Supplementary Notes #02

Data Mining and Data Warehousing

Solutions to exercises on Classification

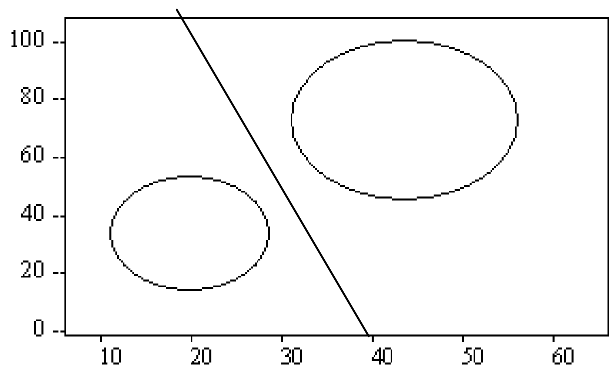
1) No need to normalize the attribute, because they are of the same type and measured on same scale.

Customer No.	Normalized Salary	Normalized Age	Approved? Yes/No	Loan Amount ('000)	Normalized Distance with Applicant
1231	0.5	0.2	Y	12	0.32
1448	0.7	0.5	Y	15	0.22
4567	0.9	0.7	N	0	0.41
7659	1.0	0.5	Y	44	0.28
5355	0.8	0.4	N	0	0.1
8800	0.7	0.4	Y	31	0.14

The decisions of the 5 nearest neighbors are {S, B, B, B, B}. Therefore, the decision should be "Buy"

Expected return = (44 + 0 + 31 +15 + 12) / 5 = 20.4

2) A linear line can separate two classes. You can find the equation of the line if you wish to.



3)

Let IL = income level, PM = payment method, FC = frequency of call, LP = any late payment and CR = credit rating

The sample X we wish to classify,
X = (IL = low, PM = cheque, FC = frequent, LP = yes)

We need to maximize $P(X \mid C_i)P(C_i)$, for $i = 1,2$. $P(C_i)$,the prior probability of each class, can be computed based on the training samples:

P(CR = Good) = 6/10 = 0.6
P(CR = Bad) = 4/10 = 0.4

To compute $P(X \mid C_i)$, for $i = 1,2$, we compute the following conditional probabilities:

P(IL = low | CR = Good) = 1/6
P(IL = low | CR = Bad) = 3/4
P(PM = cheque | CR = Good) = 2/6
P(PM = cheque | CR = Bad) = 2/4
P(FC = frequent | CR = Good) = 2/6
P(FC = frequent | CR = Bad) = 3/4
P(LP = yes | CR = Good) = 2/6
P(LP = yes | CR = Bad) = 3/4

Using the above probabilities, we obtain:

P(X | CR = Good) P(CR = Good) = $(\frac{1}{6} \times \frac{2}{6} \times \frac{2}{6} \times \frac{2}{6})\frac{6}{10}$ = 0.0037

P(X | CR = Bad) P(CR= Bad) = $(\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} \times \frac{3}{4})\frac{4}{10}$ = 0.084375

Therefore the naive Bayesian classifier predicts "CR = Bad" for sample X.

4)

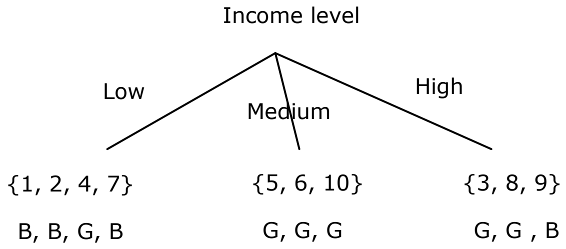
$$M(\text{"}CR\text{"}) = \left[-\frac{4}{10}\log_2\frac{4}{10}\right] + \left[-\frac{6}{10}\log_2\frac{6}{10}\right]$$

$$= 0.5288 + 0.4422$$

$$= 0.971$$

Consider the splitting according to the “Income level”

	Bad	Good	I(B,G)
Low	3	1	0.811
Medium	0	3	0
High	1	2	0.918



$$E(\text{"}IL\text{"}) = \frac{4}{10}(0.811) + \frac{3}{10}(0) + \frac{3}{10}(0.918)$$

