COMP4433 Data Mining and Data Warehousing

FAQ on Classification 1 (with reference answers)

1. Given the following table.

Parcel ID	*Origin*	*Destination*	*Type*	*Weight*
1	HK	HK	Parcel	Light
2	Kln	Kln	Letter	Light
3	NT	Kln	Letter	Light
4	HK	HK	Parcel	Heavy
5	Kln	Kln	Parcel	Light
6	NT	NT	Letter	Light
7	HK	HK	Letter	Light
8	Kln	Kln	Parcel	Heavy
9	Kln	Kln	Letter	Light
10	HK	HK	Letter	Light
11	HK	HK	Parcel	Heavy
12	Kln	Kln	Letter	Light
13	HK	HK	Letter	Light
14	Kln	Kln	Parcel	Light
15	HK	NT	Parcel	Heavy
16	NT	Kln	Letter	Light
17	HK	NT	Letter	Light
18	Kln	HK	Parcel	Light
19	HK	NT	Parcel	Heavy
20	HK	HK	Parcel	Light
21	Kln	Kln	Letter	Light
22	Kln	HK	Parcel	Heavy
23	Kln	Kln	Letter	Light
24	Kln	Kln	Letter	Light
25	HK	HK	Parcel	Light

Construct a decision tree, based on information gain, to classify the type of courier services (cf. column *Type*). You may assume that the first 20 records are available for model construction and the remaining 5 records are used to validate your answer.

Ans.

1. Determining the root attribute:

I(p,n)=(10,10)=1

Entropy for Origin

Origin	p_i	n_i	$I(p_i , n_i)$
HK	6	4	0.97
Kln	4	3	0.985
NT	0	3	0

Entropy=(10/20)*I(6,4) + (7/20)*I(4,3) + (3/20)*I(0,3)=0.83

Information_Gain(Origin)=1-0.83=0.17

Entropy for Destination

Dest.	p_i	n_i	$I(p_i , n_i)$
HK	5	3	0.954
Kln	3	5	0.954
NT	2	2	1

Information_Gain(Dest.)=1-(8/20)*0.954-(8/20)*0.954-(4/20)*1=0.0368

Entropy for Weight

Weight	p_i	N_i	$I(p_i , n_i)$
Light	5	10	0.918
Heavy	5	0	0

Information_Gain(Weight)=1-(15/20)*0.918-(5/20)*0=0.312

Hence, *Weight* is selected as the decision attribute for the root node. The above steps will be repeated to build the sub-trees.

---

2. Determining the internal node attributes:

Since we have two branches from the root and one (Weight=Heavy) can be terminated, there will only have one sub-tree under the root. Then, it is needed to determine the node attribute when (Weight=Light):
I(p,n)=(5,10)=0.918

