# COMP3411/9814 Artificial Intelligence 20T0, 2020

# Tutorial Solutions - Week 4 tutorial 7

**Week 4: Learning and Decision Trees (Week 3 Lecture 3)**

Tutorial 7: Learning, Decision Trees

## 7.1 ( Activity 8.1: Decision Trees - Open Learning)

Complete the Activity on the Decision Trees page:

Consider the task of predicting whether children are likely to be hired to play members of the Von Trapp Family in a production of The Sound of Music, based on these data:

| height | hair | eyes | hired |
|---|---|---|---|
| short | blond | blue | + |
| tall | red | blue | + |
| tall | blond | blue | + |
| tall | blond | brown | – |
| short | dark | blue | – |
| tall | dark | blue | – |
| tall | dark | brown | – |
| short | blond | brown | – |

a. Compute the information (entropy) gain for each of the three attributes (height, hair, eyes) in terms of classifying objects as belonging to the class, + or – .

There are 3 objects in class '+' and 5 in '-', so the entropy is:

Entropy(parent) = $\Sigma_i P_i \log_2 P_i$ = -(3/8)log(3/8) - (5/8)log(5/8) = 0.954

Suppose we split on height:

```
      [3,5]
      /   \
  [1,2]   [2,3]
```

Of the 3 'short' items, 1 is '+' and 2 are '-', so Entropy(short) = -(1/3)log(1/3) - (2/3)log(2/3) = 0.918

Of the 5 'tall' items, 2 are '+' and 3 are '-', so Entropy(tall) = -(2/5)log(2/5) - (3/5)log(3/5) = 0.971

The average entropy after splitting on 'height' is Entropy(height) = (3/8)(0.918) + (5/8)(0.971) = 0.951

The information gained by testing this attribute is: 0.954 - 0.951 = 0.003 (i.e. very little)

If we try splitting on 'hair' we find that the branch for 'dark' has 3 items, all '-' and the branch for 'red' has 1 item, in '+'. Thus, these branches require no further information to make a decision. The branch for 'blond' has 2 '+' and 2 '-' items and so requires 1 bit. That is,

Entropy(hair) = (3/8)(0) + (1/8)(0) + (4/8)(1) = 0.5

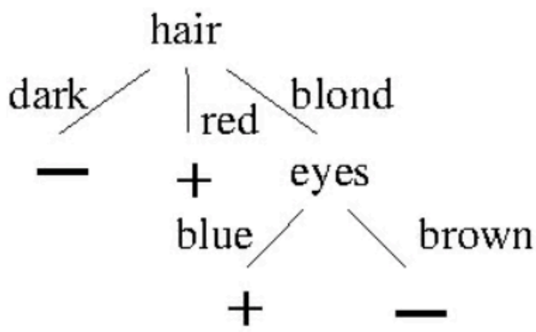
and the information gained by testing hair is 0.954 - 0.5 = 0.454 bits.

By a similar calculation, the entropy for testing 'eyes' is (5/8)(0.971) + (3/8)(0) = 0.607, so the information gained is 0.954 - 0.607 = 0.347 bits.

Thus 'hair' gives us the maximum information gain.

b. Construct a decision tree based on the minimum entropy principle.

Since the 'blond' branch for hair still contains a mixed population, we need to apply the procedure recursively to these four items. Note that we now only need to test 'height' and 'eyes' since the 'hair' attribute has already been used. If we split on 'height', the branch for 'tall' and 'short' will each contain one '+' and one '-', so the entropy gain is zero. If we split on 'eyes', the 'blue' brach contains two '+'s and the 'brown' branch two '-'s, so the tree is complete:



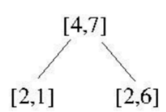

## 7.2 (Activity 8.2: Laplace Pruning - Open Learning)

Complete the Activity on the Decision Tree Pruning page:

The Laplace error estimate for pruning a node in a Decision Tree is given by:

$$E = 1 - \frac{n+1}{N+k}$$

where $N$ is the total number of items, $n$ is the number of items in the majority class and $k$ is the number of classes. Given the following sub-tree, should the children be pruned or not? Show your calculations.



The Laplace error estimate for pruning a node in a Decision Tree is given by:

$$E = 1 - \frac{n+1}{N+k}$$

where *N* is the total number of items, *n* is the number of items in the majority class and *k* is the number of classes. Given the following subtree, should the children be pruned or not? Show your calculations.


