Chapter 6

Supervised machine learning for categorical data

Professor Jung Jin Lee
Soongsil University, Korea
New Uzbekistan University, Uzbekistan

Chapter 6 Supervised machine learning for categorical data

6.1 Basic concept of supervised machine learning and classification model

6.2 Decision tree model

6.3 Naive Bayes classification model

6.4 Evaluation and comparison of classification model

6.1 Basic concept of supervised machine learning and classification model

- Classification:
  - Predicts categorical class labels
  - Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data

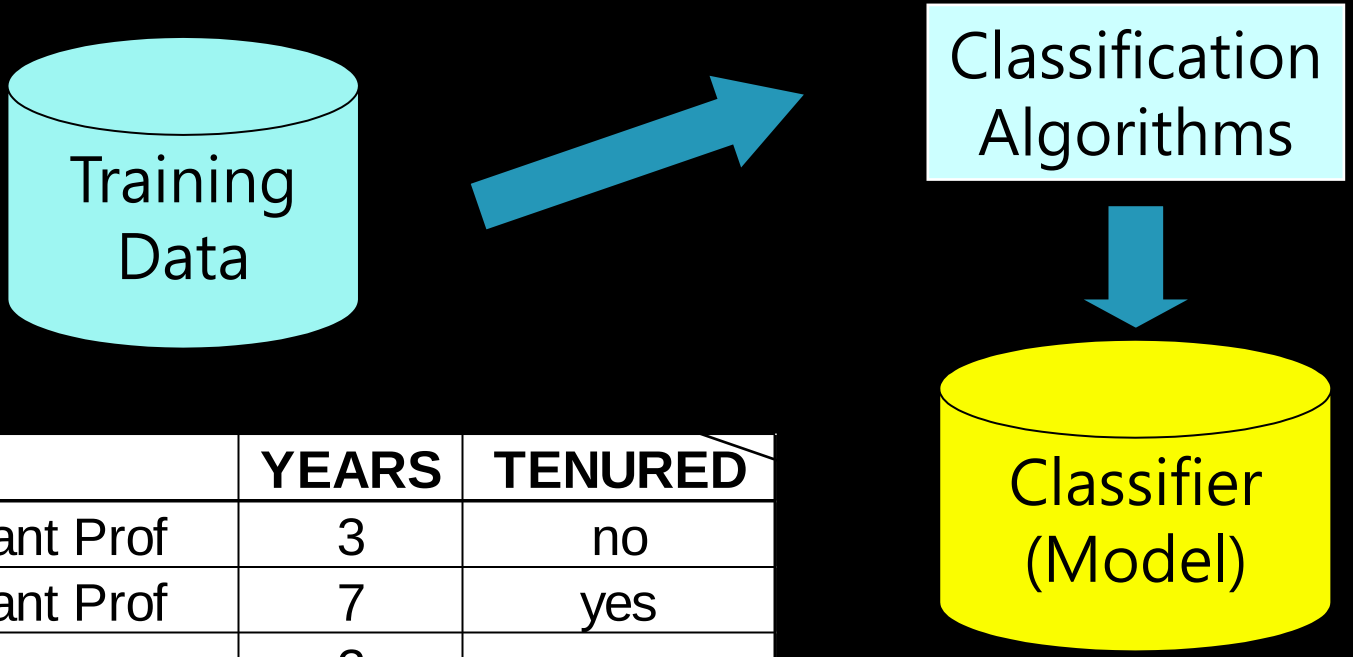
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

6.1 Basic concept of supervised machine learning and classification model

- Model construction: describing a set of predetermined classes
  - Each data is assumed to belong to a predefined class
  - The set of sample data used for model construction: training set
  - The model is represented as classification rules, decision trees, or mathematical formulae

- Model usage: for classifying unknown objects
  - Estimate the accuracy of the model
    - The known class label of a test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that the model correctly classifies
    - Test set is independent of training set, otherwise over-fitting will occur

