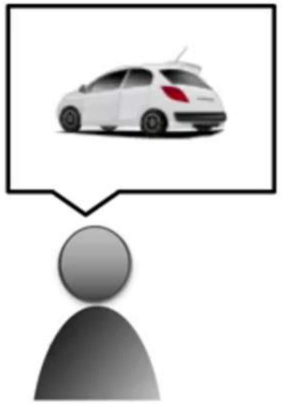
Decision Tree Example

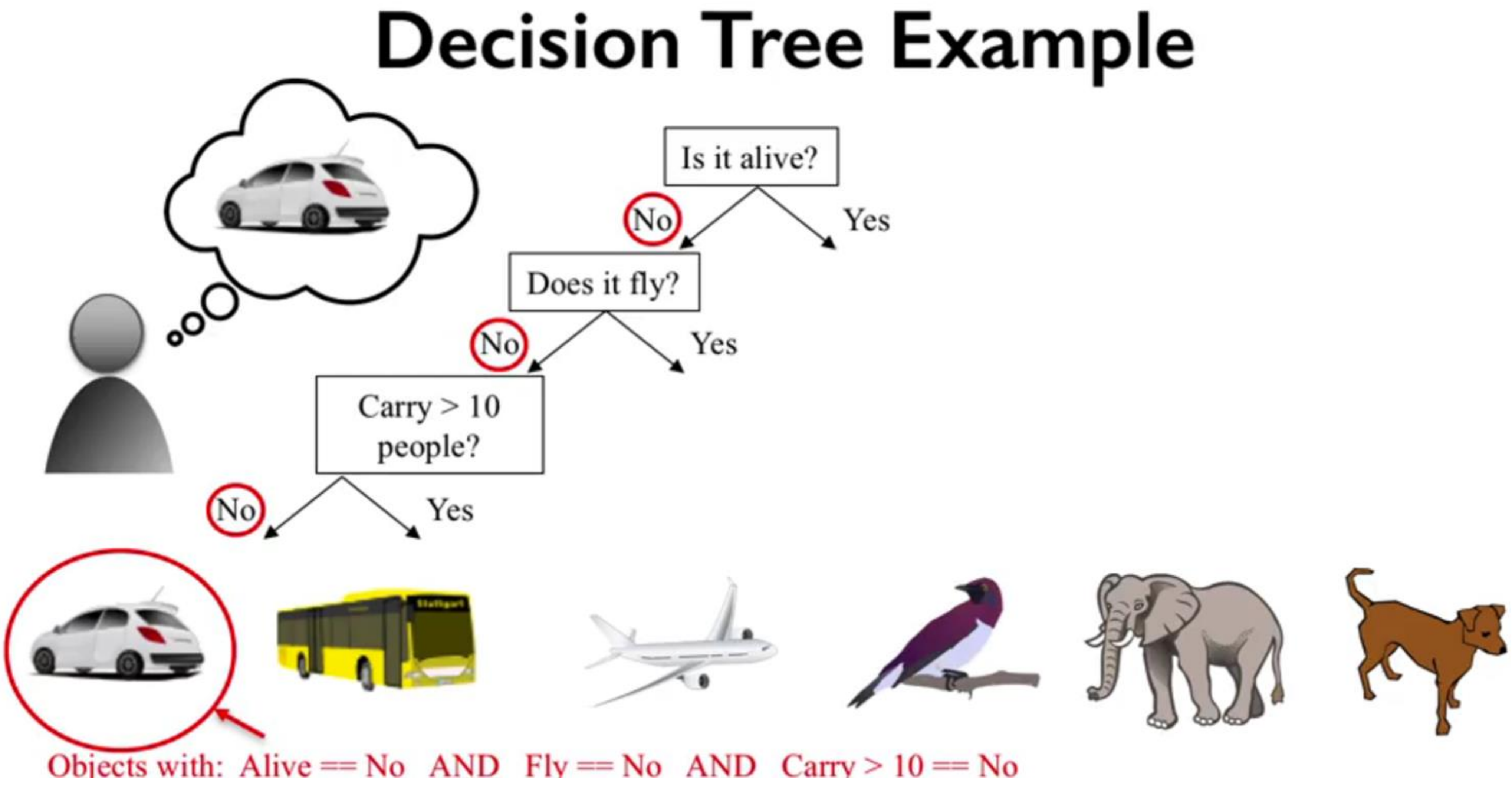
Assume a person only knows the following 6 objects:

Car Bus Airplane Bird Elephant Dog

One day, a person saw a picture of an object, then this guy tried to figure out the name of the object using a decision tree.



