Assignment 3 – Solution

Machine Learning
MSc Business Analytics

Wolfram Wiesemann

1 Individual Assignment

For the tree with the root node only, the purity is

$$H(R) = -\frac{11}{25}\log_2\left(\frac{11}{25}\right) - \frac{14}{25}\log_2\left(\frac{14}{25}\right) = 0.989.$$

If we split on the 'day' predictor, we get

$$\begin{array}{c}\text{weekday} \to 13 \text{ no, } 7 \text{ yes;}\\ \text{weekend} \to 1 \text{ no, } 4 \text{ yes.}\end{array}$$

The information gain is

$$\begin{aligned}0.989 + &\frac{20}{25}\left(0.65 \cdot \log_2(0.65) + 0.35 \cdot \log_2(0.35)\right)\\ + &\frac{5}{25}\left(0.8 \cdot \log_2(0.8) + 0.2 \cdot \log_2(0.2)\right) = 0.097.\end{aligned}$$

If we split on the 'weather' predictor, we get

$$\begin{array}{c}\text{rainy} \to 1 \text{ no, } 6 \text{ yes;}\\ \text{sunny} \to 13 \text{ no, } 5 \text{ yes.}\end{array}$$

The information gain is

$$\begin{aligned}0.989 + &\frac{7}{25}\left(0.14 \cdot \log_2(0.14) + 0.86 \cdot \log_2(0.86)\right)\\ + &\frac{18}{25}\left(0.277 \cdot \log_2(0.277) + 0.723 \cdot \log_2(0.723)\right) = 0.21.\end{aligned}$$

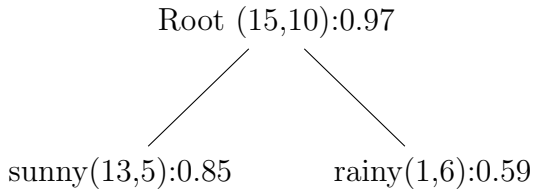
If we split on the 'time' predictor, we get

$$\begin{array}{c}\text{8am} \to 7 \text{ no 5, yes;}\\ \text{1pm} \to 7 \text{ no, } 6 \text{ yes.}\end{array}$$

The information gain is

$$0.989 + \frac{12}{25}\left(0.42 \cdot \log_2(0.42) + 0.58 \cdot \log_2(0.58)\right) + \frac{13}{25}\left(0.53 \cdot \log_2(0.53) + 0.47 \cdot \log_2(0.47)\right) = 0.056.$$

We therefore first split on the 'weather' predictor and obtain:



Here, the notation of the nodes is "predictor value (# of samples with no traffic, # of sample with traffic) : Impurity". Now let's split the left leaf node first.[1] The purity of this node is

$$-\frac{13}{18}\log_2\left(\frac{13}{18}\right) - \frac{5}{18}\log_2\left(\frac{5}{18}\right) = 0.85.$$

If we split on the 'day' predictor, we obtain

$$\begin{array}{c}\text{weekday} \to 13 \text{ no, } 2 \text{ yes;} \\ \text{weekend} \to 0 \text{ no, } 3 \text{ yes.}\end{array}$$

The information gain is

$$0.85 + \frac{15}{18}\left(0.14 \cdot \log_2(0.14) + 0.86 \cdot \log_2(0.86)\right) = 0.37.$$

Similarly, if we split on the 'time' predictor, we obtain

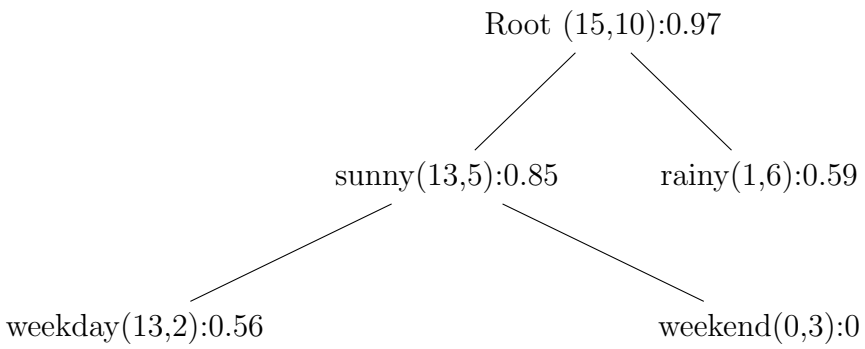
$$\begin{array}{c}\text{8am} \to 6 \text{ no, } 3 \text{ yes;} \\ \text{1pm} \to 7 \text{ no, } 2 \text{ yes.}\end{array}$$

