

A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

2020-10-4

# COMP3411

## Assignment 2

Haowei Lou

Several thin, curved lines in dark blue and light grey originate from the bottom left corner and sweep upwards and to the right.

Haowei Lou  
Z5258575

Q1

a)

<i>Example</i>	<i>Author</i>	<i>Thread</i>	<i>Length</i>	<i>Where_read</i>	<i>User_action</i>
$e_1$	known	new	long	home	skips
$e_2$	unknown	new	short	work	reads
$e_3$	unknown	followup	long	work	skips
$e_4$	known	followup	long	home	skips
$e_5$	known	new	short	home	reads
$e_6$	known	followup	long	work	skips
$e_7$	unknown	followup	short	work	skips
$e_8$	unknown	new	short	work	reads
$e_9$	known	followup	long	home	skips
$e_{10}$	known	new	long	work	skips
$e_{11}$	unknown	followup	short	home	skips
$e_{12}$	known	new	long	work	skips
$e_{13}$	known	followup	short	home	reads
$e_{14}$	known	new	short	work	reads
$e_{15}$	known	new	short	home	reads
$e_{16}$	known	followup	short	work	reads
$e_{17}$	known	new	short	home	reads
$e_{18}$	unknown	new	short	work	reads
$e_{19}$	unknown	new	long	work	?
$e_{20}$	unknown	followup	short	home	?

The maximum information gain split tree is:

Split on author:

There are 9 skips and 9 reads in the data set.

$$\text{Entropy (parent)} = \sum_i P_i \log_2 P_i = -\frac{9}{18} \log_2 \left(\frac{9}{18}\right) - \frac{9}{18} \log_2 \left(\frac{9}{18}\right) = -\log_2 \left(\frac{9}{18}\right) = -\log_2 2^{-1} = 1$$

For the 12 knowns, 6 of them are skips and 6 are reads.

$$\text{Entropy (known)} = \sum_i P_i \log_2 P_i = -\frac{6}{12} \log_2 \frac{6}{12} - \frac{6}{12} \log_2 \frac{6}{12} = 1$$

For the 6 unknowns, 3 of them are skips and 3 are reads.

$$\text{Entropy (unknown)} = \sum_i P_i \log_2 P_i = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1$$

$$\text{The average entropy after splitting on 'Author', Entropy(Author)} = \frac{12}{18} + \frac{6}{18} = 1$$

Information gained is  $1 - 1 = 0$

Split on thread:

There are 9 skips and 9 reads in the data set.

$$\text{Entropy (parent)} = \sum_i P_i \log_2 P_i = -\frac{9}{18} \log_2 \left(\frac{9}{18}\right) - \frac{9}{18} \log_2 \left(\frac{9}{18}\right) = 1$$

For the 10 news, 3 of them are skip and 7 of them are read

$$\text{Entropy (new)} = \sum_i P_i \log_2 P_i = -\frac{3}{10} \log_2 \left(\frac{3}{10}\right) - \frac{7}{10} \log_2 \left(\frac{7}{10}\right) = 0.8812908992$$

For the 8 followups, 6 of them are skip and 2 of them are read

$$\text{Entropy (followup)} = \sum_i P_i \log_2 P_i = -\frac{6}{8} \log_2 \left(\frac{6}{8}\right) - \frac{2}{8} \log_2 \left(\frac{2}{8}\right) = 0.8112781245$$

$$\text{The average entropy after splitting on 'Thread, Entropy(Thread)} = \frac{10}{18}(0.8812908992) + \frac{8}{18}(0.8112781245) = 0.8501741104 \approx 0.85$$

Information gained by testing this attribute is:  $1 - 0.85 = 0.15$ , which is minor.