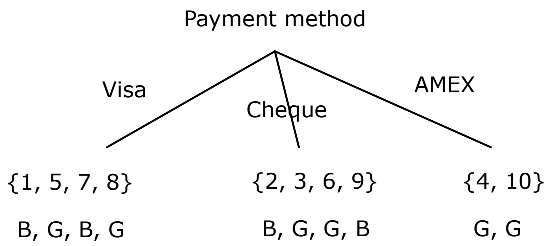
$$= 0.3244 + 0 + 0.2754$$

$$= 0.5998$$

$$Gain(\text{"}IL\text{"}) = 0.971 - 0.5998 = 0.3712$$

Consider the splitting according to the “Payment method”

	Bad	Good	I(B,G)
Visa	2	2	1
Cheque	2	2	1
AMEX	0	2	0



$$E(\text{"}PM\text{"}) = \frac{4}{10}(1) + \frac{4}{10}(1) + \frac{2}{10}(0)$$

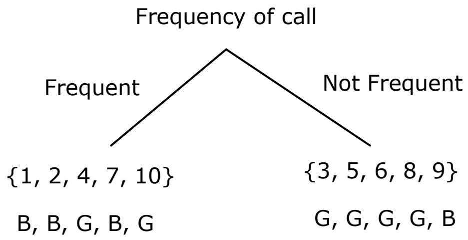
$$= 0.4 + 0.4 + 0$$

$$= 0.8$$

$$Gain(\text{"}PM\text{"}) = 0.971 - 0.8 = 0.171$$

Consider the splitting according to the “Frequency of call”

	Bad	Good	I(B,G)
Frequent	3	2	0.971
Not Frequent	1	4	0.722

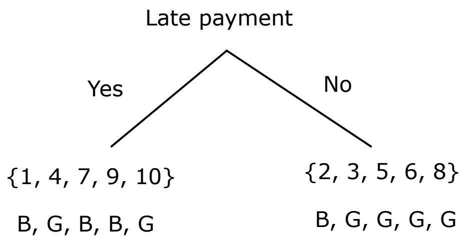


$$E(\text{"}FC\text{"}) = \frac{5}{10}(0.971) + \frac{5}{10}(0.722)$$
$$= 0.4855 + 0.361$$
$$= 0.8465$$

$$Gain(\text{"}FC\text{"}) = 0.971 - 0.8465 = 0.1245$$

Consider the splitting according to the “Late payment”

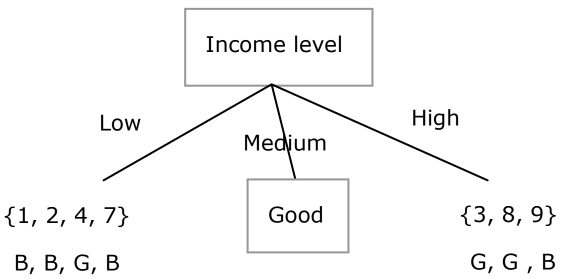
	Bad	Good	I(B,G)
Yes	3	2	0.971
No	1	4	0.722



$$E(\text{"}LP\text{"}) = \frac{5}{10}(0.971) + \frac{5}{10}(0.722)$$
$$= 0.4855 + 0.361$$
$$= 0.8465$$

$$Gain(\text{"}LP\text{"}) = 0.971 - 0.8465 = 0.1245$$

Among four information gain values, the “Income level” has the largest value. Therefore, select “Income level” as the root of the decision tree.