The person asked many yes-no questions



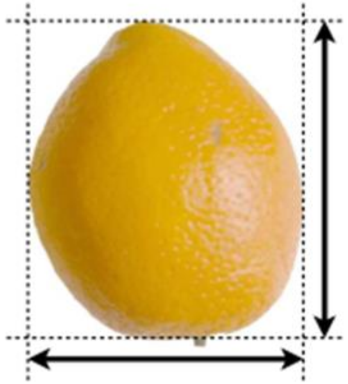
A data sample has three binary features: [alive, fly, carry_more_than_10]

Decision Tree Example

Feature Vector x

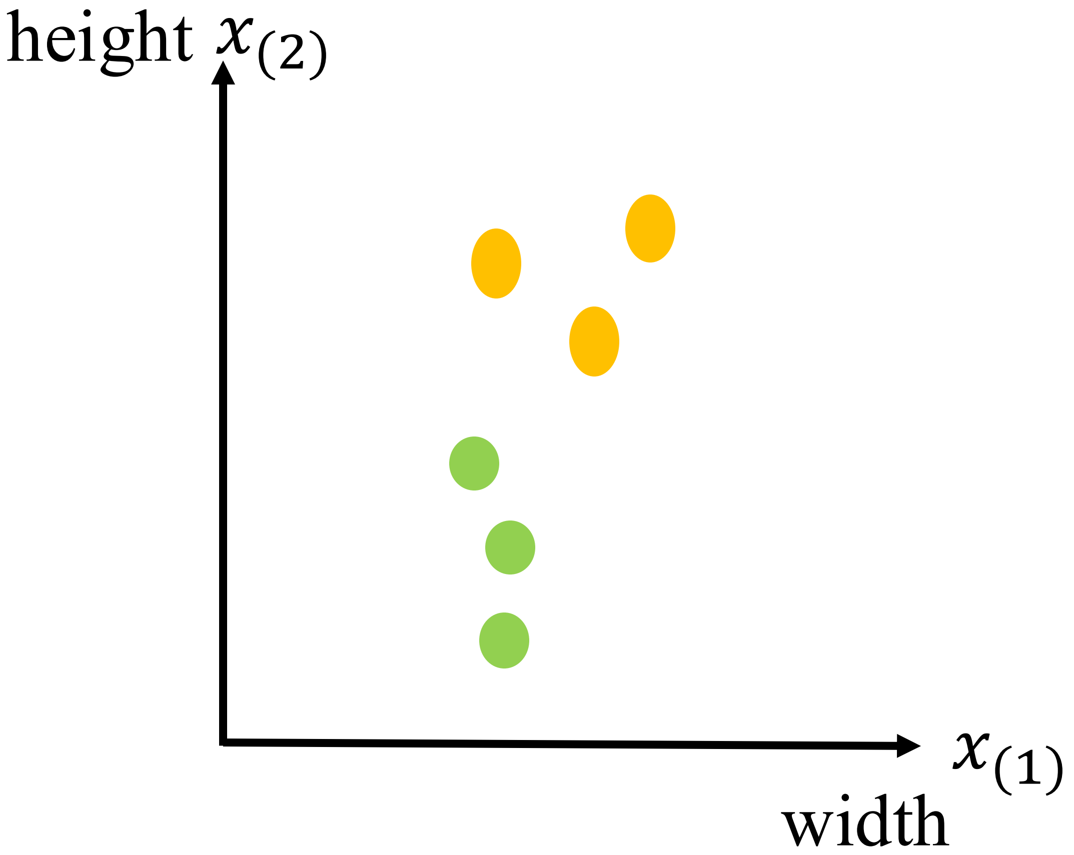
$$x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix} \quad \begin{matrix} \text{width} \\ \text{height} \end{matrix}$$