Error(Parent) = 1 - (7+1)/(11+2) = 1 - 8/13 = 5/13 = 0.385

Error(Left) = 1 - (2+1)/(3+2) = 1 - 3/5 = 2/5 = 0.4

Error(Right) = 1 - (6+1)/(8+2) = 1 - 7/10 = 3/10 = 0.3

Backed Up Error = (3/11)*(0.4) + (8/11)*(0.3) = 0.327 < 0.385

Since Error of Parent is larger than Backed Up Error ⇒ Don't Prune

7.3 Construct a Decision Tree for the following set of examples.

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

What class is assigned to the instance {D15, Sunny, Hot, High, Weak}?

1.

(i) $Values(Outlook) = \{sunny, overcast, rain\}$
$S = [9+, 5-]$
$S_{sunny} \leftarrow [2+, 3-]$
$S_{overcast} \leftarrow [4+, 0-]$
$S_{rain} \leftarrow [3+, 2-]$
$Gain(S, Outlook) = Entropy(S) - \sum_{v=\{sunny,overcast,rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$
$= Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - \frac{4}{14} Entropy(S_{overcast}) - \frac{5}{14} Entropy(S_{rain})$
$= 0.940 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$
$= 0.247$
$Entropy(S) = Entropy([9+, 5-]) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14}$
$= 0.940$
$Entropy(S_{sunny}) = Entropy([2+, 3-]) = -\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5} = 0.971$
$Entropy(S_{overcast}) = Entropy([4+, 0-]) = -\frac{4}{4} log_2 \frac{4}{4} - \frac{0}{4} log_2 \frac{0}{4} = 0$
$Entropy(S_{rain}) = Entropy([3+, 2-]) = -\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5} = 0.971$

(ii) $Values(Temperature) = \{hot, mild, cool\}$
$S = [9+, 5-]$
$S_{hot} \leftarrow [2+, 2-]$
$S_{mild} \leftarrow [4+, 2-]$
$S_{cool} \leftarrow [3+, 1-]$
$Gain(S, Temperature) = Entropy(S) - \sum_{v=\{hot,mild,cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$
$= Entropy(S) - \frac{4}{14} Entropy(S_{hot}) - \frac{6}{14} Entropy(S_{mild}) - \frac{4}{14} Entropy(S_{cool})$
$= 0.940 - \frac{4}{14} \times 1.00 - \frac{6}{14} \times 0.918 - \frac{4}{14} \times 0.232$
$= 0.029$
$Entropy(S_{hot}) = Entropy([2+, 2-]) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} = 1.00$
$Entropy(S_{mild}) = Entropy([4+, 2-]) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0.918$
$Entropy(S_{cool}) = Entropy([3+, 1-]) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.232$

(iii) $Values(Humidity) = \{high, normal\}$
$S = [9+, 5-]$
$S_{high} \leftarrow [3+, 4-]$
$S_{normal} \leftarrow [6+, 1-]$
$Gain(S, Humidity) = Entropy(S) - \sum_{v=\{high,normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$
$= Entropy(S) - \frac{7}{14} Entropy(S_{high}) - \frac{7}{14} Entropy(S_{normal})$
$= 0.940 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592$
$= 0.152$
$Entropy(S_{high}) = Entropy([3+, 4-]) = -\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7} = 0.985$
$Entropy(S_{mild}) = Entropy([6+, 1-]) = -\frac{6}{7} log_2 \frac{6}{7} - \frac{1}{7} log_2 \frac{1}{7} = 0.592$

(iv) $Values(Wind) = \{weak, strong\}$
$S = [9+, 5-]$
$S_{weak} \leftarrow [6+, 2-]$
$S_{strong} \leftarrow [3+, 3-]$
$Gain(S, Wind) = Entropy(S) - \sum_{v=\{weak,strong\}} \frac{|S_v|}{|S|} Entropy(S_v)$
$= Entropy(S) - \frac{8}{14} Entropy(S_{weak}) - \frac{6}{14} Entropy(S_{strong})$
$= 0.940 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1.00$
$= 0.048$
$Entropy(S_{weak}) = Entropy([6+, 2-]) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{6}{8} log_2 \frac{6}{8} = 0.811$
$Entropy(S_{strong}) = Entropy([3+, 3-]) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = 1.00$

2.