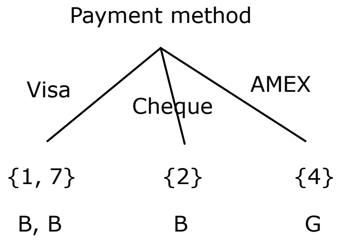


Consider the "Low" branch of the root node

$$M(\text{"}CR\text{"}) = \left[-\frac{1}{4}\log_2\frac{1}{4}\right] + \left[-\frac{3}{4}\log_2\frac{3}{4}\right]$$
$$= 0.5 + 0.3113$$
$$= 0.8113$$

Consider the splitting according to the "Payment method"

	Bad	Good	I(B,G)
Visa	2	0	0
Cheque	1	0	0
AMEX	0	1	0

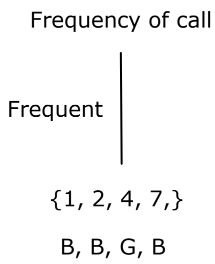


$$E(\text{"}PM\text{"}) = \frac{2}{4}(0) + \frac{1}{4}(0) + \frac{1}{4}(0) = 0$$

$$Gain(\text{"}PM\text{"}) = 0.8113 - 0 = 0.8113$$

Consider the splitting according to the "Frequency of call"

	Bad	Good	I(B,G)
Frequent	3	1	0.811
Not Frequent	0	0	0

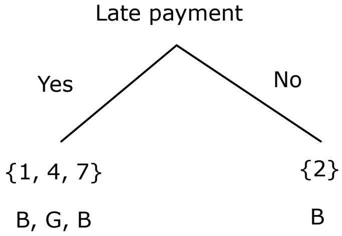


$$E(\text{"}FC\text{"}) = \frac{4}{4}(0.811) + \frac{0}{4}(0) = 0.811$$

$$Gain(\text{"}FC\text{"}) = 0.811 - 0.811 = 0$$

Consider the splitting according to the "Late payment"

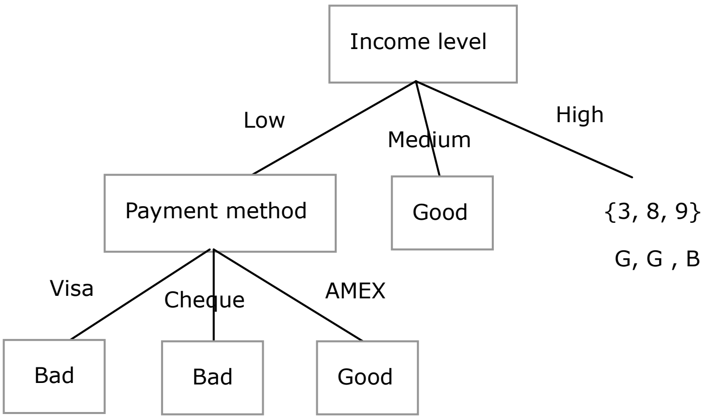
	Bad	Good	I(B,G)
Yes	2	1	0.918
No	1	0	0



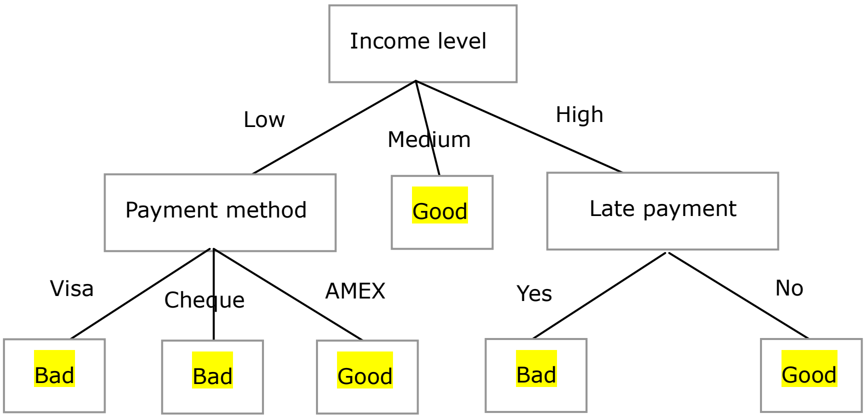
$$E(\text{"}LP\text{"}) = \frac{3}{4}(0.918) + \frac{1}{4}(0) = 0.6885$$

$$Gain("LP") = 0.811 - 0.6885 = 0.1225$$

Among three information gain values, the “Payment method” has the largest value. Therefore, select “Payment method” as the node in the “Low” branch.



Consider the remaining entries, the final decision tree should be:



According to the decision tree above, the sample X is predicted as “Bad”.