Entropy for Origin

Origin	p_i	n_i	$I(p_i , n_i)$
HK	2	4	0.918
Kln	3	3	1
NT	0	3	0

Entropy=(6/15)*I(2,4) + (6/15)*I(3,3) + (3/15)*I(0,3)=0.767

Information_Gain(Origin)=0.918-0.767=0.151

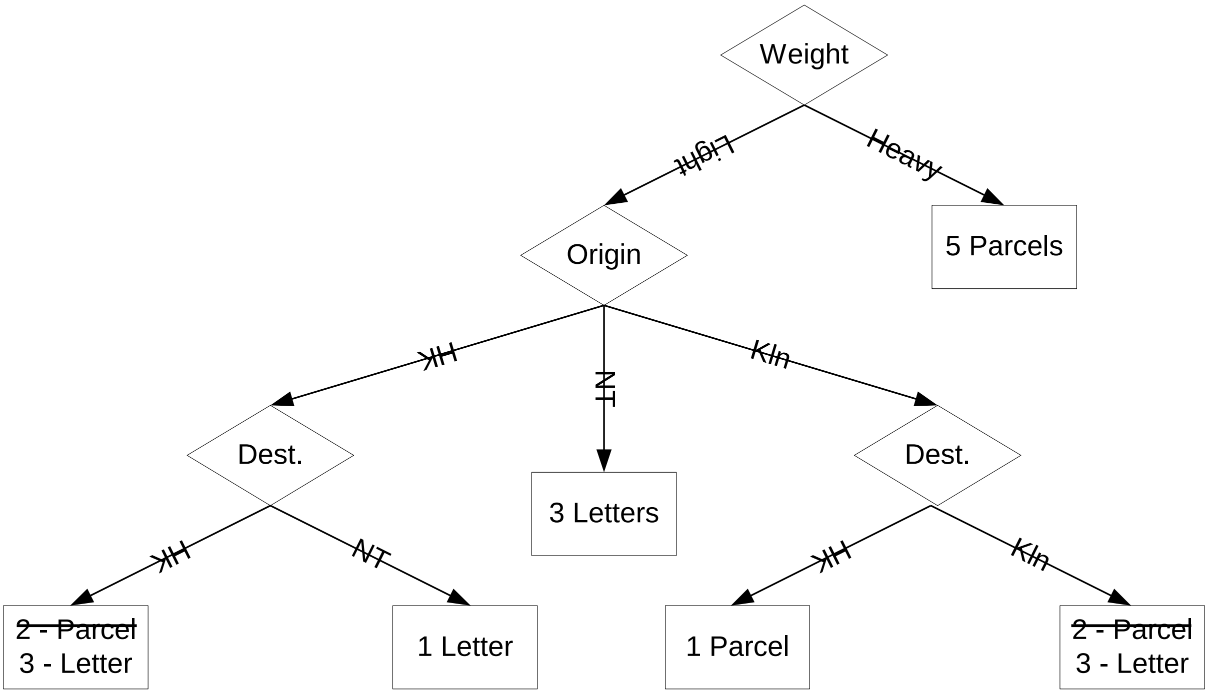
Entropy for Destination

Dest.	p_i	n_i	$I(p_i, n_i)$
HK	3	3	1
Kln	2	5	0.863
NT	0	2	0

Entropy=(6/15)*I(3,3) + (7/15)*I(2,5) + (2/15)*I(0,2)=0.803

Information_Gain(Dest.)= 0.918-0.803=0.115

This time, *Origin* is selected as the decision attribute for this node. The final decision tree can then be built by taking the last available attribute into considerations and it is shown below.



Except the last record, i.e., parcel ID 25, all testing records can be classified correctly. The classification rate on the testing data is then equal to 80%.

2. You are working for the FRIDAY telecom company and are given some customer records. Your manager asks you to find the classification rule(s) for high and low usage customers. The data are given below.

Customer ID	Monthly Income	Age	Education	Marital Status	Usage
9100123	Low	Old	University	Married	Low
9303034	High	Young	College	Single	High
9210126	Medium	Young	College	Married	High
9142020	Medium	Old	High School	Single	Low
9910111	High	Old	University	Single	High
9576732	Low	Old	High School	Married	Low

a) Suppose you take use of the decision tree to solve the problem. What are the (theoretical) maximum and minimum depths of the tree being formed?

Ans.

For a database consisting of four feature attributes, the theoretical maximum and minimum depths of the tree being formed are 4 and 0 respectively.

b) Construct a decision tree, based on information gain, to classify customers as "high usage" and "low usage". Show your steps.

Ans.

I(p,n)=I(3,3)=1

M.Income	I(pi,ni)
Low	0
Medium	1
High	0

Age	I(pi,ni)
Old	0.81
Young	0

Education	I(pi,ni)
University	1
College	0
High School	0

Mari. Status	I(pi,ni)
Single	0.92
Married	0.92

Information_Gain(Monthly Income)=1-0.33=0.67
Information_Gain(Age)=1-0.67*0.81=0.46
Information_Gain(Education)=1-0.33=0.67
Information_Gain(Marital Status)=1-0.92=0.08

Hence, either *Monthly Income* or *Education* can be selected as the decision attribute for the root node.
Let the *Monthly Income* be selected. The *high* and *low* branches will terminate with high usage and low usage respectively. For the *medium* branch, any of the *Age*, *Education*, and *Marital Status* can be selected as the decision attribute for the second level.
On the other hand, let the *Education* be selected. The *College* and *High School* branches will terminate with high usage and low usage respectively. For the *University* branch, any of the *Monthly Income* and *Marital Status* (but not *Age*) can be selected as the decision attribute for the second level.

c) Extract the classification rules from the decision tree constructed in part (b).

Ans.

One solution is:
IF *Monthly Income* is *High* THEN *Usage* is *High*
IF *Monthly Income* is *Low* THEN *Usage* is *Low*
IF *Monthly Income* is *Medium* AND *Age* is *Young* THEN *Usage* is *High*
IF *Monthly Income* is *Medium* AND *Age* is *Old* THEN *Usage* is *Low*

Another solution is:
IF *Monthly Income* is *High* THEN *Usage* is *High*
IF *Monthly Income* is *Low* THEN *Usage* is *Low*
IF *Monthly Income* is *Medium* AND *Marital Status* is *Married* THEN *Usage* is *High*
IF *Monthly Income* is *Medium* AND *Marital Status* is *Single* THEN *Usage* is *Low*

d) Based on the results of parts (b) & (c), classify the following customer record.