Split by length:

There are 9 skips and 9 reads in the data set.

$$\text{Entropy (parent)} = \sum_i P_i \log_2 P_i = -\frac{9}{18} \log_2 \left(\frac{9}{18}\right) - \frac{9}{18} \log_2 \left(\frac{9}{18}\right) = 1$$

For the 7 longs, 7 of them are skip and 0 of them are read

$$\text{Entropy(long)} = \sum_i P_i \log_2 P_i = -\frac{7}{7} \log_2 \left(\frac{7}{7}\right) - \frac{0}{7} \log_2 \left(\frac{0}{7}\right) = 0$$

For the 11 shorts, 2 of them are skip and 9 of them are read.

$$\text{Entropy(short)} = \sum_i P_i \log_2 P_i = -\frac{2}{11} \log_2 \left(\frac{2}{11}\right) - \frac{9}{11} \log_2 \left(\frac{9}{11}\right) = 0.6840384356$$

$$\text{The average entropy after splitting on 'Length, Entropy(Length)} = \frac{7}{18}(0) + \frac{11}{18}(0.6840384356) = 0.4180234884 \approx 0.42$$

Information gained by testing this attribute is:  $1 - 0.42 = 0.58$ , which is big.

Split by Whereread:

There are 9 skips and 9 reads in the data set.

$$\text{Entropy (parent)} = \sum_i P_i \log_2 P_i = -\frac{9}{18} \log_2 \left(\frac{9}{18}\right) - \frac{9}{18} \log_2 \left(\frac{9}{18}\right) = 1$$

For the 8 homes, 4 of them are skip and 4 of them are read

$$\text{Entropy(home)} = \sum_i P_i \log_2 P_i = -\frac{4}{8} \log_2 \left(\frac{4}{8}\right) - \frac{4}{8} \log_2 \left(\frac{4}{8}\right) = 1$$

For the 10 works, 5 of them are skip and 5 of them are read

$$\text{Entropy(work)} = \sum_i P_i \log_2 P_i = -\frac{5}{10} \log_2 \left(\frac{5}{10}\right) - \frac{5}{10} \log_2 \left(\frac{5}{10}\right) = 1$$

$$\text{The average entropy after splitting on 'Whereread, Entropy(Whereread)} = \frac{8}{18} + \frac{10}{18} = 1$$

Information gained is  $1 - 1 = 0$

By compare these four information gained. select Length to be the first node will maximise the total information gained.

Next node:

Split by author:

There are 2 skips and 9 reads in the data set after 7 rows that include 'long' have been deducted.

$$\text{Entropy}(\text{parent}) = \sum_i P_i \log_2 P_i = -\frac{9}{11} \log_2 \left(\frac{9}{11}\right) - \frac{2}{11} \log_2 \left(\frac{2}{11}\right) = 0.6840384356 \approx 0.68$$

For the 6 knowns, 0 of them are skip and 6 of them are read

$$\text{Entropy}(\text{known}) = \sum_i P_i \log_2 P_i = -\frac{0}{6} \log_2 \left(\frac{0}{6}\right) - \frac{6}{6} \log_2 \left(\frac{6}{6}\right) = 0$$

For the 5 unknowns, 2 of them are skip and 3 of them are read

$$\text{Entropy}(\text{unknown}) = \sum_i P_i \log_2 P_i = -\frac{2}{5} \log_2 \left(\frac{2}{5}\right) - \frac{3}{5} \log_2 \left(\frac{3}{5}\right) = 0.9709505945$$

$$\text{The average entropy after splitting on 'Author', Entropy}(\text{author}) = \frac{6}{11} (0) + \frac{5}{11} (0.9709505945) = 0.4413411793 \approx 0.44$$

$$\text{Information gained is } 0.68 - 0.44 = 0.24$$

Split by thread:

There are 2 skips and 9 reads in the data set after 7 rows that include 'long' have been deducted.

$$\text{Entropy}(\text{parent}) = \sum_i P_i \log_2 P_i = -\frac{9}{11} \log_2 \left(\frac{9}{11}\right) - \frac{2}{11} \log_2 \left(\frac{2}{11}\right) = 0.6840384356 \approx 0.68$$

For the 7 news, 0 of them are skip and 7 of them are read

$$\text{Entropy}(\text{new}) = \sum_i P_i \log_2 P_i = -\frac{0}{7} \log_2 \left(\frac{0}{7}\right) - \frac{7}{7} \log_2 \left(\frac{7}{7}\right) = 0$$