$y \quad = 1 \quad$ It is an apple

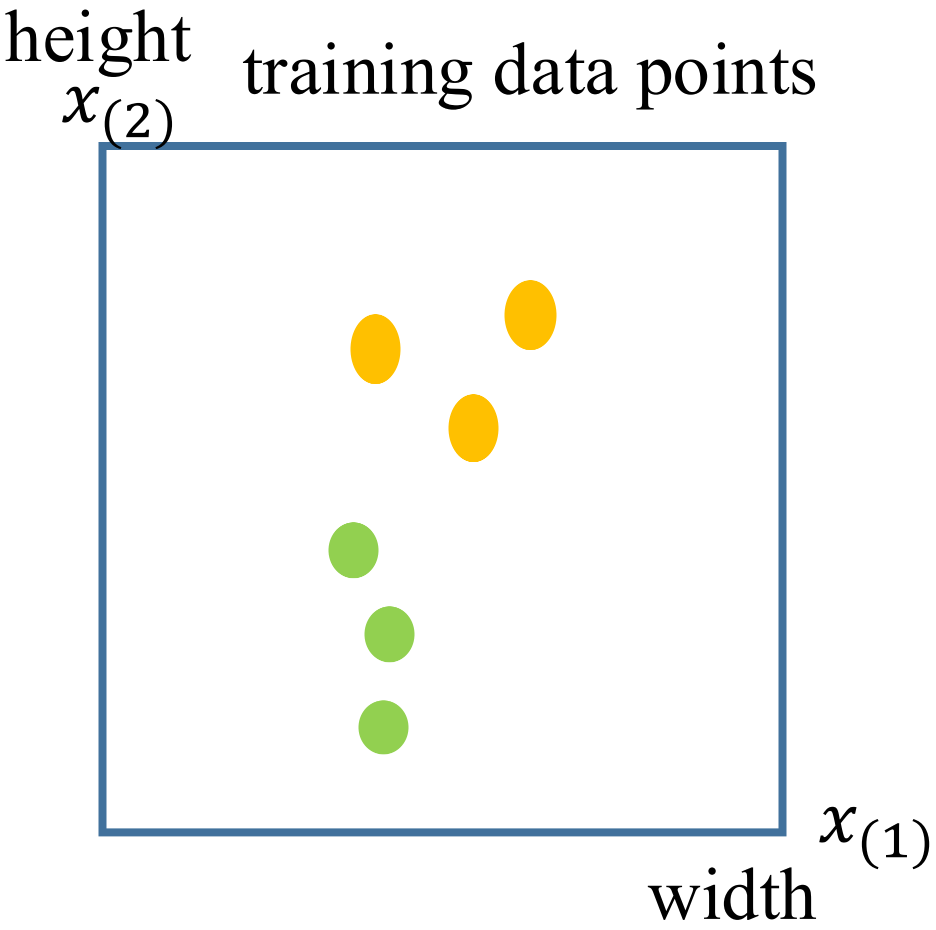


$y \quad = 0 \quad$ It is a lemon

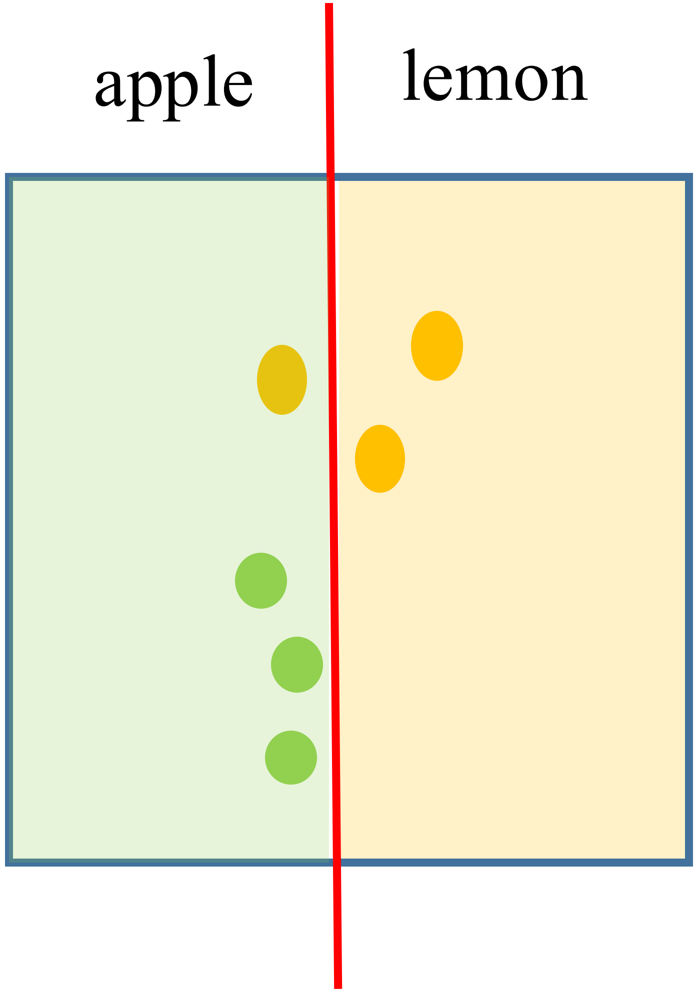
Feature Space (Input Space)



