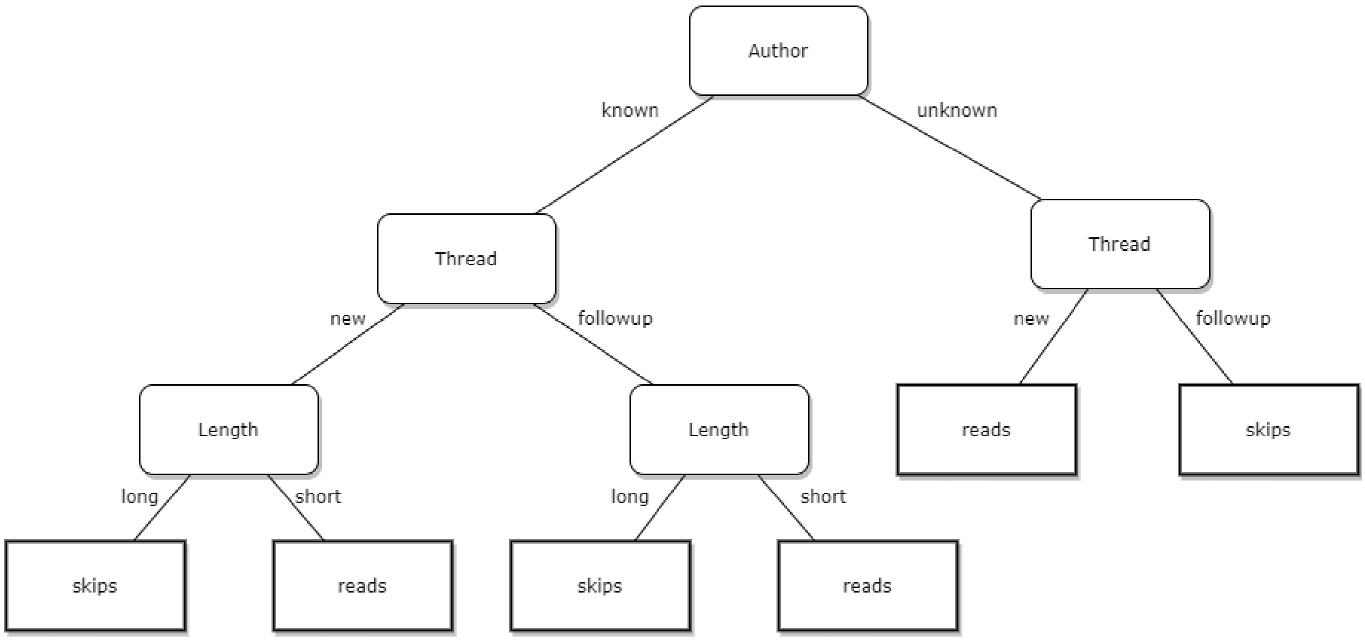
z5261536 Jianjun(Jay) Chen

COMP3411 Assignment 2

JianJun Jay Chen (z5261536)

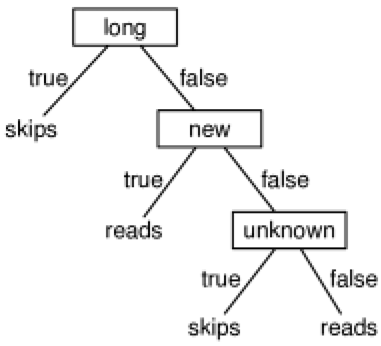
Question 1.1

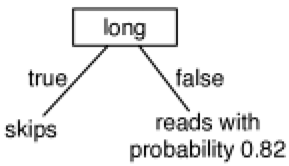
(a) The tree found when the features are in the order $[Author, Thread, Length, WhereRead]$ is: **(function 1)**



This tree represents a different function than that found with the maximum information gain split. (function 2)

From the tree of Figure 7.6 (function 2):





z5261536 Jianjun(Jay) Chen

We know function 2 was built by selecting a feature that gives the maximum information gain which also has fewer splitting compares to function 1. Since the stopping criterion is that all of the examples have the same classification, and by calculating the entropy of function2 we know the $Entropy(parent) = 1$ and the average entropy after splitting on 'Author' $Entropy(Author) = 1$, hence the information gained by this attribute is: 1 – 1 = 0. On the other hand, the average entropy after splitting on 'Length' $Entropy(Length) = \left(\frac{7}{18}\right) * 0 + \left(\frac{11}{18}\right) * \left(-\left(\frac{9}{11}\right) log\left(\frac{9}{11}\right) - \left(\frac{2}{11}\right) log\left(\frac{2}{11}\right)\right) = 0 + \frac{11}{18} * 0.684 = 0.418$, and the information gained by testing $Length$ is: 1 – 0.418 = 0.582, much greater than testing Author.