For the 4 followups, 2 of them are skip and 2 of them are read

$$\text{Entropy}(\text{followup}) = \sum_i P_i \log_2 P_i = -\frac{2}{4} \log_2 \left(\frac{2}{4}\right) - \frac{2}{4} \log_2 \left(\frac{2}{4}\right) = 1$$

$$\text{The average entropy after splitting on 'Thread', Entropy}(\text{Thread}) = \frac{7}{11} (0) + \frac{4}{11} (1) = 0.3636363636 \approx 0.36$$

$$\text{Information gained is } 0.68 - 0.36 = 0.32$$

Split by Whereread:

There are 2 skips and 9 reads in the data set after 7 rows that include 'long' have been deducted.

$$\text{Entropy}(\text{parent}) = \sum_i P_i \log_2 P_i = -\frac{9}{11} \log_2 \left(\frac{9}{11}\right) - \frac{2}{11} \log_2 \left(\frac{2}{11}\right) = 0.6840384356 \approx 0.684$$

For the 5 homes, 1 of them are skip and 4 of them are read

$$\text{Entropy}(\text{home}) = \sum_i P_i \log_2 P_i = -\frac{1}{5} \log_2 \left(\frac{1}{5}\right) - \frac{4}{5} \log_2 \left(\frac{4}{5}\right) = 0.7219280949$$

For the 6 works, 1 of them are skip and 5 of them are read

$$\text{Entropy}(\text{work}) = \sum_i P_i \log_2 P_i = -\frac{1}{6} \log_2 \left(\frac{1}{6}\right) - \frac{5}{6} \log_2 \left(\frac{5}{6}\right) = 0.6500224216$$

$$\text{The average entropy after splitting on 'Thread', Entropy}(\text{Whereread}) = \frac{5}{11} (0.7219280949) + \frac{6}{11} (0.6500224216) = 0.6827068186 \approx 0.683$$

Information gained is  $0.684 - 0.683 = 0.001$

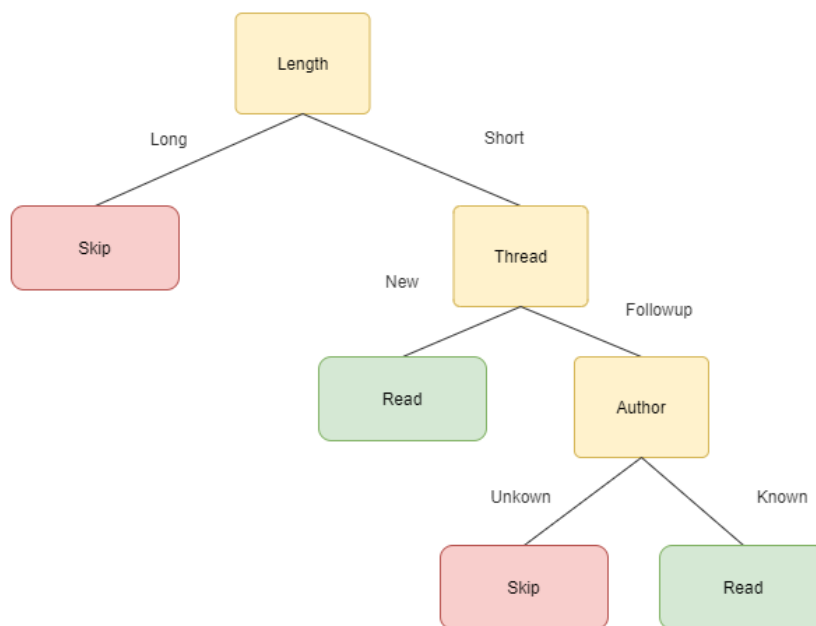
By compare these three information gained. select thread to be the second node will maximise the total information gained.