The information gain is

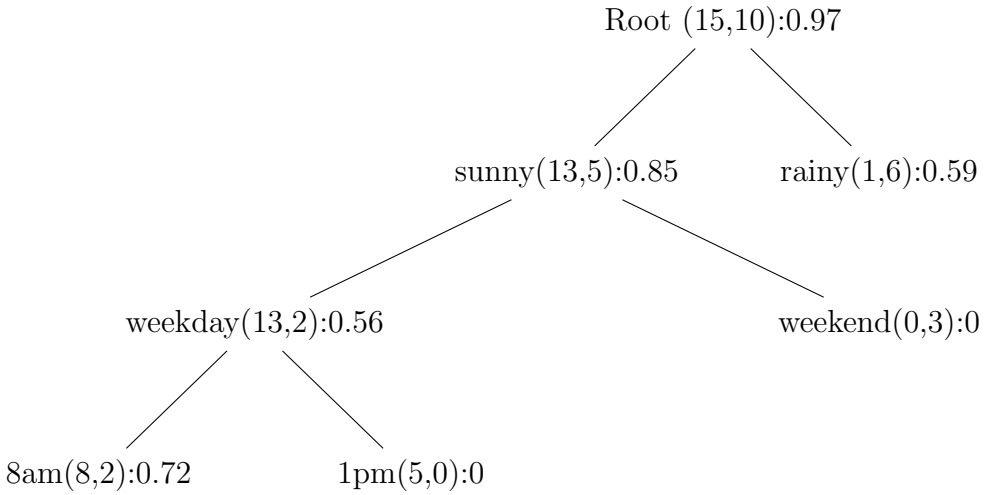
$$0.85 + \frac{9}{18}\left(0.66 \cdot \log_2(0.66) + 0.33 \cdot \log_2(0.33)\right) + \frac{9}{18}\left(0.28 \cdot \log_2(0.28) + 0.72 \cdot \log_2(0.72)\right) = 0.011.$$

Therefore we split on the 'day' predictor:

[1] Note that since we are growing a complete tree, we do not have to select the node with the best split – each 'splittable' node will be split anyway.



We can now split the leftmost leaf node on the remaining 'time' predictor:



Finally we need to decide how to split the 'rainy(1,6):0.59' node. The purity of the node is 0.59. If we split on the 'day' predictor, we obtain

$$\begin{array}{c}\text{weekday} \rightarrow 0 \text{ no, } 5 \text{ yes;} \\ \text{weekend} \rightarrow 1 \text{ no, } 1 \text{ yes}\end{array}$$

with an information gain of

$$0.59 + \frac{2}{7}\left(0.5 \cdot \log_2(0.5) + 0.5 \cdot \log_2(0.5)\right) = 0.274.$$

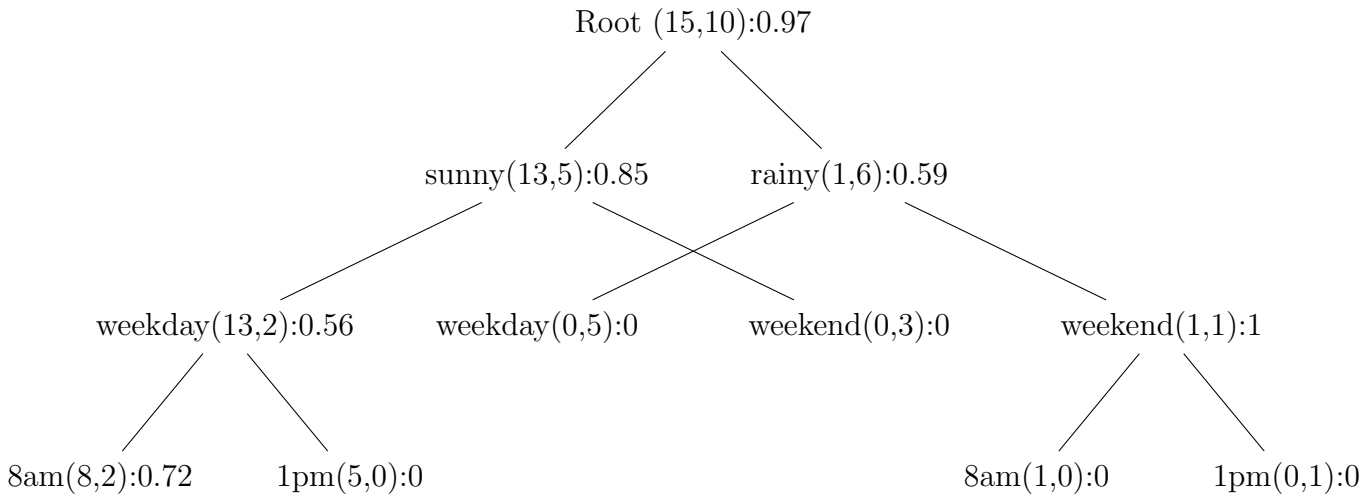
If we split on the 'time' predictor, we obtain

$$\begin{array}{c}\text{8am} \rightarrow 1 \text{ no, } 2 \text{ yes;} \\ \text{1pm} \rightarrow 0 \text{ no, } 4 \text{ yes}\end{array}$$

with an information gain of

$$0.59 + \frac{3}{7}\left(0.66 \cdot \log_2(0.66) + 0.33 \cdot \log_2(0.33)\right) = 0.2.$$

We thus first split on the 'day' predictor and then on the 'time' predictor:



The tree misclassifies 2 'traffic' samples out of the 25 training examples. The sensitivity is 2/11, and the specificity is 14/14.

The predictions on the test set are:

day	**weather**	**time**	**traffic**	**prediction**
weekend	rainy	8am	no	no
weekday	sunny	8am	yes	no
weekend	sunny	1pm	yes	yes
weekday	sunny	8am	no	no
weekend	sunny	1pm	yes	yes
weekday	rainy	8am	no	yes
weekday	sunny	8am	yes	no
weekday	sunny	1pm	no	no
weekday	sunny	1pm	no	no
weekday	sunny	1pm	no	no
weekend	rainy	1pm	yes	yes
weekday	sunny	8am	yes	no
weekday	sunny	1pm	no	no
weekday	rainy	1pm	yes	yes
weekday	sunny	1pm	no	no

The classification tree misclassifies 4 out of the 15 samples. The sensitivity is 4/7, and the specificity is 7/8.