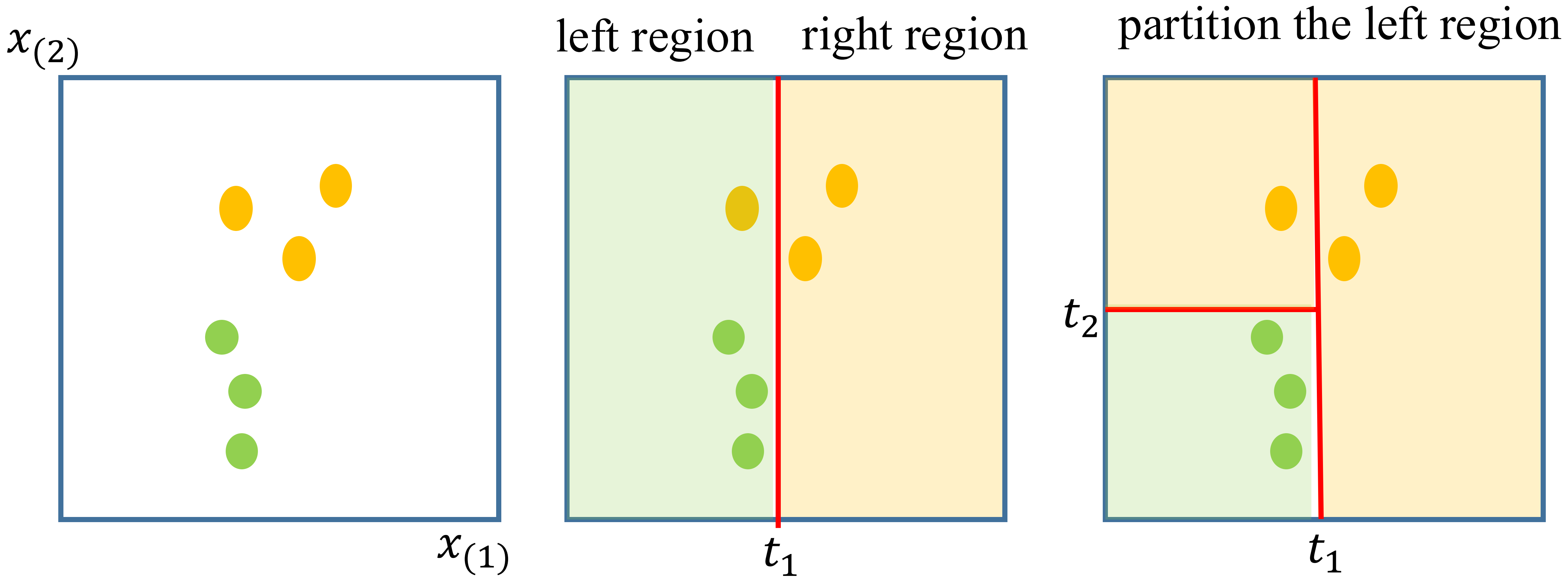
Decision Tree Example



split the feature space