6.1 Basic concept of supervised machine learning and classification model



Training Data

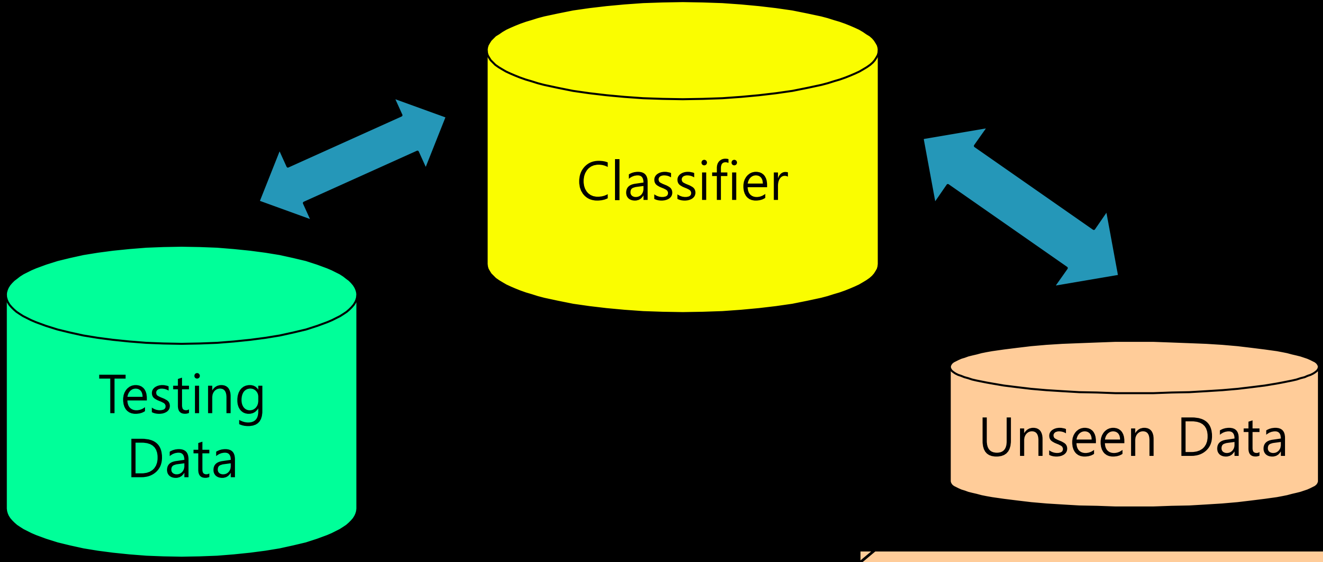
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

6.1 Basic concept of supervised machine learning and classification model



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

Tenured?