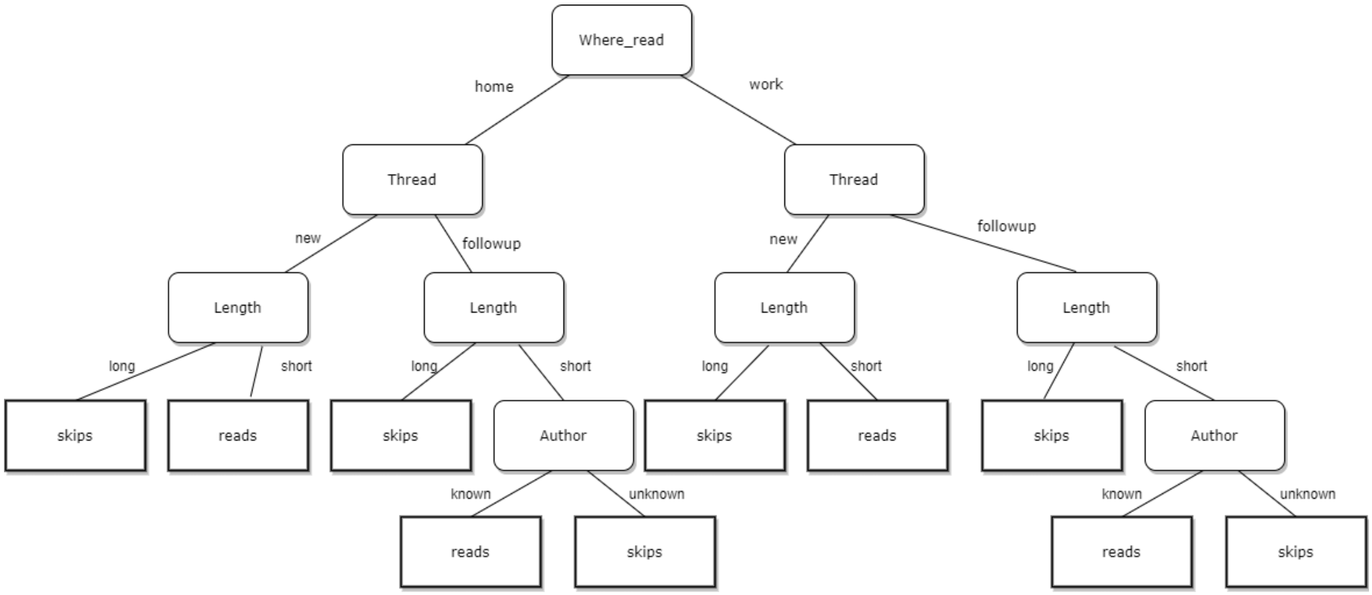
Thus, the learning algorithm in the order $[length, Thread, Author]$ gives us the maximum information gain, and more efficient than function 1 that in the order $[Author, Thread, Length, WhereRead]$.

Besides, we can use example $e19$ $[unknown, new, long, work]$ to test the result for both functions, if we put this example into the first function, $function$ 1 will give us output $User\_action$ = $\boldsymbol{reads}$ **(unknown->new->reads)** while the $function$ 2 will give us result $User\_action$ = $\boldsymbol{skips}$**. (long->skips)**

Therefore, this tree represents a different function since they produce different output in the same situation.