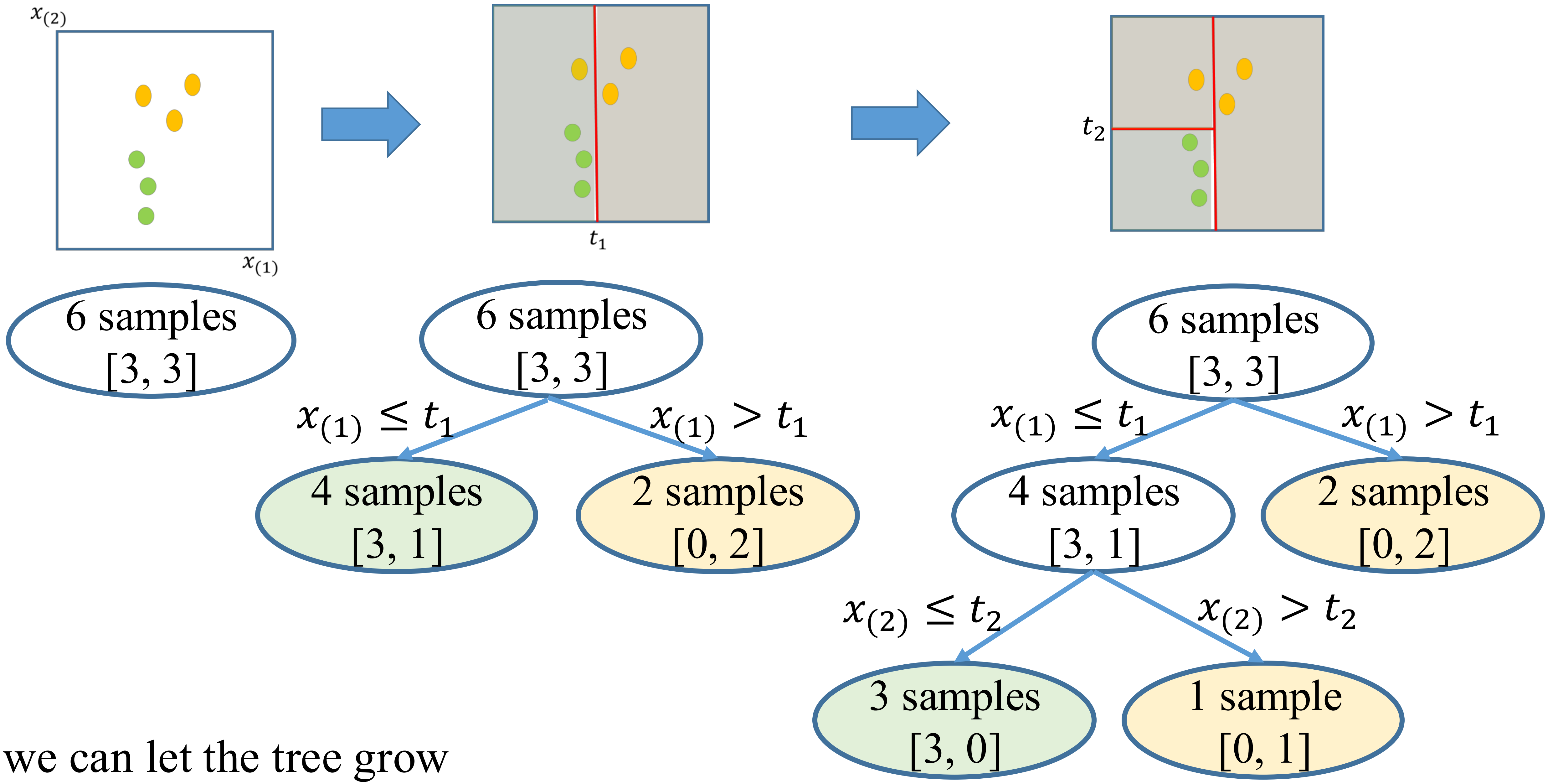
$x_{(1)} \leq t_1$: apple (majority voting)
because there are 3 apples and 1 lemon