↓

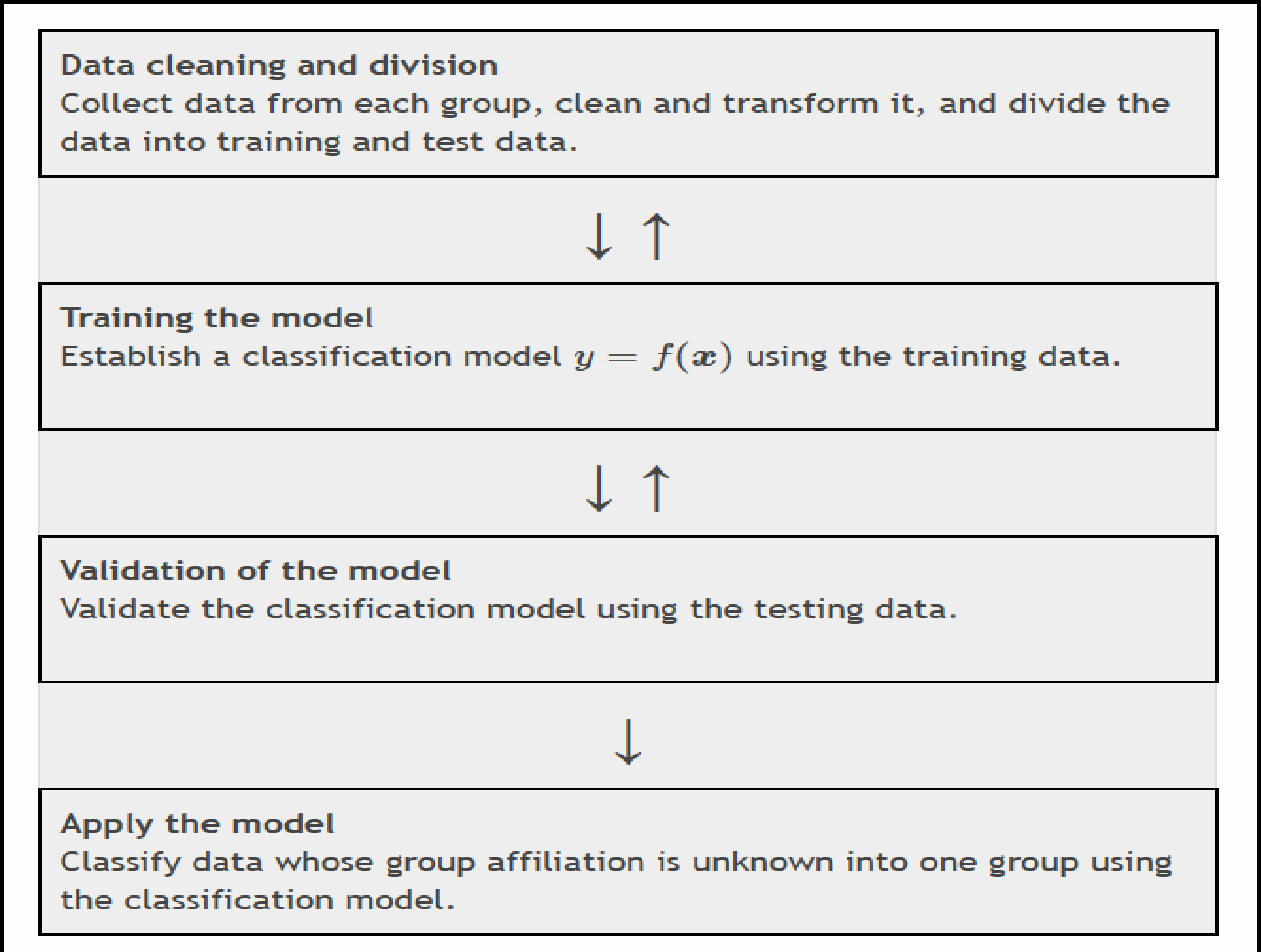
Yes

6.1 Basic concept of supervised machine learning and classification model

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

6.1 Basic concept of supervised machine learning and classification model

❖ General process of classification



Data cleaning and division
Collect data from each group, clean and transform it, and divide the data into training and test data.

↓ ↑

Training the model
Establish a classification model $y = f(x)$ using the training data.

↓ ↑

Validation of the model
Validate the classification model using the testing data.

↓

Apply the model
Classify data whose group affiliation is unknown into one group using the classification model.

6.1 Basic concept of supervised machine learning and classification model

❖ General process of classification

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
- Data transformation
  - Generalize and/or normalize data