Next node:

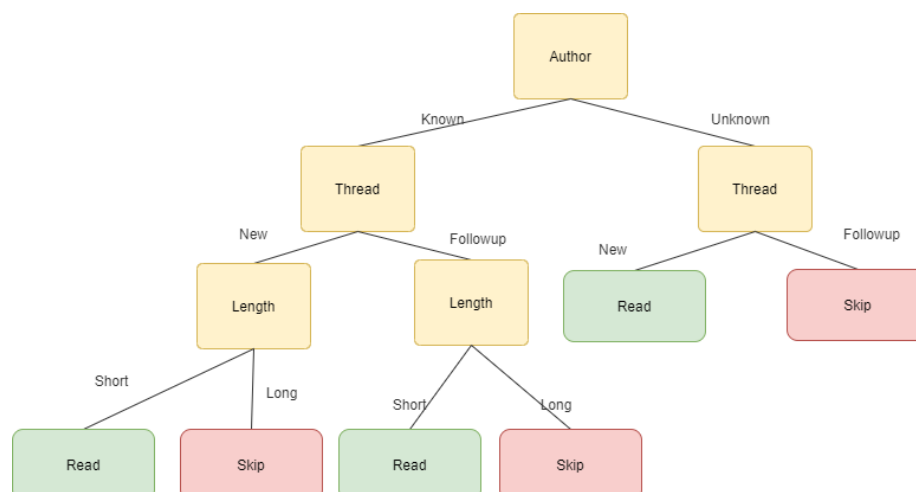
Split by author:

There are 2 skips and 2 reads in the data set after 14 rows that include 'long' and 'thread' have been deducted. And we have a clear classification if split by author, thus the third node should be author.

Optimal decision tree gained by maximizing information gain is:



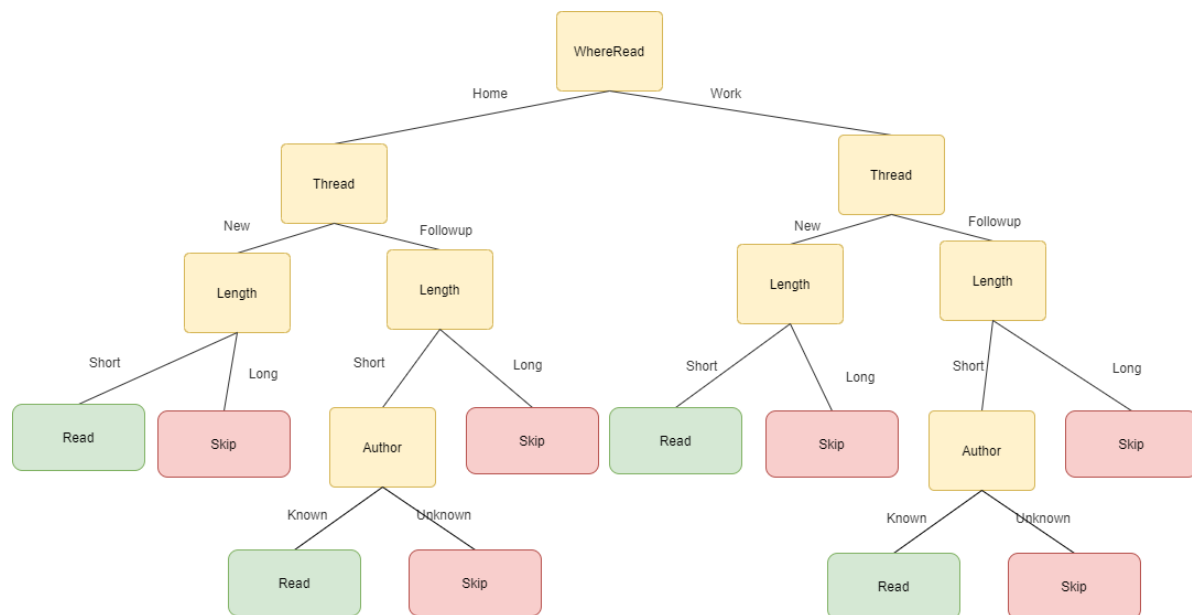
The order of features in new tree is: Author->Thread->Length->Where\_read->User\_actions



The tree found in this part is different with the tree found by maximum information gain. Functions that they represent are also different. For example, e19 in table, <unknown,new,long,work>, optimal tree will go through long->skip, which give answer 'skip'. And the other one will go through unknown->new->read, which give answer 'read'.

b)

The order of feature is <WhereRead, Thread,Length,Author> and tree found is:



This tree has the same function as the maximum information gained tree, where, <unknown,new,long,work> = skip, which is the same as the maximum information gained tree, but different from the tree found in part a.

c)

There is no tree that can correctly classify the training example but represent a different function than those found.

