WPI – CS4445 Data Mining and Knowledge Discovery in Databases A08

Homework 1: Data Pre-processing, Mining, and Evaluation of Decision Trees

Amro Khasawneh

Part I

1. [60pts] **Construction of ID3 decision tree**

hair	eggs	toothed	legs	type
no	no	yes	0	type1
no	yes	yes	4	type5
no	yes	no	6	type6
no	yes	yes	0	type4
no	yes	no	6	type6
yes	no	yes	4	type1
yes	no	yes	4	type1
no	yes	no	2	type2
no	yes	yes	0	type4
no	yes	no	6	type6
no	yes	yes	4	type5
no	yes	yes	0	type4
yes	no	yes	4	type1
yes	yes	no	6	type6
no	yes	no	2	type2

To select the first attribute with which to split, we calculate the *information gain* for each attribute and choose the one that gains the most information to split on (equivalent to selecting the attribute with the lowest *entropy*).

?

Before splitting, the entropy over all the instances is:

$$\left(-\frac{4}{15}\times\log_2\left(\frac{4}{15}\right)\right)+\left(-\frac{2}{15}\times\log_2\left(\frac{2}{15}\right)\right)+\left(-\frac{0}{15}\right)+\left(-\frac{3}{15}\times\log_2\left(\frac{3}{15}\right)\right)+\left(-\frac{2}{15}\times\log_2\left(\frac{2}{15}\right)\right)+\left(-\frac{4}{15}\times\log_2\left(\frac{4}{15}\right)\right)+\left(-\frac{0}{15}\right)=2.257$$

Now we calculate the entropy for the attribute: hair, eggs, toothed, legs.

- hair:

hair	type						
	type1	type2	type3	type4	type5	type6	type7
no (11)	1	2	0	3	2	3	0
yes (4)	3	0	0	0	0	1	0

$$\frac{11}{15}\times\left[\left(-\frac{1}{11}\times\log_2\left(\frac{1}{11}\right)\right)+\left(-\frac{2}{11}\times\log_2\left(\frac{2}{11}\right)\right)+\left(-\frac{0}{11}\right)+\left(-\frac{3}{11}\times\log_2\left(\frac{3}{11}\right)\right)+\left(-\frac{2}{11}\times\log_2\left(\frac{2}{11}\right)\right)+\left(-\frac{3}{11}\times\log_2\left(\frac{3}{11}\right)\right)+\left(-\frac{0}{11}\right)\right]+\frac{4}{15}$$
$$\times\left[\left(-\frac{3}{4}\times\log_2\left(\frac{3}{4}\right)\right)+4\times\left(-\frac{0}{4}\right)+\left(-\frac{1}{4}\times\log_2\left(\frac{1}{4}\right)\right)+\left(-\frac{0}{4}\right)\right]=0.558$$

- eggs:

eggs	type						
	type1	type2	type3	type4	type5	type6	type7
no (4)	4	0	0	0	0	0	0
yes (11)	0	2	0	3	2	4	0

$$\frac{4}{15}\times\left[\left(-\frac{4}{4}\times\log_2\left(\frac{4}{4}\right)\right)+6\times\left(-\frac{0}{4}\right)\right]+\frac{11}{15}\times\left[\left(-\frac{0}{11}\right)+\left(-\frac{2}{11}\times\log_2\left(\frac{2}{11}\right)\right)+\left(-\frac{0}{11}\right)+\left(-\frac{3}{11}\times\log_2\left(\frac{3}{11}\right)\right)+\left(-\frac{2}{11}\times\log_2\left(\frac{2}{11}\right)\right)+\left(-\frac{4}{11}\times\log_2\left(\frac{4}{11}\right)\right)+\left(-\frac{0}{11}\right)\right]=1.42$$

- toothed:

toothed	type						
	type1	type2	type3	type4	type5	type6	type7
no (6)	0	2	0	0	0	4	0
yes (9)	4	0	0	3	2	0	0

$$\frac{6}{15}\times\left[\left(-\frac{0}{6}\right)+\left(-\frac{2}{6}\times\log_2\left(\frac{2}{6}\right)\right)+3\times\left(-\frac{0}{6}\right)+\left(-\frac{4}{6}\times\log_2\left(\frac{4}{6}\right)\right)+\left(-\frac{0}{6}\right)\right]+\frac{9}{15}\times\left[\left(-\frac{4}{9}\times\log_2\left(\frac{4}{9}\right)\right)+2\times\left(-\frac{0}{9}\right)+\left(-\frac{3}{9}\times\log_2\left(\frac{3}{9}\right)\right)+\left(-\frac{2}{9}\times\log_2\left(\frac{2}{9}\right)\right)+2\times\left(-\frac{0}{9}\right)\right]=1.286$$