z5261536 Jianjun(Jay) Chen

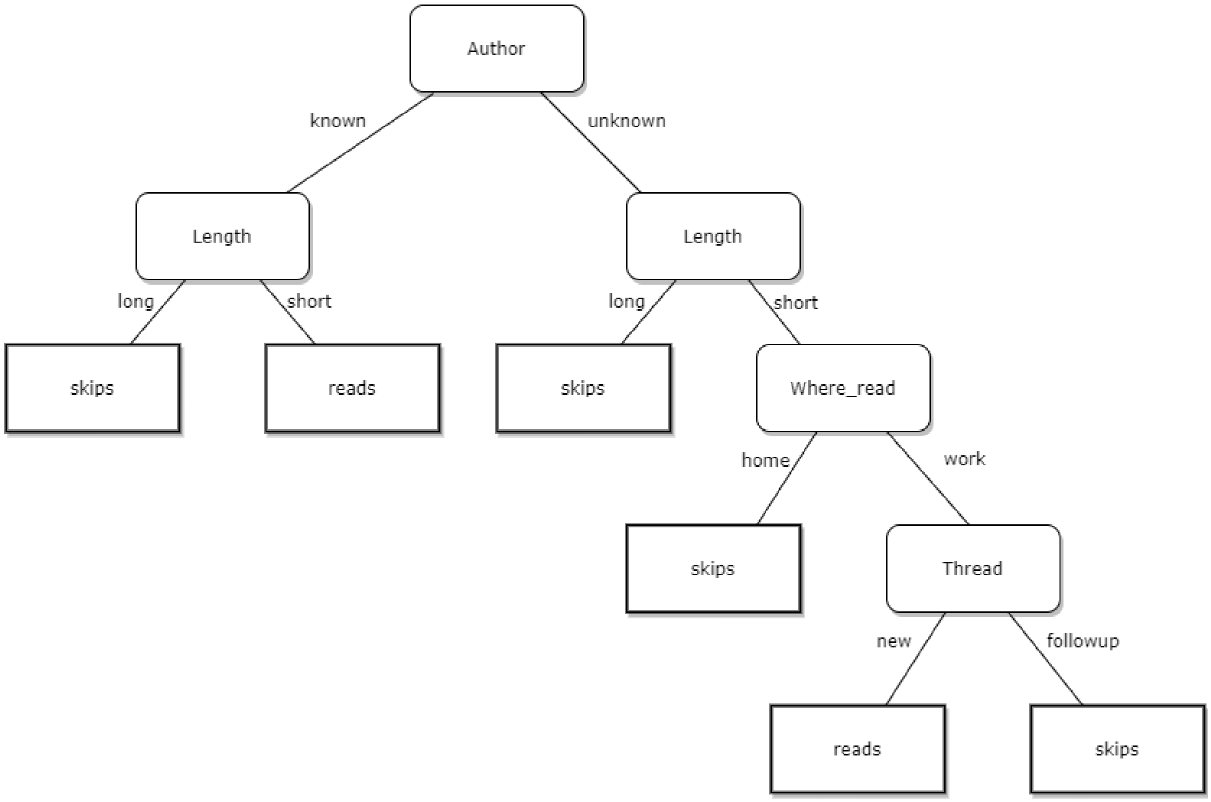
(b) Similarly, the tree fond when the features are in the order $[WhereRead, Thread, Length, Author]$ is: **(function 3)**



This tree represents the same function as that found with the maximum information gain split. Whatever the $where\_read$ is $home$ or $work$ or $Thread$ is new or $followup$, if the $Length = long$, $User\_action = skips$. And if the $Thread = new$ and $Length = short$ or $Author = known$, then $User\_action = \boldsymbol{reads}$, otherwise, $User\_action = \boldsymbol{skips}$. It produces the same output as function 2 for the same condition, hence, it represents the same function as that found with the maximum information gain split.

z5261536 Jianjun(Jay) Chen

(c) Yes, there is. For example, the tree found when the features are in the order $[Author, Length, WhereRead, Thread]$: **(function 4)**



When the case is $[Author\ =\ unknown, Thread = new, Length\ =\ short, Where\_read\ =\ home]$. This function will return $skips$, but all the previous function (1, 2, 3) will return $reads$.

Thus, this tree correctly classifies the training examples but represents a different function than those found by the preceding algorithms.

z5261536 Jianjun(Jay) Chen

Question 1.2

In this question, we need to use the J48 implementation in Weka to create and train the decision trees based on the **Adult data set.** The task is to analyse the data and build a model to predict whether income exceeds $50K/year based on the provided data set.

Firstly, I combined the data from $adult.data$ and $adult.test$ into a data file and change these data into complete data sets format by adding $@relation$, $@attribute$ and $@data$. Then, I convert the data file into the $ARFF$ format.

```
@relation adult

@attribute age numeric
@attribute workclass {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
@attribute fnlwgt numeric
@attribute education {Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th,
@attribute education-num numeric
@attribute marital-status {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
@attribute occupation {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct
@attribute relationship {Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}
@attribute race {White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}
@attribute sex {Female, Male}
@attribute capital-gain numeric
@attribute capital-loss numeric
@attribute hours-per-week numeric
@attribute native-country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, S
@attribute class {>50K, <=50K}


@data
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
```

For training and testing data, I chose the J48 algorithm under the “Tree” subcategory in the classify module, And I used both cross-validation with default value 10-fold and percentage split with default value 66% training data and 34% test data.