(i) $Values(Temperature) = \{hot, mild, cool\}$
$S_{sunny} = [2+, 3-]$
$S_{sunny,hot} \leftarrow [0+, 2-]$
$S_{sunny,mild} \leftarrow [1+, 1-]$
$S_{sunny,cool} \leftarrow [1+, 0-]$
$Gain(S_{sunny}, Temperature) = Entropy(S_{sunny}) - \sum_{v=\{hot,mild,cool\}} \frac{|S_{sunny,v}|}{|S_{sunny}|} Entropy(S_{sunny,v})$
$= Entropy(S) - \frac{2}{5}Entropy(S_{sunny,hot}) - \frac{2}{5}Entropy(S_{sunny,mild}) - \frac{1}{5}Entropy(S_{sunny,cool})$
$= 0.971 - \frac{2}{5} \times 0.00 - \frac{2}{5} \times 1.00 - \frac{1}{5} \times 0.00$
$= 0.571$
$Entropy(S_{sunny}) = Entropy([2+, 3-]) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$
$Entropy(S_{sunny,hot}) = Entropy([0+, 2-]) = -\frac{0}{2}log_2\frac{0}{2} - \frac{2}{2}log_2\frac{2}{2} = 0.00$
$Entropy(S_{sunny,mild}) = Entropy([1+, 1-]) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1.00$
$Entropy(S_{sunny,cool}) = Entropy([1+, 0-]) = -\frac{1}{1}log_2\frac{1}{1} - \frac{0}{1}log_2\frac{0}{1} = 0.00$

(ii) $Values(Humidity) = \{high, normal\}$
$S_{sunny} = [2+, 3-]$
$S_{sunny,high} \leftarrow [0+, 3-]$
$S_{sunny,normal} \leftarrow [2+, 0-]$
$Gain(S, Humidity) = Entropy(S_{sunny}) - \sum_{v=\{high,normal\}} \frac{|S_{sunny,v}|}{|S_{sunny}|} Entropy(S_{sunny,v})$
$= Entropy(S_{sunny}) - \frac{3}{5}Entropy(S_{sunny,high}) - \frac{2}{5}Entropy(S_{sunny,normal})$
$= 0.971 - \frac{3}{5} \times 0.00 - \frac{2}{5} \times 0.00$
$= 0.971$
$Entropy(S_{sunny,high}) = Entropy([0+, 3-]) = -\frac{0}{3}log_2\frac{0}{3} - \frac{3}{3}log_2\frac{3}{3} = 0.00$
$Entropy(S_{sunny,mild}) = Entropy([2+, 0-]) = -\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2} = 0.00$

(iii) $Values(Wind) = \{weak, strong\}$
$S_{sunny} = [2+, 3-]$
$S_{weak} \leftarrow [1+, 2-]$
$S_{strong} \leftarrow [1+, 1-]$
$Gain(S, Wind) = Entropy(S_{sunny}) - \sum_{v=\{weak,strong\}} \frac{|S_{sunny,v}|}{|S_{sunny}|} Entropy(S_{sunny,v})$
$= Entropy(S) - \frac{3}{5}Entropy(S_{sunny,weak}) - \frac{2}{5}Entropy(S_{strong})$
$= 0.971 - \frac{3}{5} \times 0.918 - \frac{2}{5} \times 1.00$
$= 0.020$
$Entropy(S_{weak}) = Entropy([1+, 2-]) = -\frac{1}{3}log_2\frac{1}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.918$
$Entropy(S_{strong}) = Entropy([1+, 1-]) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1.00$

3.

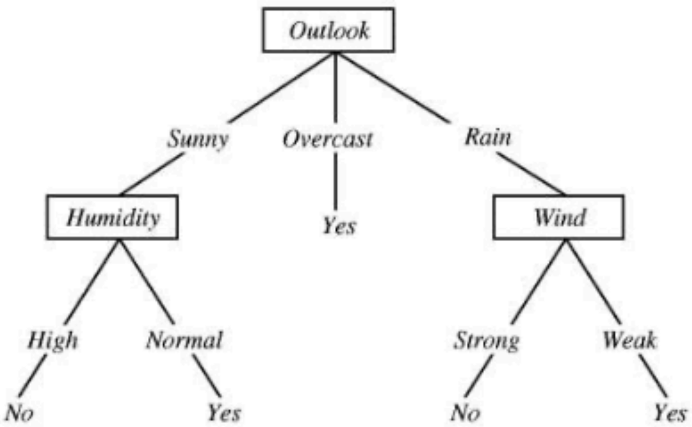
(i) $Values(Temperature) = \{hot, mild, cool\}$
$S_{rain} = [2+, 3-]$
$S_{rain,hot} \leftarrow [0+, 0-]$
$S_{rain,mild} \leftarrow [2+, 1-]$
$S_{rain,cool} \leftarrow [1+, 1-]$
$Gain(S_{rain}, Temperature) = Entropy(S_{rain}) - \sum_{v=\{hot,mild,cool\}} \frac{|S_{rain,v}|}{|S_{rain}|} Entropy(S_{rain,v})$
$= Entropy(S) - \frac{0}{5}Entropy(S_{rain,hot}) - \frac{3}{5}Entropy(S_{rain,mild}) - \frac{2}{5}Entropy(S_{rain,cool})$
$= 0.971 - \frac{0}{5} \times 0.00 - \frac{3}{5} \times 0.918 - \frac{2}{5} \times 1.00$
$= 0.020$
$Entropy(S_{rain}) = Entropy([3+, 2-]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$
$Entropy(S_{rain,hot}) = Entropy([0+, 0-]) = -\frac{0}{0}log_2\frac{0}{0} - \frac{0}{0}log_2\frac{0}{0} = 0.00$
$Entropy(S_{rain,mild}) = Entropy([2+, 1-]) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.918$
$Entropy(S_{rain,cool}) = Entropy([1+, 1-]) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1.00$