Let read be 'True' and Skip be 'False'

The tree found in maximum information gained can be represented in logic by:

$$(short \text{ AND } new) \text{ OR } (short \text{ AND } \neg new \text{ AND } known)$$

After simplification:

$$(short \text{ AND } new) \text{ OR } (short \text{ AND } known)$$

Tree found in a)

$$(known \text{ AND } new \text{ AND } short) \text{ OR } (known \text{ AND } \neg new \text{ AND } short) \text{ OR } (\neg known \text{ AND } New)$$

After simplification:

$$(\neg known \text{ AND } new) \text{ OR } (short \text{ AND } known)$$

Tree found in b)

$$(home \text{ AND } new \text{ AND } short) \text{ OR } (home \text{ AND } \neg new \text{ AND } short \text{ AND } known) \text{ OR } (\neg home \text{ AND } new \text{ AND } short) \text{ OR } (\neg home \text{ AND } \neg new \text{ AND } short \text{ AND } known)$$

After simplification:

$$(short \text{ AND } new) \text{ OR } (short \text{ AND } known)$$

It has shown that features that matter are Length, Author and Thread in this decision tree.

Thus, ways to permute these three features are:

[Author,Thread,Length],[Author,Length,Thread],[Thread,Author,Length],[Thread,Length,Author],[Length,Author,Thread],[Length,Thread,Author].

Both[Length, Thread, Author], [Thread, Length, Author], [Author, Length, Thread] are same in logic expression because they combine Length and Thread together, and then Author. And has already been found in preceding.

Then there are two possible different decision tree function for this question, which is [Author, Thread, Length], [Thread, Length, Author]. Which has already been found in a) and b). Therefore, there are no tree that correctly classifies the training examples but represents a different function than those found by the preceding algorithms.

## Q2

My code for developing decision tree is as follow:

```
import numpy as np
import pandas as pd
from sklearn import preprocessing
from sklearn import tree
from sklearn.preprocessing import LabelEncoder

data = pd.read_csv('./adult.data',header=None) #load data
features = data.iloc[:, :-1].values
answer = data.iloc[:, -1].values

test = pd.read_csv('./adult.test',header=None,skiprows=1) #load test
features_test = data.iloc[:, :-1].values
answer_test = data.iloc[:, -1].values

encode_data = LabelEncoder()#encode data
encode_data.fit(features[:, 1])
features[:, 1] = encode_data.fit_transform(features[:, 1])
features[:, 3] = encode_data.fit_transform(features[:, 3])
features[:, 5] = encode_data.fit_transform(features[:, 5])
features[:, 6] = encode_data.fit_transform(features[:, 6])
features[:, 7] = encode_data.fit_transform(features[:, 7])
features[:, 8] = encode_data.fit_transform(features[:, 8])
features[:, 9] = encode_data.fit_transform(features[:, 9])
features[:, 13] = encode_data.fit_transform(features[:, 13])

decision_tree = tree.DecisionTreeClassifier(criterion='entropy',splitter='best',max_depth=8,min_samples_split=30,random_state=0)
decision_tree = decision_tree.fit(features,answer)

encode_data.fit(features_test[:, 1])
features_test[:, 1] = encode_data.fit_transform(features_test[:, 1])
features_test[:, 3] = encode_data.fit_transform(features_test[:, 3])
features_test[:, 5] = encode_data.fit_transform(features_test[:, 5])
features_test[:, 6] = encode_data.fit_transform(features_test[:, 6])
features_test[:, 7] = encode_data.fit_transform(features_test[:, 7])
features_test[:, 8] = encode_data.fit_transform(features_test[:, 8])
features_test[:, 9] = encode_data.fit_transform(features_test[:, 9])
features_test[:, 13] = encode_data.fit_transform(features_test[:, 13])
prediction = decision_tree.predict(features_test)
i = 0
correct = 0
while i < len(prediction):
    if prediction[i] == answer_test[i]:
        correct += 1
    i += 1
print('The accuracy rate is:', "{:.2%}".format(correct/i))
```



Haowei Lou z5258575

The accuracy of my method is about 85% in test dataset.