6.1 Basic concept of supervised machine learning and classification model

❖ Evaluating classification model

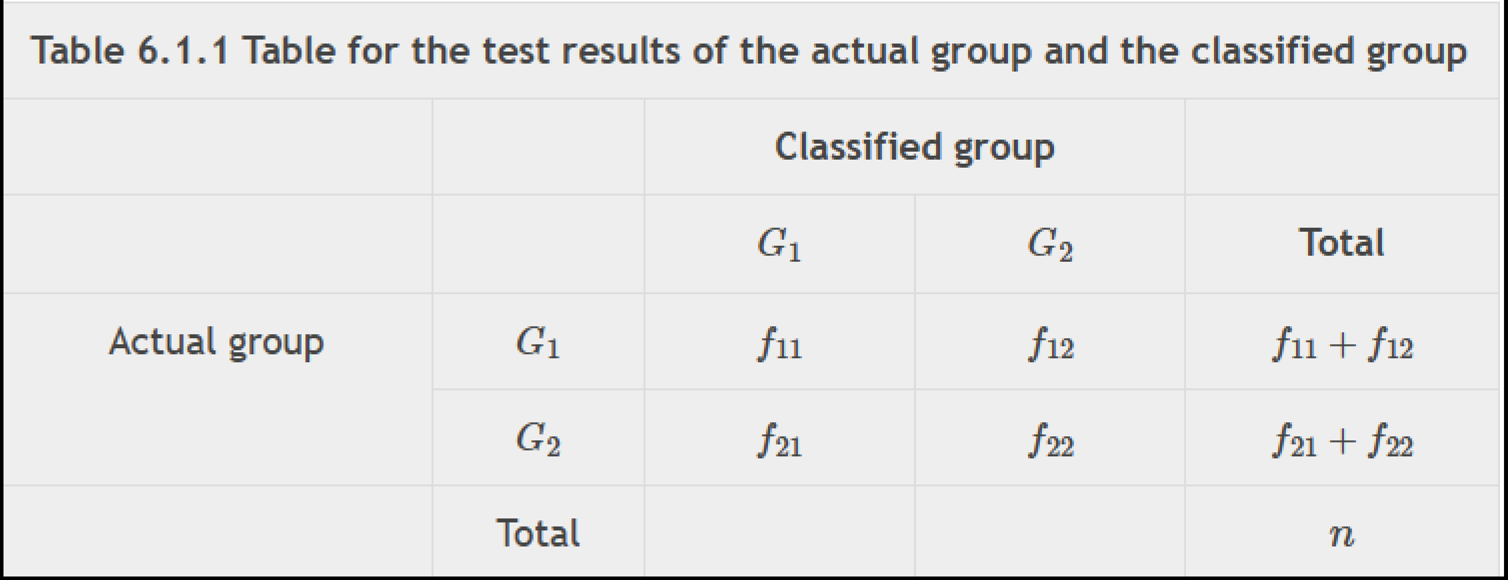
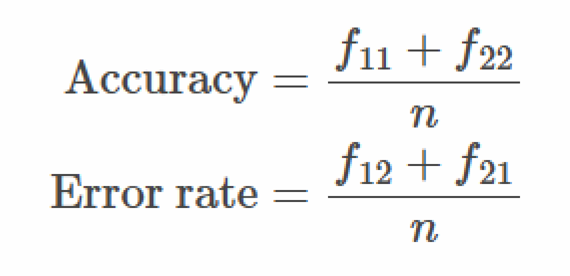


Table 6.1.1 Table for the test results of the actual group and the classified group

		Classified group		
		G_1	G_2	Total
Actual group	G_1	f_{11}	f_{12}	$f_{11} + f_{12}$
	G_2	f_{21}	f_{22}	$f_{21} + f_{22}$
	Total			n



$$\text{Accuracy} = \frac{f_{11} + f_{22}}{n}$$

$$\text{Error rate} = \frac{f_{12} + f_{21}}{n}$$

6.1 Basic concept of supervised machine learning and classification model

❖ Spliting method for training and testing data

- Holdout method
  - divides the entire data set into two non-overlapping data sets and holding out one as training data and the other as testing data. (½ training, ½ testing)
  - Repeat r times and calculate the average accuracy.

$$\text{Overall accuarcy} = \frac{1}{r}\Sigma_{i=1}^{r}(Accuracy)_i$$

- Cross validation method
  - Change the role of training and testing data
  - K-fold cross validation is possible