(ii) $Values(Humidity) = \{high, normal\}$
$S_{rain} = [3+, 2-]$

$S_{rain,high} \leftarrow [1+, 1-]$
$S_{rain,normal} \leftarrow [1+, 1-]$
$Gain(S, Humidity) = Entropy(S_{rain}) - \sum_{v=\{high,normal\}} \frac{|S_{rain,v}|}{|S_{rain}|} Entropy(S_{rain,v})$
$= Entropy(S_{rain}) - \frac{2}{5} Entropy(S_{rain,high}) - \frac{3}{5} Entropy(S_{rain,normal})$
$= 0.971 - \frac{2}{5} \times 1.00 - \frac{3}{5} \times 0.551$
$= 0.020$
$Entropy(S_{rain,high}) = Entropy([1+, 1-]) = -\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2} = 1.00$
$Entropy(S_{rain,mild}) = Entropy([2+, 1-]) = -\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3} = 0.551$

$S_{rain,high} \leftarrow [1+, 1-]$
$S_{rain,normal} \leftarrow [1+, 1-]$
$Gain(S, Humidity) = Entropy(S_{rain}) - \sum_{v=\{high,normal\}} \frac{|S_{rain,v}|}{|S_{rain}|} Entropy(S_{rain,v})$
$= Entropy(S_{rain}) - \frac{2}{5} Entropy(S_{rain,high}) - \frac{3}{5} Entropy(S_{rain,normal})$
$= 0.971 - \frac{2}{5} \times 1.00 - \frac{3}{5} \times 0.551$
$= 0.020$
$Entropy(S_{rain,high}) = Entropy([1+, 1-]) = -\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2} = 1.00$
$Entropy(S_{rain,mild}) = Entropy([2+, 1-]) = -\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3} = 0.551$

(iii) $Values(Wind) = \{weak, strong\}$
$S_{rain} = [3+, 2-]$
$S_{weak} \leftarrow [3+, 0-]$
$S_{strong} \leftarrow [0+, 2-]$
$Gain(S, Wind) = Entropy(S_{rain}) - \sum_{v=\{weak,strong\}} \frac{|S_{rain,v}|}{|S_{rain}|} Entropy(S_{rain,v})$
$= Entropy(S) - \frac{3}{5} Entropy(S_{rain,weak}) - \frac{2}{5} Entropy(S_{strong})$
$= 0.971 - \frac{3}{5} \times 0.00 - \frac{2}{5} \times 0.00$
$= 0.971$
$Entropy(S_{weak}) = Entropy([3+, 0-]) = -\frac{3}{3} log_2 \frac{3}{3} - \frac{0}{3} log_2 \frac{0}{3} = 0.00$
$Entropy(S_{strong}) = Entropy([0+, 2-]) = -\frac{0}{2} log_2 \frac{0}{2} - \frac{2}{2} log_2 \frac{2}{2} = 0.00$




So the example is assigned the *No* class.

7.4 Consider a Naive Bayes classifier for the same set of examples. What class is now assigned to the instance {D15, Sunny, Hot, High, Weak}?

$P(Play|Outlook, Temperature, Humidity, Wind) = \frac{P(Outlook, Temperature, Humidity, Wind|Play).P(Play)}{P(Outlook, Temperature, Humidity, Wind)}$

Similarly for $P(\neg Play|Outlook, Temperature, Humidity, Wind)$

$P(Sunny, Hot, High, Weak|Play).P(Play)$
$= P(Sunny|Play).P(Hot|Play).P(High|Play).P(Weak|Play).P(Play)$ by independence
$= 2/9 * 2/9 * 3/9 * 6/9 * 9/14$
$= 0.00705$

$P(Sunny, Hot, High, Weak|\neg Play).P(\neg Play)$
$= P(Sunny|\neg Play).P(Hot|\neg Play).P(High|\neg Play).P(Weak|\neg Play).P(\neg Play)$ by independence
$= 3/5 * 2/5 * 4/5 * 2/5 * 5/14$
$= 0.02743$

Again the example is assigned the *No* class.