- legs:

legs	type						
	type1	type2	type3	type4	type5	type6	type7
0 (4)	1	0	0	3	0	0	0
2 (2)	0	2	0	0	0	0	0
4 (5)	3	0	0	0	2	0	0
5 (0)	0	0	0	0	0	0	0
6 (4)	0	0	0	0	0	4	0
8 (0)	0	0	0	0	0	0	0

$$\frac{4}{15}\times\left[\left(-\frac{1}{4}\times\log_2\left(\frac{1}{4}\right)\right)+2\times\left(-\frac{0}{4}\right)+\left(-\frac{3}{4}\times\log_2\left(\frac{3}{4}\right)\right)+3\times\left(-\frac{0}{4}\right)\right]+\frac{2}{15}\times\left[\left(-\frac{0}{2}\right)+\left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right)+5\times\left(-\frac{0}{2}\right)\right]+\frac{5}{15}\times\left[\left(-\frac{3}{5}\times\log_2\left(\frac{3}{5}\right)\right)+3\times\left(-\frac{0}{5}\right)+\left(-\frac{2}{5}\times\log_2\left(\frac{2}{5}\right)\right)+2\times\left(-\frac{0}{5}\right)\right]+\frac{0}{15}+\frac{4}{15}\times\left[5\times\left(-\frac{0}{4}\right)+\left(-\frac{4}{4}\times\log_2\left(\frac{4}{4}\right)\right)+\left(-\frac{0}{4}\right)\right]+\frac{0}{15}=0.54$$

Therefore at this stage, the information gains for the above attributes are:

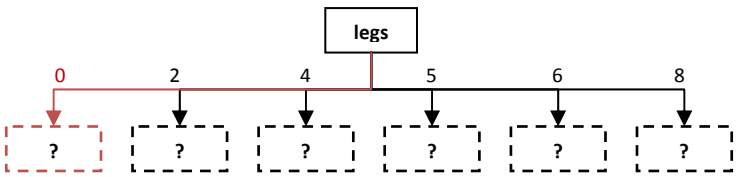
$$gain(hair)=2.257-0.558=1.699$$

$$gain(eggs)=2.257-1.42=0.837$$

$$gain(toothed)=2.257-1.286=0.971$$

$$gain(legs)=2.257-0.54=1.717$$

So we select *legs* as the splitting attribute at the root of the tree, and we continue recursively on the children nodes.



First we consider *legs=0*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	no	yes	0	type1
no	yes	yes	0	type4
no	yes	yes	0	type4
no	yes	yes	0	type4

The entropy over the above instances is:

$$\left(-\frac{1}{4}\times \log_2\left(\frac{1}{4}\right)\right) + 2\times\left(-\frac{0}{4}\right) + \left(-\frac{3}{4}\times \log_2\left(\frac{3}{4}\right)\right) + 3\times\left(-\frac{0}{4}\right) = 0.811$$

Now we calculate the entropy for the attribute: hair, eggs, and toothed over these instances.

- hair:

hair	type						
	type1	type2	type3	type4	type5	type6	type7
no (4)	1	0	0	3	0	0	0
yes (0)	0	0	0	0	0	0	0

$$\frac{4}{4}\times\left[\left(-\frac{1}{4}\times \log_2\left(\frac{1}{4}\right)\right) + 2\times\left(-\frac{0}{4}\right) + \left(-\frac{3}{4}\times \log_2\left(\frac{3}{4}\right)\right) + 3\times\left(-\frac{0}{4}\right)\right] + \frac{0}{4} = 0.811$$

- eggs:

eggs	type						
	type1	type2	type3	type4	type5	type6	type7
no (1)	1	0	0	0	0	0	0
yes (3)	0	0	0	3	0	0	0

$$\frac{1}{4}\times\left[\left(-\frac{1}{1}\times \log_2\left(\frac{1}{1}\right)\right) + 6\times\left(-\frac{0}{1}\right)\right] + \frac{3}{4}\times\left[3\times\left(-\frac{0}{3}\right) + \left(-\frac{3}{3}\times \log_2\left(\frac{3}{3}\right)\right) + 3\times\left(-\frac{0}{3}\right)\right] = 0.0$$