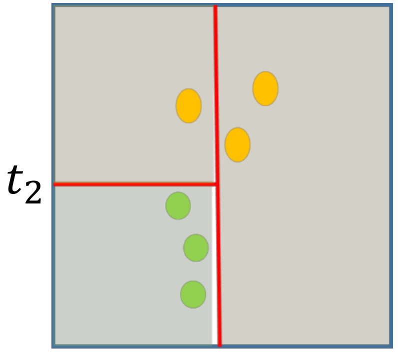
a sample has two features $x_{(1)}$ and $x_{(2)}$

$x_{(1)} \le t_1$: apple
$x_{(1)} > t_1$: lemon

Given $x_{(1)} \le t_1$,
$x_{(2)} \le t_2$: apple
$x_{(2)} > t_2$: lemon

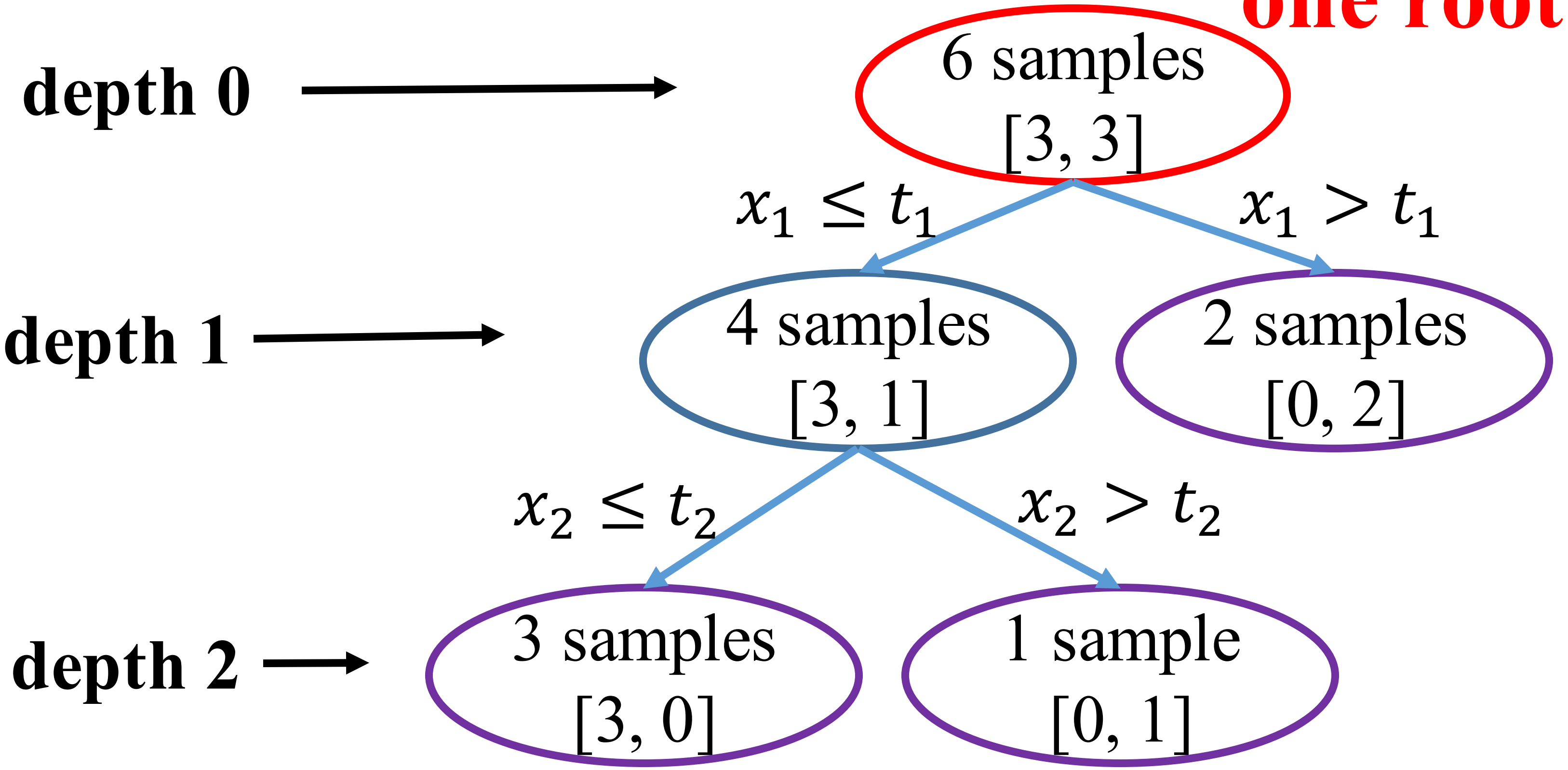


we can let the tree grow another layer



We can further expand this tree such that every leaf node only contains one data sample