Customer ID	Monthly Income	Age	Education	Marital Status
9100100	Medium	Unknown	University	Married

Ans.

According to the rule:
IF *Monthly Income* is *Medium* AND *Marital Status* is *Married* THEN *Usage* is *High*
The customer should be classified as high usage one.

However, according to another set of rules:
IF *Education* is *College* THEN *Usage* is *High*
IF *Education* is *High School* THEN *Usage* is *Low*
IF *Education* is *University* AND *Marital Status* is *Married* THEN *Usage* is *Low*
IF *Education* is *University* AND *Marital Status* is *Single* THEN *Usage* is *High*
The customer should be classified as low usage one, instead.

Both are acceptable, according to the selected rule from the selected decision tree. A more complicated solution should take all possible classification results into consideration and then apply the majority voting to determine the final solution. In addition, the "importance" of the classification rule can also be taken into consideration. For example, if the rule
IF *Monthly Income* is *Medium* AND *Marital Status* is *Married* THEN *Usage* is *High*
is responsible for 3 training database records while the rule
IF *Education* is *University* AND *Marital Status* is *Married* THEN *Usage* is *Low*
is responsible for 2 training database records, we should classify the customer as high usage.

3. Given Table 1 below showing Hang Seng Bank's daily closing price and trend information (which is simply labelled as "Up" if today's closing price is higher than the previous trading day's closing price, "Level" if today's closing price is the same as the previous trading day's closing price and "Down" if today's closing price is lower than the previous trading day's closing price), construct a regression tree (decision tree for regression) based on the extracted data in Table 2 to predict the closing prices of 3/1/2007, 4/1/2007 and 5/1/2007. Compute the mean absolute deviation (MAD) of these three trading days' prediction. Show your final regression tree.

Table 1: Hang Seng Bank's Stock Information

Date	Closing Price	Trend
1/12/2006	103.5	---
4/12/2006	103.8	Up
5/12/2006	104.1	Up
6/12/2006	104.2	Up
7/12/2006	104.1	Down
8/12/2006	104.3	Up
11/12/2006	104.5	Up
12/12/2006	104.5	Level
13/12/2006	104.5	Level
14/12/2006	105	Up
15/12/2006	105.1	Up
18/12/2006	104.3	Down
19/12/2006	104.6	Up
20/12/2006	105.1	Up
21/12/2006	105.4	Up
22/12/2006	105.6	Up
27/12/2006	105.9	Up
28/12/2006	106	Up
29/12/2006	106.3	Up
2/1/2007	106.3	Level
3/1/2007	106.9	Up
4/1/2007	106.7	Down
5/1/2007	106.9	Up

Table 2. Feature Engineered Hang Seng Bank Price Information

Date	Trend				
	Today-3	Today-2	Today-1	Today	Tomorrow's Price
7/12/2006	Up	Up	Up	Down	104.3
8/12/2006	Up	Up	Down	Up	104.5
11/12/2006	Up	Down	Up	Up	104.5
12/12/2006	Down	Up	Up	Level	104.5
13/12/2006	Up	Up	Level	Level	105
14/12/2006	Up	Level	Level	Up	105.1
15/12/2006	Level	Level	Up	Up	104.3
18/12/2006	Level	Up	Up	Down	104.6
19/12/2006	Up	Up	Down	Up	105.1
20/12/2006	Up	Down	Up	Up	105.4
21/12/2006	Down	Up	Up	Up	105.6
22/12/2006	Up	Up	Up	Up	105.9
27/12/2006	Up	Up	Up	Up	106
28/12/2006	Up	Up	Up	Up	106.3
29/12/2006	Up	Up	Up	Up	106.3

Answer

Standard deviation for one attribute (Tomorrow's Price, TP).

TP
104.3
104.6
104.5
105
104.5
104.5
105.1
104.3
105.1
105.4
105.6
105.9
106
106.3
106.3

COUNT	15
AVERAGE	105.16
SD	0.694550214
CV	0.66%

Standard deviation for two attributes (target and predictor).