- toothed:

toothed	type						
	type1	type2	type3	type4	type5	type6	type7
no (0)	0	0	0	0	0	0	0
yes (4)	1	0	0	3	0	0	0

$$\frac{0}{4} + \frac{4}{4}\times\left[\left(-\frac{1}{4}\times \log_2\left(\frac{1}{4}\right)\right) + 2\times\left(-\frac{0}{4}\right) + \left(-\frac{3}{4}\times \log_2\left(\frac{3}{4}\right)\right) + 3\times\left(-\frac{0}{4}\right)\right] = 0.811$$

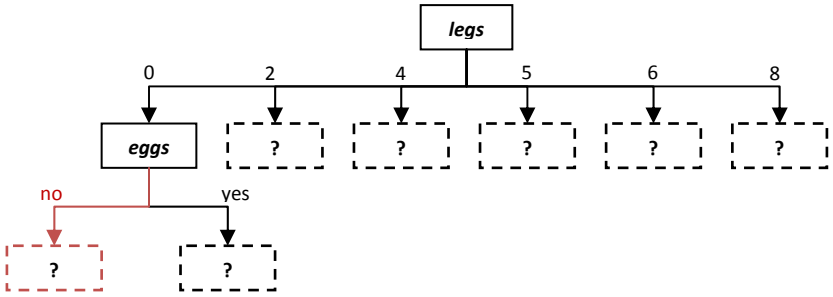
Therefore at this stage, the information gains for the above attributes are:

$$gain(hair) = 0.811 - 0.811 = 0.0$$

$$gain(eggs) = 0.811 - 0.0 = 0.811$$

$$gain(toothed) = 0.811 - 0.811 = 0.0$$

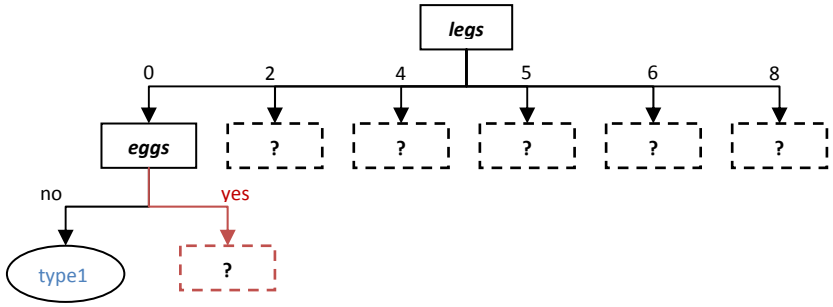
So we select *eggs* as the splitting attribute, and we continue recursively on the children nodes.



Now we consider *eggs=no*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	no	yes	0	type1

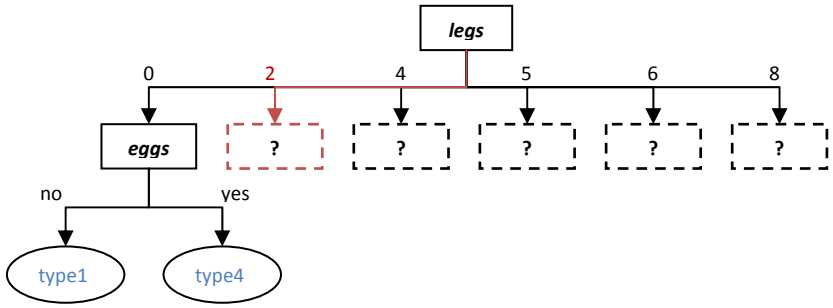
We reached a homogeneous (pure) node with no entropy (that is it contains instances that all have the same classification); therefore we label this node with *type1*.



