RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Tutorial Problems Week 3

1. Assume that 3 fold cross validation is to be used with the following data. Give two possible outcomes for the composition of the the training and test sets.

No	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	7.0	3.2	4.7	1.4	Iris-versicolor
4	6.4	3.2	4.5	1.5	Iris-versicolor
5	6.9	3.1	4.9	1.5	Iris-versicolor
6	6.3	3.3	6.0	2.5	Iris-virginica
7	5.8	2.7	5.1	1.9	Iris-virginica
8	7.1	3.0	5.9	2.1	Iris-virginica

2. The following is the output from a data file with various car attributes and numeric class miles per gallon (mpg). Study the output and answer the following questions.

```
@relation 'autoMpg.names'
@attribute cylinders { 8, 4, 6, 3, 5}
@attribute displacement real
@attribute horsepower real
@attribute weight real
@attribute acceleration real
@attribute model { 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82}
@attribute origin { 1, 3, 2}
@attribute mpg real

=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation: autoMpg.names


M5 pruned model tree:
(using smoothed linear models)

displacement <= 190.5 : LM1 (227/37.092%)
displacement >  190.5 :
|   horsepower <= 127 :
|   |   displacement <= 241 : LM2 (44/14.999%)
|   |   displacement >  241 : LM3 (31/27.34%)
|   horsepower >  127 : LM4 (96/15.987%)

LM num: 1
mpg =
```

```
        0.0009 * displacement
        - 0.0785 * horsepower
        - 0.0066 * weight
        + 0.0966 * model=75,71,76,74,77,78,79,81,82,80
        + 1.5034 * model=71,76,74,77,78,79,81,82,80
        + 2.4334 * model=77,78,79,81,82,80
        + 3.8499 * model=79,81,82,80
        + 0.0982 * model=81,82,80
        + 0.1256 * model=82,80
        + 1.86 * model=80
        + 1.6562 * origin=2,3
        + 44.8206

LM num: 2
mpg =
        -0.0104 * displacement
        + 0.0187 * horsepower
        - 0.0033 * weight
        + 0.1257 * model=75,71,76,74,77,78,79,81,82,80
        + 2.7325 * model=76,74,77,78,79,81,82,80
        - 1.6417 * model=74,77,78,79,81,82,80
        + 0.3584 * model=77,78,79,81,82,80
        + 0.34 * model=78,79,81,82,80
        + 0.1812 * model=79,81,82,80
        + 1.1726 * model=81,82,80
        + 0.1634 * model=82,80
        + 0.2178 * origin=2,3
        + 29.3347

LM num: 3
mpg =
        0.0085 * displacement
        - 0.0075 * horsepower
        - 0.0036 * weight
        + 0.1257 * model=75,71,76,74,77,78,79,81,82,80
        + 0.7804 * model=76,74,77,78,79,81,82,80
        - 0.3793 * model=74,77,78,79,81,82,80
        + 0.3584 * model=77,78,79,81,82,80
        + 2.3842 * model=78,79,81,82,80
        + 0.1812 * model=79,81,82,80
        + 6.9945 * model=81,82,80
        + 0.1634 * model=82,80
        + 0.2178 * origin=2,3
        + 28.3813

LM num: 4
mpg =
        0.0024 * displacement
        - 0.0278 * horsepower
```

```
        - 0.0016 * weight
        - 0.3653 * acceleration
        + 0.6228 * model=75,71,76,74,77,78,79,81,82,80
        + 0.1255 * model=76,74,77,78,79,81,82,80
        + 0.5877 * model=74,77,78,79,81,82,80
        + 1.0532 * model=77,78,79,81,82,80
        + 1.1897 * model=78,79,81,82,80
        + 0.1812 * model=79,81,82,80
        + 0.5364 * model=81,82,80
        + 0.1634 * model=82,80
        + 0.2178 * origin=2,3
        + 28.7128

Number of Rules : 4
```

(a) Look at the attribute 'model'. The values are numeric but it is not coded as real. Why is this?

(b) What about the attribute 'cylinders'?

(c) Sketch the model as a tree.

(d) Explain the procedure for classifying a new instance.

(e) What would be the prediction for the instance cylinders=8, displacement=307, horsepower=130, weight=3504, acceleration=12, model=70, origin=1?

3. Calculate the mean absolute error of the following prediction results

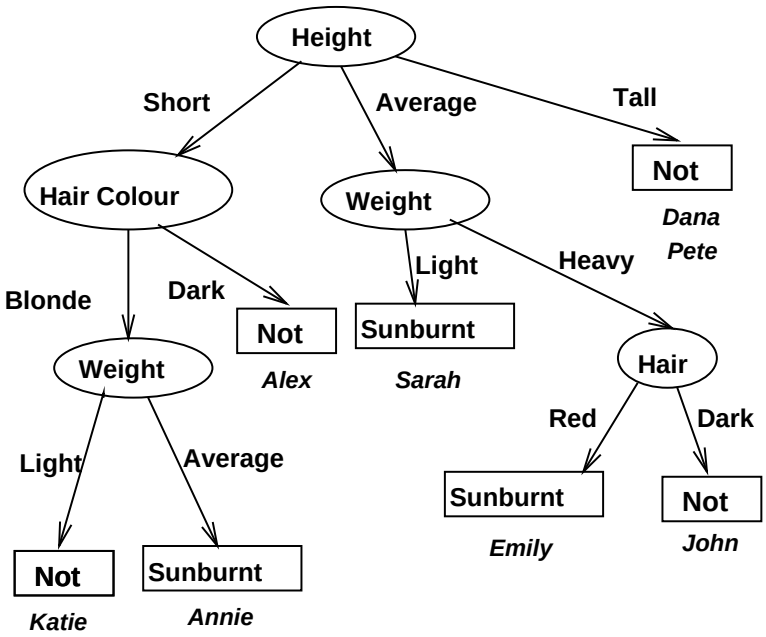
Actual mpg	Predicted mpg
30	25
25	30
30	50
45	40