6.1 Basic concept of supervised machine learning and classification model

❖ Spliting method for training and testing data

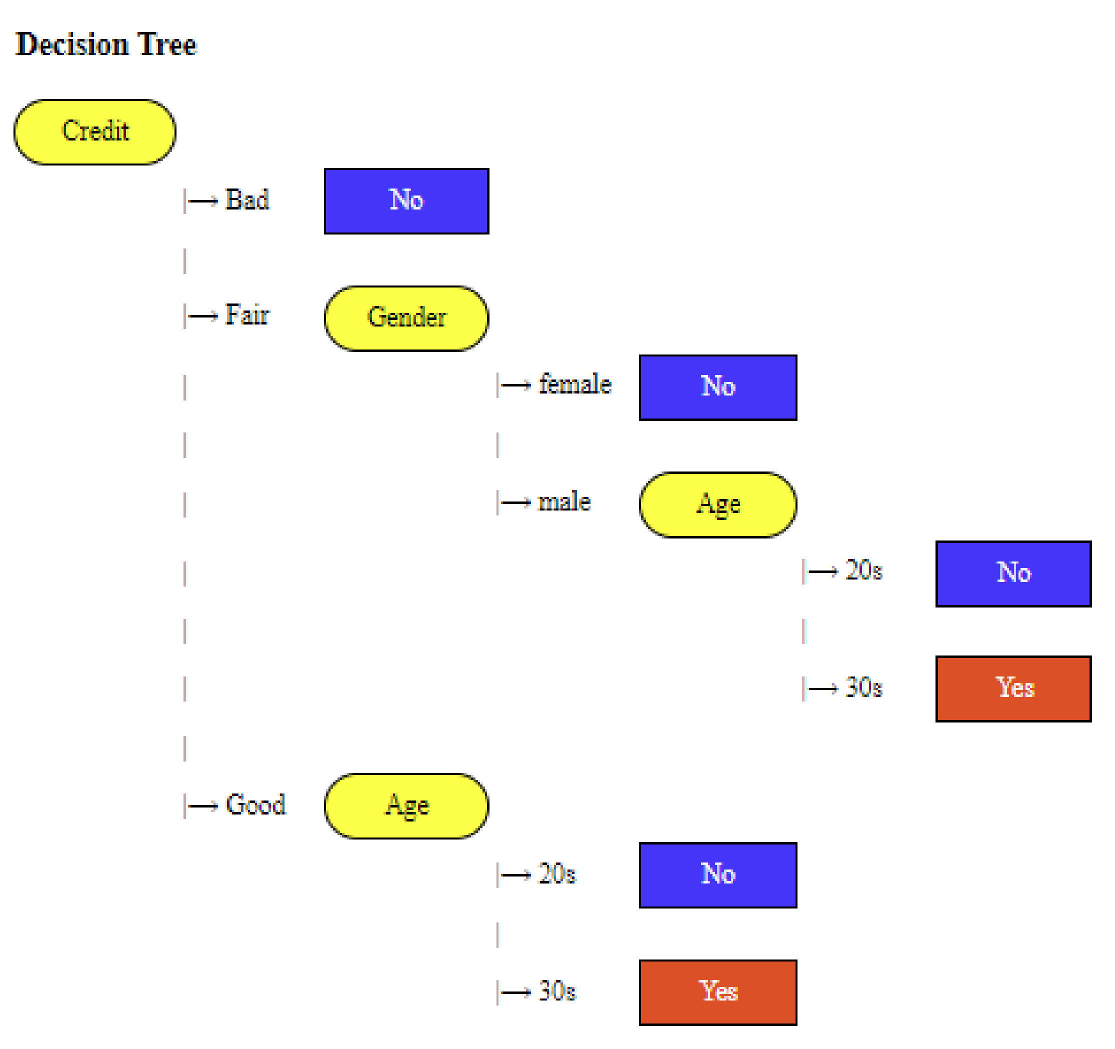
- Bootstrap method
  - Sampling with replacement for training data
  - If the total number of data is N, the number of sample extracted for training data using the bootstrap sampling is 63.2% in average
  - Probability that each data is selected by the bootstrap sampling is $1 - (1 - 1/N)^N$

6.2 Decision tree model

- Decision Tree
  - tree-shaped drawing of a classification function consisting of decision rules
  - Ovals represent nodes: tests for variables

    top node: root node
  - Branches: values of the tested variables
  - Rectangles: final classified groups which are called leaves

