

COMP4433 Data Mining and Data Warehousing

FAQ on Classification 1

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.

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

- 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?
- Construct a decision tree, based on information gain, to classify customers as “high usage” and “low usage”. Show your steps.
- Extract the classification rules from the decision tree constructed in part (b).
- 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

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' predictions. 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				Tomorrow's Price
	Today-3	Today-2	Today-1	Today	
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