4. Build the confusion matrix from the following classification results:

Actual	Predicted
Benign	Benign
Benign	Benign
Benign	Malignant
Malignant	Benign
Benign	Malignant

5. Explain why it is not possible to get a confusion matrix for the data of question 3.

6. Convert the following decision tree to 'line-printer form'.



7. Draw the decision tree on the next page in tree form.

8. How will these examples be classified by the following tree:

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
a	47.75	8	u	g	c	v	7.875	t	t	06	t	g	000	1260
a	27.42	14.5	u	g	x	h	3.085	t	t	01	f	g	120	11
a	27.42	3.00	t	g	x	h	3.085	t	t	-1	f	g	320	200

Which combinations of attribute values will result in the decision being made at the line labelled THIS LINE?

```
'A9' = 't':
|   'A15' > 400 : '+' (131.0/5.0)
|   'A15' <= 400 :
|   |   'A3' <= 0.29 : '+' (11.0/1.3)
|   |   'A3' > 0.29 :
|   |   |   'A11' > 0 : '+' (114.0/22.4)
|   |   |   'A11' <= 0 :
|   |   |   |   'A3' <= 2.335 :
|   |   |   |   |   'A2' > 53.33 : '+' (2.1/1.1)
|   |   |   |   |   'A2' <= 53.33 :
|   |   |   |   |   |   'A14' <= 230 : '-' (18.0/2.5)
|   |   |   |   |   |   'A14' > 230 :
|   |   |   |   |   |   |   'A15' <= 12 : '-' (16.9/6.9)
|   |   |   |   |   |   |   'A15' > 12 : '+' (3.0/1.1)
|   |   |   |   'A3' > 2.335 :
|   |   |   |   |   'A14' > 395 : '-' (5.1/1.2)
|   |   |   |   |   'A14' <= 395 :
|   |   |   |   |   |   'A14' > 320 : '+' (5.1/1.3)
|   |   |   |   |   |   'A14' <= 320 :
|   |   |   |   |   |   |   'A14' > 250 : '-' (5.1/2.3)
|   |   |   |   |   |   |   'A14' <= 250 :
|   |   |   |   |   |   |   |   'A4' = 'u': '+' (35.8/11.2)
|   |   |   |   |   |   |   |   'A4' = 'y': '-' (14.0/7.8)
|   |   |   |   |   |   |   |   'A4' = 'l': '+' (0.0)
|   |   |   |   |   |   |   |   'A4' = 't': '+' (0.0)
'A9' = 'f':
|   'A3' > 0.165 : '-' (298.0/15.1)
|   'A3' <= 0.165 :
|   |   'A7' = 'h': '-' (0.0)
|   |   'A7' = 'bb': '+' (1.2/0.9)
|   |   'A7' = 'j': '+' (1.2/0.9)
|   |   'A7' = 'n': '+' (1.2/0.9)                *******THIS LINE****
|   |   'A7' = 'z': '-' (0.0)
|   |   'A7' = 'dd': '-' (0.0)
|   |   'A7' = 'ff': '-' (5.0/1.9)
|   |   'A7' = 'o': '-' (0.0)
|   |   'A7' = 'v':
|   |   |   'A2' <= 35.58 : '-' (18.7/5.3)
|   |   |   'A2' > 35.58 : '+' (3.6/1.3)
```

9. Suppose that you are to deal with the situation described in the following table, which gives the history of eight past production runs in a factory:

Run	Supervisor	Operator	Machine	Overtime	output
1	Patrick	Joe	a	no	high
2	Patrick	Sam	b	yes	low
3	Thomas	Jim	b	no	low
4	Patrick	Jim	b	no	high
5	Sally	Joe	c	no	high
6	Thomas	Sam	c	no	low
7	Thomas	Joe	c	no	low
8	Patrick	Jim	a	yes	low

   (a) Construct a decision tree from this table. When it is necessary to choose an attribute for splitting use the order Supervisor, Operator, Machine, Overtime.

   (b) Construct a decision tree from this table using the basic decision tree algorithm with disorder to determine the factors that influence output.

   (c) What would you control to ensure high output: supervisor, operator, machine or overtime?

   (d) Repeat using information gain.

   (e) Repeat using information ratio.

10. [Optional] Apply the J48 algorithm to this data, then compare your result to a run of the J48 program on this data.

	Age	Prescription	Astigmatic	Tears	Lenses
1	8	myope	no	reduced	NO
2	12	myope	yes	normal	HARD
3	10	hypermetrope	no	normal	SOFT
4	14	hypermetrope	yes	reduced	NO
5	15	myope	no	normal	SOFT
6	24	myope	yes	reduced	NO
7	35	hypermetrope	yes	reduced	NO
8	55	hypermetrope	yes	normal	NO
9	56	myope	no	normal	NO
10	60	myope	yes	normal	HARD
11	65	hypermetrope	no	reduced	NO
12	72	hypermetrope	no	normal	SOFT

11. [Optional] A data set of 100 examples has 10 attributes with 5 different values each.

   (a) How many possible decision trees are there for this data set?

   (b) Estimate how long it would take to exhaustively search for the best tree.

   (c) Suppose we restrict attention to 2 level trees only. How long would it take?

12. [Optional] What is the computational complexity of the j48 algorithm?