Besides, there are a lot of different pruning settings against overfitting such as confidenceFactor, minNumObj, subtreeRaising etc. The confidence factor is used for pruning and the smaller the value incur more pruning, minNumObj is the minimum number of instances per leaf, default value 2, sometimes it’s better not to continue splitting if the nodes get very small. And we can also consider whether to use the subtree raising operation during pruning, it will affect the complexity and speed of the algorithm as well. I changed some of the pruning settings and compared the results in the table below. (Model time, Tree size, Accuracy)

Algorithm	ConfiFactor	MinObjNum	Model time	# of leaves	Tree size	Accuracy
(default)	0.25	2	1.93s	696	911	86.09%
	0.15	2	1.77s	281	378	86.15%
	0.05	2	1.93s	59	68	85.80%
	0.25	5	1.37s	361	475	86.08%
	0.25	10	1.11s	224	297	86.13%
	0.25	15	0.98s	180	238	86.06%
	0.15	5	1.25s	249	324	86.11%
	0.15	10	1.11s	137	186	86.15%
	0.15	15	0.96s	152	201	85.99%
	0.05	5	1.29s	49	74	85.77%
	0.05	10	1.07s	75	110	85.78%
	0.05	15	0.99s	21	36	85.71%
	0.05	20	1s	61	88	85.67%

I first control the value of MinObjNum (default = 2) and change the value of ConfidenceFactor, then I control the ConfidenceFactor (default = 0.25) and change the value MinObjNum. After that, I changed both of them to find the optimal combination for the size of the tree and the accuracy. When the size of the tree is getting smaller and smaller, the accuracy is also decreasing. In the end, I decided to use the tree with ConfidenceFactor = 0.05 and MinObjNum = 15. Although this tree doesn't have the highest accuracy, it has the smallest size with an acceptable accuracy rate.

z5261536 Jianjun(Jay) Chen

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.05 -M 15
Relation:     adult
Instances:    48842
Attributes:   15
              age
              workclass
              fnlwgt
              education
              education-num
              marital-status
              occupation
              relationship
              race
              sex
              capital-gain
              capital-loss
              hours-per-week
              native-country
              class
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===

J48 pruned tree
------------------

capital-gain <= 6849
|   marital-status = Married-civ-spouse
|   |   capital-loss <= 1844
|   |   |   education-num <= 11
|   |   |   |   capital-gain <= 5060: <=50K (13798.0/3783.0)
|   |   |   |   capital-gain > 5060: >50K (116.0/5.0)
|   |   |   education-num > 11
|   |   |   |   hours-per-week <= 30
|   |   |   |   |   sex = Female: >50K (159.0/64.0)
|   |   |   |   |   sex = Male: <=50K (386.0/111.0)
|   |   |   |   hours-per-week > 30
|   |   |   |   |   age <= 33
|   |   |   |   |   |   age <= 25: <=50K (121.0/28.0)
|   |   |   |   |   |   age > 25
|   |   |   |   |   |   |   education-num <= 12: <=50K (129.0/40.0)
|   |   |   |   |   |   |   education-num > 12: >50K (840.0/364.0)
```

```
|   |   |   |   |   age > 33: >50K (4113.0/1305.0)
|   |   capital-loss > 1844
|   |   |   capital-loss <= 1980: >50K (857.0/18.0)
|   |   |   capital-loss > 1980
|   |   |   |   capital-loss <= 2163: <=50K (104.0)
|   |   |   |   capital-loss > 2163
|   |   |   |   |   education-num <= 12
|   |   |   |   |   |   education-num <= 9: <=50K (43.0/11.0)
|   |   |   |   |   |   education-num > 9
|   |   |   |   |   |   |   capital-loss <= 2392: <=50K (16.0/6.0)
|   |   |   |   |   |   |   capital-loss > 2392: >50K (16.0/4.0)
|   |   |   |   |   education-num > 12: >50K (89.0/5.0)
|   marital-status = Divorced: <=50K (6454.0/498.0)
|   marital-status = Never-married: <=50K (15908.0/530.0)
|   marital-status = Separated: <=50K (1505.0/76.0)
|   marital-status = Widowed: <=50K (1485.0/96.0)
|   marital-status = Married-spouse-absent: <=50K (613.0/44.0)
|   marital-status = Married-AF-spouse: <=50K (35.0/12.0)
capital-gain > 6849: >50K (2055.0/28.0)
```

Number of Leaves : 21

Size of the tree : 36

Time taken to build model: 0.99 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	41864	85.7131 %
Incorrectly Classified Instances	6978	14.2869 %
Kappa statistic	0.5706	
Mean absolute error	0.2096	
Root mean squared error	0.3249	
Relative absolute error	57.5685 %	
Root relative squared error	76.1472 %	
Total Number of Instances	48842	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.576	0.054	0.769	0.576	0.659	0.580	0.875	0.751	>50K

0.946 0.424 0.876 0.946 0.910 0.580 0.875 0.943
<=50K
Weighted Avg. 0.857 0.336 0.851 0.857 0.850 0.580 0.875
0.897

=== Confusion Matrix ===

a b <--
classified as6732
4955 |a = >50K
2023 35132 | b = <=50K