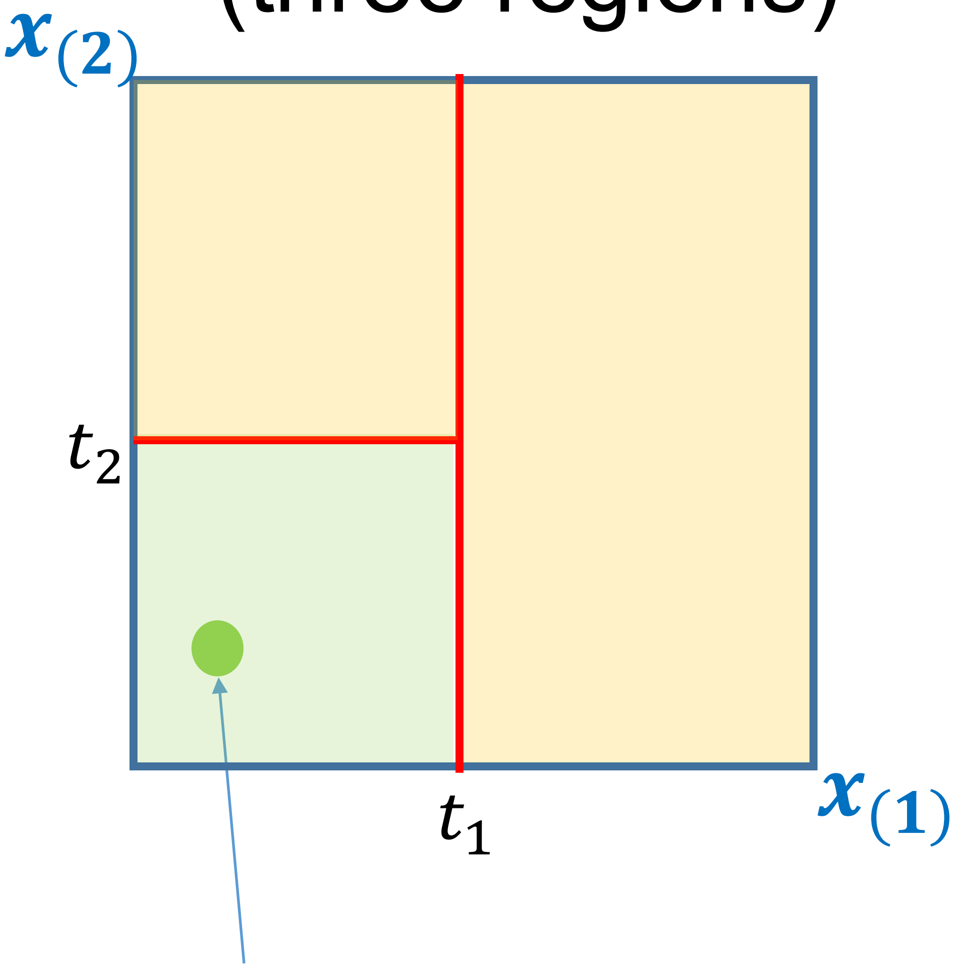
one root node



3 leaf (terminal) nodes do not have child nodes

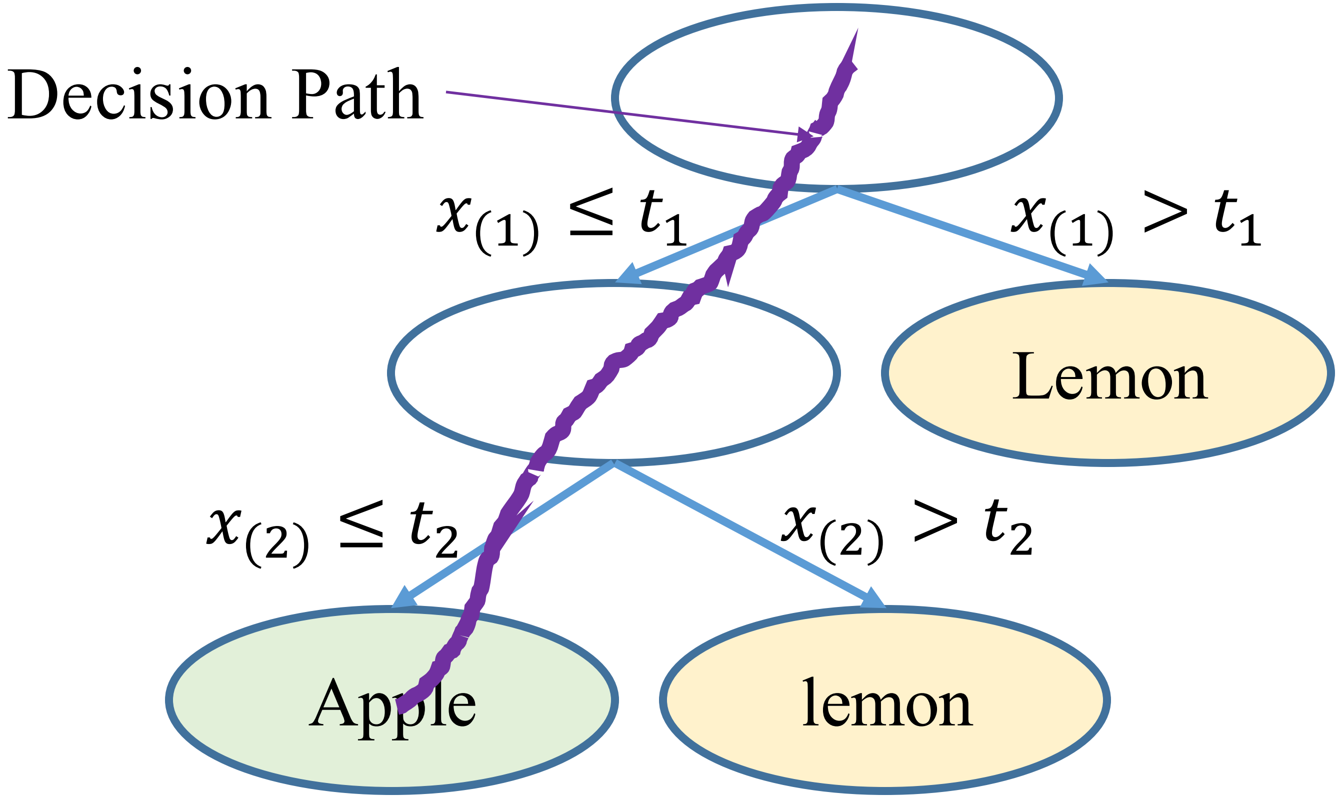
The depth of the tree is 2

partition of input space (three regions)