Consider "Today" (T)

		SD	COUNT
T	UP	0.69187	11
	LEVEL	0.25	2
	DOWN	0.15	2

S(TP,T)	0.5607
SDR(TP,T)	0.13385

SDR: Standard Deviation Reduction

Consider "Today – 1" (T-1)

		SD	COUNT
T-1	UP	0.7797	11
	LEVEL	0.05	2
	DOWN	0.3	2

S(TP,T-1)	0.61845
SDR(TP,T-1)	0.0761

Consider "Today – 2" (T-2)

		SD	COUNT
T-2	UP	0.72841	11
	LEVEL	0.4	2
	DOWN	0.45	2

S(TP,T-2)	0.6475
SDR(TP,T-2)	0.04705

Consider “Today – 3” (T-3)

		SD	COUNT
T-3	UP	0.69473	11
	LEVEL	0.15	2
	DOWN	0.55	2

S(TP,T-3)	0.6028
SDR(TP,T-3)	0.09175

The attribute “Today” (T) has the largest SDR and is chosen to be the “decision node” at root.

For branches “Level” and “Down”, each of them has only two instances (even smaller than the number of label of that attribute, so we simply stop the splitting and form the leaf nodes with the average amounts directly.

```
Today = Down:  104.45 (mean of 104.3 and 104.6)
    |
    | = Level: 104.75 (mean of 104.5 and 105)
    |
    | = Up     (further splitting)
```

Further split the branch “Up” and we have the following.

For branch “Up” of “Today”

Consider T-1

		SD	COUNT
T-1	UP	0.71927	8
	LEVEL	0	1
	DOWN	0.3	2

S(TP,T-1)	0.57765
SDR(TP,T-1)	0.11422

Consider T-2

		SD	COUNT
T-2	UP	0.61578	7
	LEVEL	0.4	2
	DOWN	0.45	2

S(TP,T-2)	0.5464
SDR(TP,T-2)	0.14546

Consider T-3

		SD	COUNT
T-3	UP	0.66685	9
	LEVEL	0	1
	DOWN	0	1

S(TP,T-3)	0.54561
SDR(TP,T-3)	0.14626

"T-3" has the largest SDR and is assigned to the new node.

```
Today = Down:  104.45
Today = Level: 104.75
Today = Up
    |   T-3 = Down:  105.6 (only 1 sample)
    |   T-3 = Level: 104.3 (only 1 sample)
    |   T-3 = Up (further splitting)
```

Further split the branch "Up" and we have the following.

For branch "Up" of "Today – 3"

Consider "T-1"

		SD	COUNT
T-1	UP	0.62893	6
	LEVEL	0	1
	DOWN	0.3	2

S(TP,T-1)	0.48595
SDR(TP,T-1)	0.1809

Consider "T-2"

		SD	COUNT
T-2	UP	0.66437	6
	LEVEL	0	1
	DOWN	0.45	2

S(TP,T-2)	0.54291
SDR(TP,T-2)	0.12394

Attribute "T-1" is chosen.

```
Today = Down:  104.45
Today = Level: 104.75
```

```
Today = Up
     |    T-3 = Down:  105.6
     |    T-3 = Level: 104.3
     |    T-3 = Up
            |    T-1 = Down:  104.8
            |    T-1 = Level: 105.1
            |    T-1 = Up (further splitting)
```

For branch "Up" of "T-1", there is only one attribute left which is "T-2". We simply use it as the new node. Finally, the decision tree may look like the one below.

Final Decision Tree for Regression

```
Today = Down:  104.45
Today = Level: 104.75
Today = Up
     |    T-3 = Down:  105.6
     |    T-3 = Level: 104.3
     |    T-3 = Up
            |    T-1 = Down:  104.8
            |    T-1 = Level: 105.1
            |    T-1 = Up
                   |    T-2 = Down:  104.95
                   |    T-2 = Level: N/A
                   |    T-2 = Up:    106.125
```

A better fitted (also potentially overfitted) tree could take into considerations of non-zero SD to further split the samples.

For prediction/regression, a simple approach can be just making use of the mean value of the leaf node that the testing/unseen sample is destined. For 2/1/2007, Today=Level and hence the predicted tomorrow's price value is 104.75 (see the tree above). For 4/1/2007, Today=Down and hence the predicted tomorrow's price value is 104.45. If you have further split this leaf node into two for the two samples (7/12/2006 and 18/12/2006), this unseen sample should be predicted by the 7/12/2006 sample, i.e., tomorrow's price value is 104.3. For 3/1/2007, Today=Up, T-3=Up, T-1=Level, hence tomorrow's price value is 105.1.

Date	**T-3**	**T-2**	**T-1**	**T**	**TP**	**Prediction**
2/1/2007	Up	Up	Up	Level	106.9	104.75
3/1/2007	Up	Up	Level	Up	106.7	105.1
4/1/2007	Up	Level	Up	Down	106.9	104.3 (or 104.45)

So, the MAD value for these three days of prediction is

MAD = (|106.9-104.75|+|106.7-105.1|+|106.9-104.3|)/3= 0.21167