Next we consider *eggs=yes*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	yes	yes	0	type4
no	yes	yes	0	type4
no	yes	yes	0	type4

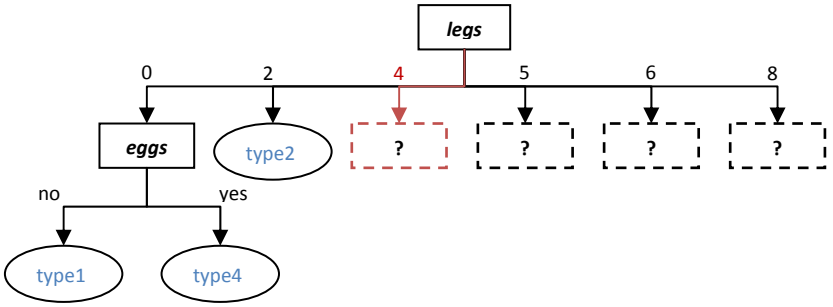
We reached a homogeneous (pure) node with no entropy; therefore we label this node with *type4.*



Next we consider *legs=2*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	yes	no	2	type2
no	yes	no	2	type2

We reached a homogeneous (pure) node with no entropy; therefore we label this node with *type2.*



Then we consider *legs=4*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	yes	yes	4	type5
yes	no	yes	4	type1
yes	no	yes	4	type1
no	yes	yes	4	type5
yes	no	yes	4	type1

The entropy over the above instances is:

$$\left(-\frac{3}{5}\times\log_2\left(\frac{3}{5}\right)\right)+3\times\left(-\frac{0}{5}\right)+\left(-\frac{2}{5}\times\log_2\left(\frac{2}{5}\right)\right)+2\times\left(-\frac{0}{5}\right)=0.971$$

Now we calculate the entropy for the attribute: hair, eggs, and toothed over these instances.

- hair:

hair	type						
	type1	type2	type3	type4	type5	type6	type7
no (2)	0	0	0	0	2	0	0
yes (3)	3	0	0	0	0	0	0

$$\frac{2}{5}\times\left[4\times\left(-\frac{0}{2}\right)+\left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right)+2\times\left(-\frac{0}{2}\right)\right]+\frac{3}{5}\times\left[\left(-\frac{3}{3}\times\log_2\left(\frac{3}{3}\right)\right)+6\times\left(-\frac{0}{3}\right)\right]=0.0$$

- eggs:

eggs	type						
	type1	type2	type3	type4	type5	type6	type7
no (3)	3	0	0	0	0	0	0
yes (2)	0	0	0	0	2	0	0

$$\frac{3}{5}\times\left[\left(-\frac{3}{3}\times\log_2\left(\frac{3}{3}\right)\right)+6\times\left(-\frac{0}{3}\right)\right]+\frac{2}{5}\times\left[4\times\left(-\frac{0}{2}\right)+\left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right)+2\times\left(-\frac{0}{2}\right)\right]=0.0$$

- toothed:

toothed	type						
	type1	type2	type3	type4	type5	type6	type7
no (0)	0	0	0	0	0	0	0
yes (5)	3	0	0	0	2	0	0

$$\frac{0}{5}+\frac{5}{5}\times\left[\left(-\frac{3}{5}\times\log_2\left(\frac{3}{5}\right)\right)+3\times\left(-\frac{0}{5}\right)+\left(-\frac{2}{5}\times\log_2\left(\frac{2}{5}\right)\right)+2\times\left(-\frac{0}{5}\right)\right]=0.971$$

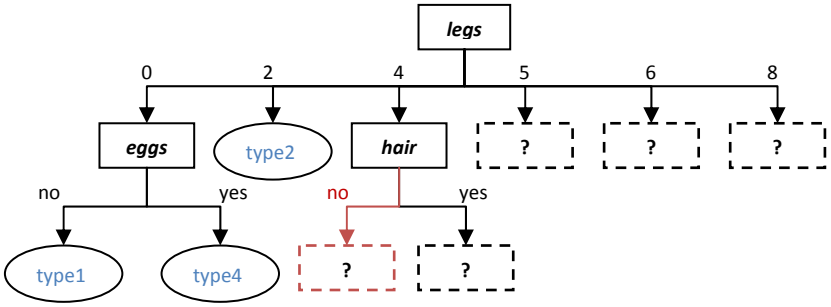
Therefore at this stage, the information gains for the above attributes are:

$$gain(hair) = 0.971 - 0.0 = 0.971$$

$$gain(eggs) = 0.971 - 0.0 = 0.971$$

$$gain(toothed) = 0.971 - 0.971 = 0.0$$

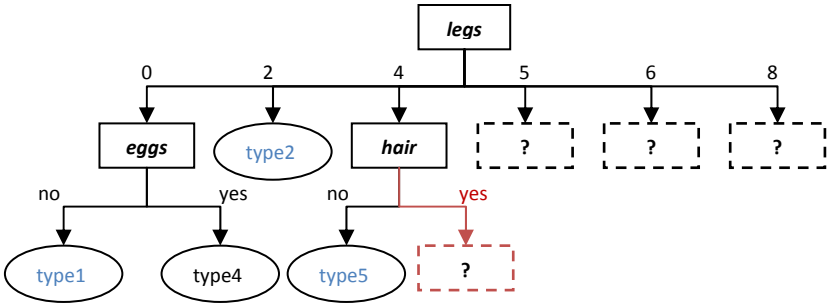
Since both *hair* and *eggs* have the same information gain value 0.971, we can select either one for the next level of the tree. We select *hair* (since it is the first one to appear in the list of attributes) as the splitting attribute, and we continue recursively on the children nodes.



Next we consider *hair=no*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	yes	yes	4	type5
no	yes	yes	4	type5

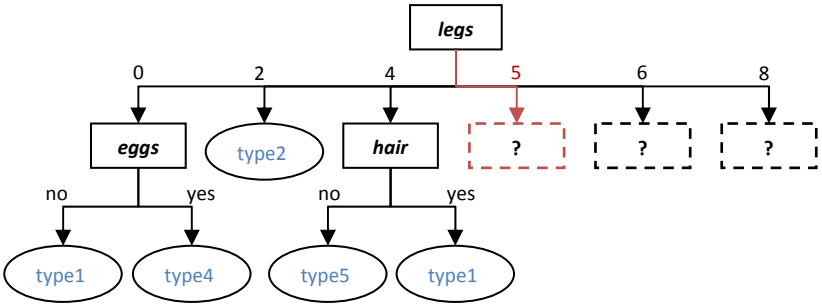
We reached a homogeneous (pure) node with no entropy; therefore we label this node with *type5.*



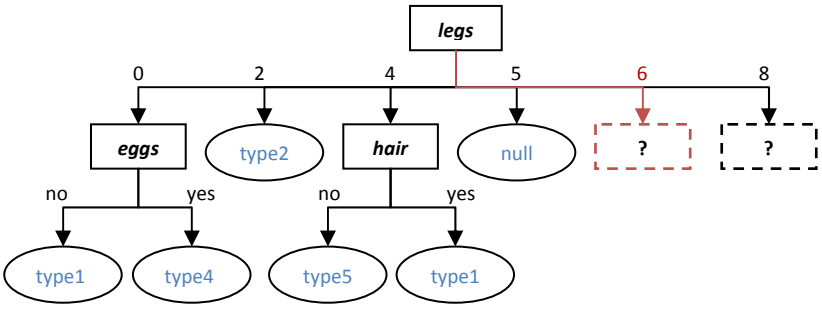
Next we consider *hair=yes*. We have the following instances at this point:

hair	eggs	toothed	legs	type
yes	no	yes	4	type1
yes	no	yes	4	type1
yes	no	yes	4	type1

We reached a homogeneous (pure) node with no entropy; therefore we label this node with *type1.*



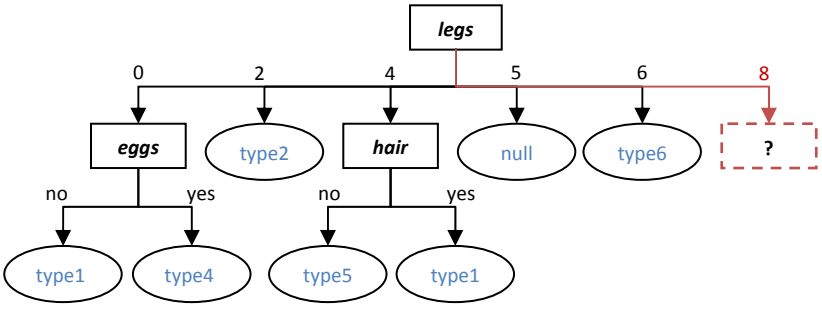
Next we consider *legs=5*. There are no instances reaching this node. Weka System would label this node as *null*. Alternatively, we can choose the majority class over the instances at the parent of this node (which would be the root, with majority class tied between *type1* and *type6*, and selecting *type1* the first to appear as the prediction class).