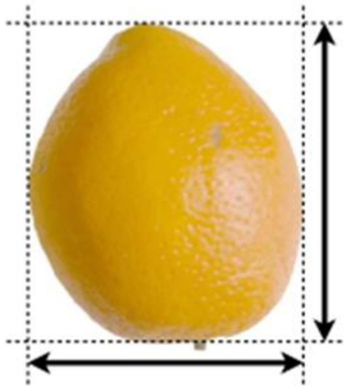
a new data sample x

use the decision tree T2



The data sample x falls into a region/node

Apply Decision Tree to the fruit dataset



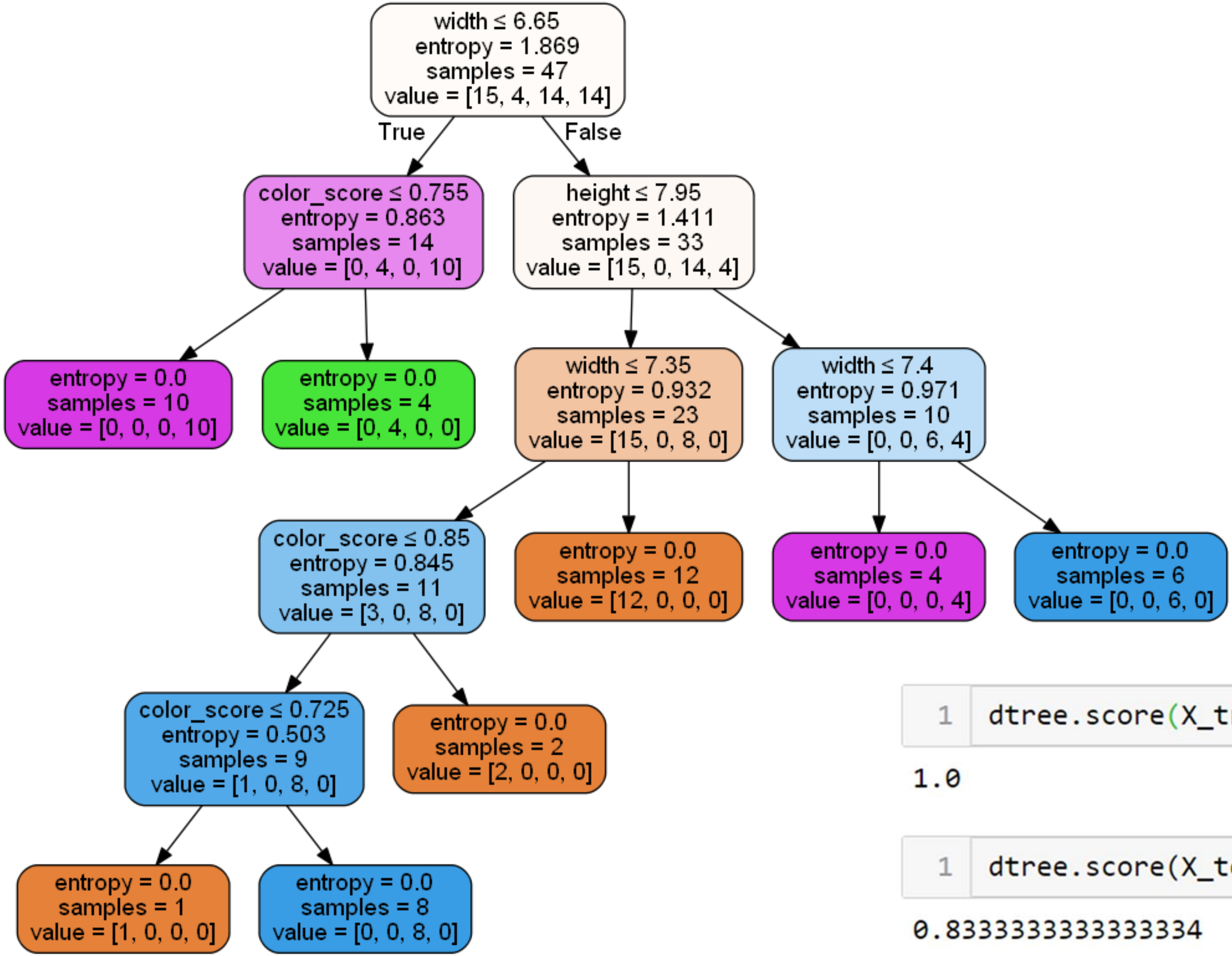
The fruits dataset (a large table)
Each row contains the information of a fruit sample/instance

fruit label	fruit_name	subtype	mass (g)	width (cm)	height (cm)	color_score
1	apple	granny_smith	192	8.4	7.3	0.55
4	lemon	spanish_belsan	194	7.2	10.3	0.70

http://usapple.org/the-industry/apple-varieties/

The feature vector of a fruit sample: [width, height, color_score]

1:apple,
2:mandarin
3:orange
4:lemon




```
dtree.score(X_train, Y_train)
```

```
1.0
```

```
dtree.score(X_test, Y_test)
```

```
0.8333333333333334
```

Use a trained decision tree for classification

- Assume that a tree has been trained (fitted) on the training set.
- A new sample x needs to be classified
- The sample x will eventually fall into one of the leaf nodes.
- The majority class of the training samples in the leaf node is C.
- The predicted class label of the new sample x is C