + (2^2 + 3^2)/5) = 6.4

X2>=46 Gini = 5 - ((2^2 + 2^2)/4 + (0^2 + 1^2)/1) = 6.0

X3>=14 Gini = 5 - ((2^2 + 2^2)/4 + (0^2 + 1^2)/1) = 6.0

X3>=8 Gini = 5 - ((1^2 + 2^2)/3 + (1^2 + 1^2)/2) = 6.3

X3>=1 Gini = 5 - ((0^2 + 1^2)/1 + (2^2 + 2^2)/4) = 6.0

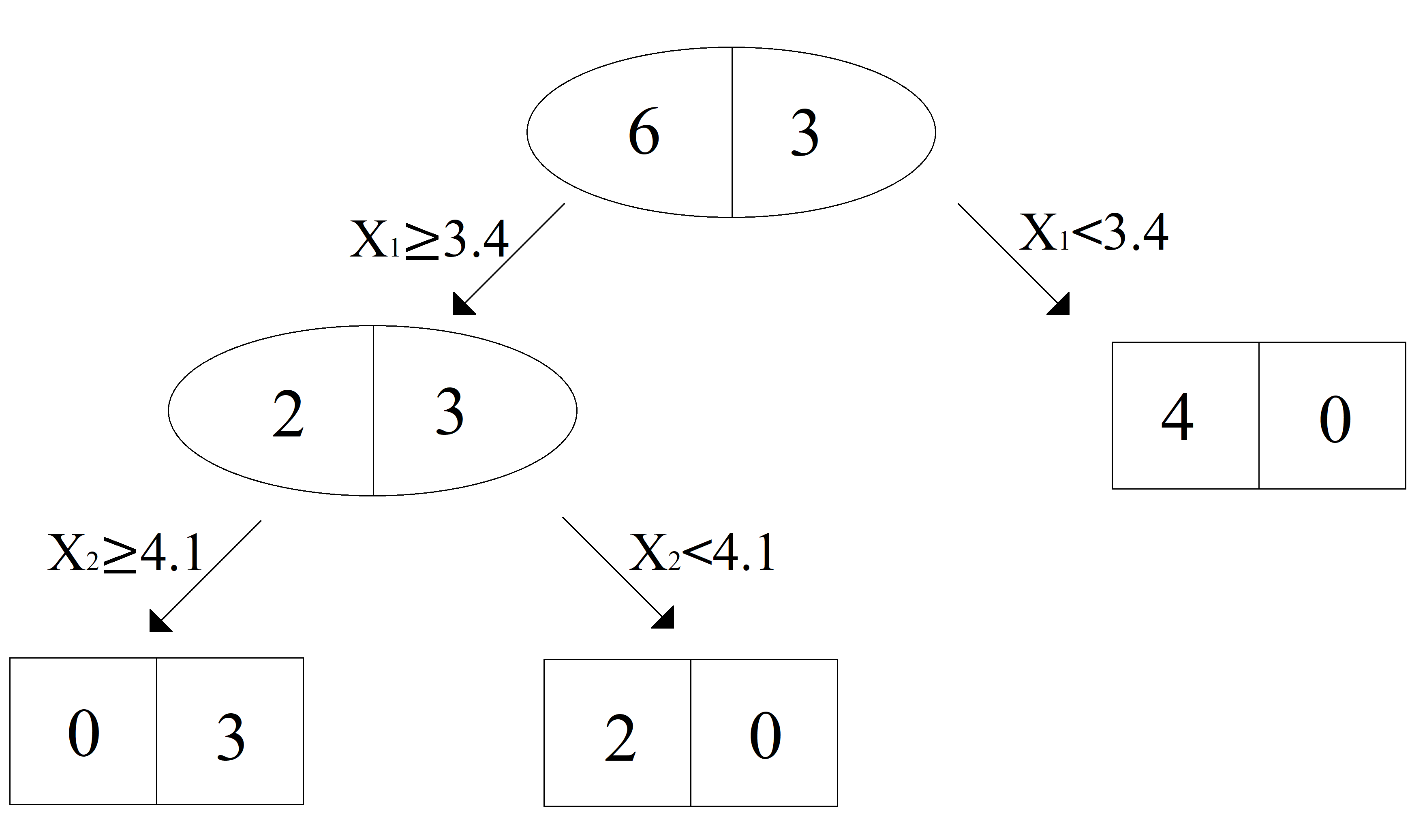
X3>=1 Gini = 5 - ((0^2 + 1^2)/1 + (2^2 + 2^2)/4) = 6.0

X3>=-7.1 Gini = 5 - ((0^2 + 0^2)/0 + (2^2 + 3^2)/5) = 6.4

Локально оптимальное решение это x2 ≥ 4.1

В этом случае мы получим ещё две ветви, соответствующие набору из двух

элементов первого класс и набору из трёх элементов второго класса. Так как в наборах нет элементов разных классов, то задача построения дерева решений выполнена.



**X1<12.5**

**X1>=12.5**

Классификация выполняется в соответствии с выделенными правилами:

Y = 1, если x1 ≥ 12.5 и x2 ≥ 4.1, иначе Y = 0

| **X1** | **X2** | **X3** | **Y** |
| --- | --- | --- | --- |
| -70 | -80 | 1,12 | 0 |
| -4 | 100 | -7,4 | 0 |
| 80 | -44,7 | 1 | 0 |
| 64,2 | 102,4 | 14 | 1 |