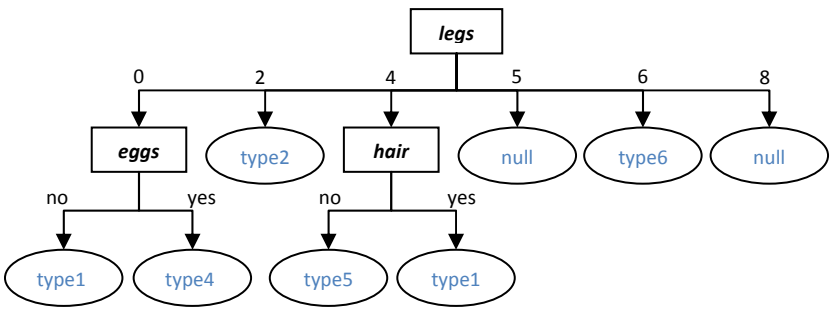
Then we consider *legs=6*. We have the following instances at this point:

hair	eggs	toothed	legs	type
no	yes	no	6	type6
no	yes	no	6	type6
no	yes	no	6	type6
yes	yes	no	6	type6

We reached a homogeneous (pure) node with no entropy; therefore we label this node with *type6.*



Next we consider *legs=8*. There are no instances reaching this node. Again, we would label this node as *null*. Alternatively, we can choose the majority class over the instances at the parent of this node (which would be the root, with *type1* as the majority class).



2. Accuracy on test data [10 pts]

animal	hair	eggs	toothed	legs	actual class	predicted class
bass	no	yes	yes	0	type4	**type4**
buffalo	yes	no	yes	4	type1	**type1**
chicken	no	yes	no	2	type2	**type2**
crayfish	no	yes	no	6	type7	**type6**
deer	yes	no	yes	4	type1	**type1**
dove	no	yes	no	2	type2	**type2**
goat	yes	no	yes	4	type1	**type1**
pike	no	yes	yes	0	type4	**type4**
toad	no	yes	yes	4	type5	**type5**
vampire	yes	no	yes	2	type1	**type2**

Classification accuracy: [5 pts]

Correctly Classified Instances	8	80 %
Incorrectly Classified Instances	2	20 %
Total Number of Instances	10	

Confusion Matrix: [10 pts]

```
a b c d e f g   <-- classified as
3 1 0 0 0 0 0 | a = type1
0 2 0 0 0 0 0 | b = type2
0 0 0 0 0 0 0 | c = type3
0 0 0 2 0 0 0 | d = type4
0 0 0 0 1 0 0 | e = type5
0 0 0 0 0 0 0 | f = type6
0 0 0 0 0 1 0 | g = type7
```

3. Testing

1) [5 pts]

animal	hair	eggs	toothed	legs	predicted class
scorpion	no	no	no	8	**null**

As it was explained during the construction of the decision tree, if training instances do not reach a certain branch, we can either output a *null* prediction, or we can predict the majority class of the parent node.
In this case, we get either *null* or *type1,* if the majority class approach is used.

2) [10 pts]

animal	hair	eggs	toothed	Legs	predicted class
no-name	yes	no	no	?	**type1**

Assume this time J4.8 instead of ID3 on the tree constructed above. Since at the root node the value of *legs* is missing, this instance is split down the branch weighted according to the probability **P**(*legs=X*) of each *legs* value over the train dataset.

$w_{legs=0}=4/15$, $w_{legs=2}=2/15$, $w_{legs=4}=5/15$, $w_{legs=5}=0/15$, $w_{legs=6}=4/15$, $w_{legs=8}=0/15$

From here we continue down each branch, and get the prediction weighted by the previous probabilities:

legs=0 branch: since *eggs=no* then we get *type=**type1***
legs=2 branch: we get *type=**type2***
legs=4 branch: since *hair=yes* then we get *type=**type1***
legs=5 branch: we get ***null***
legs=6 branch: we get *type=**type6***
legs=8 branch: we get ***null***

branch	weight	type=
legs=0	4/15	type1
legs=2	2/15	type2
legs=4	5/15	type1
legs=5	0/15	null
legs=6	4/15	type6
legs=8	0/15	null

By summing the results we get:

type=	probability
type1	4/15+5/15 = 9/15
type2	2/15
type4	0
type5	0
type6	4/15
type8	0
null	0

Then the predicted class would be the most probable one, namely ***type1***.