6.2 Decision tree model



Decision Tree

Credit

|→ Bad No

|→ Fair Gender

|→ female No

|→ male Age

|→ 20s No

|→ 30s Yes

|→ Good Age

|→ 20s No

|→ 30s Yes

6.2 Decision tree model

❖ Decision tree algorithm

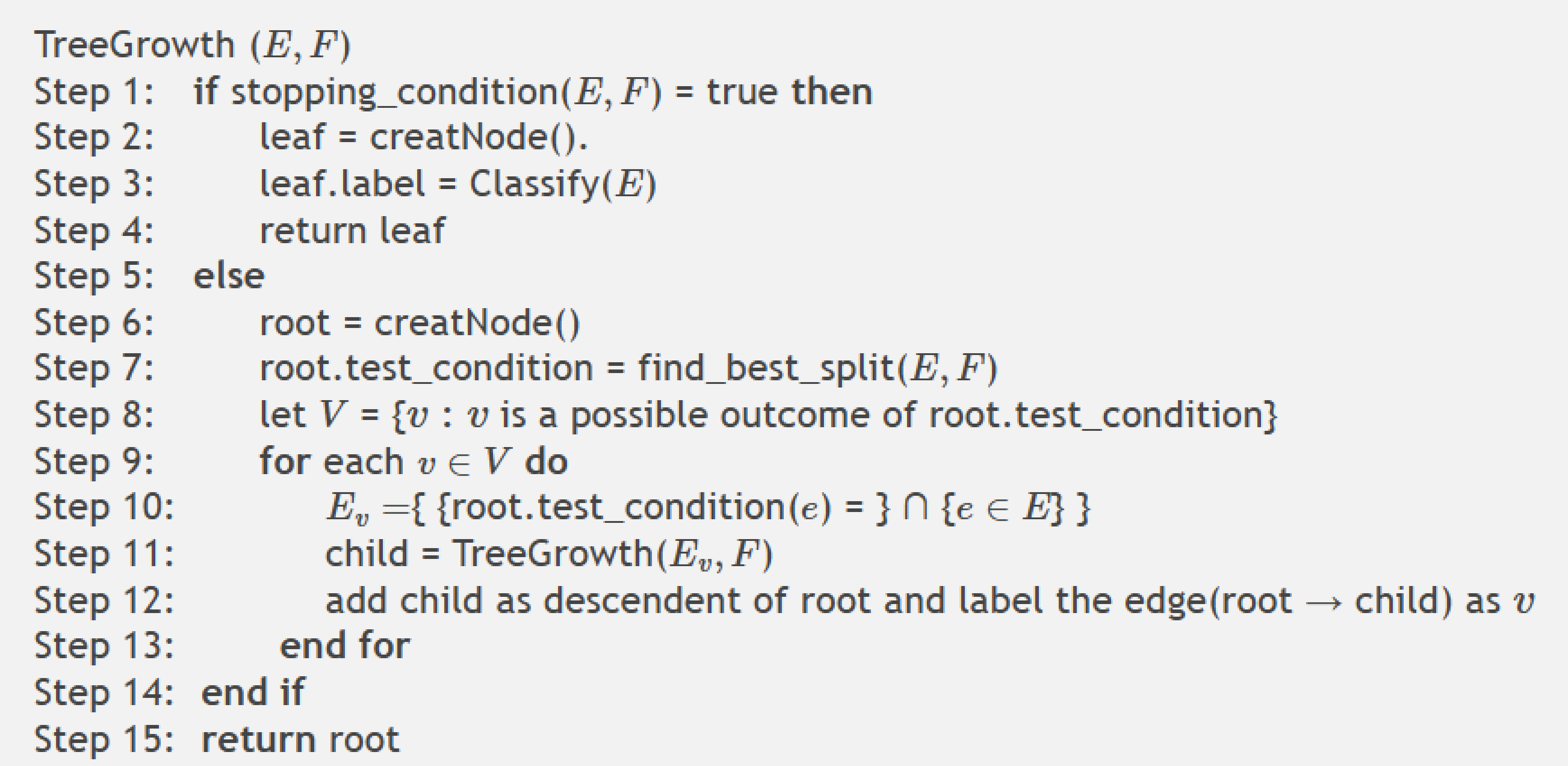
- The number of cases for creating a decision tree is exponentially proportional to the number of variables and values of each variable
- How do you quickly find a decision tree with high classification accuracy?
- It is not easy to find the global optimum in the entire number of decision tree cases
- We create an algorithm that is locally optimal

  => Form a decision tree by selecting the optimal variable at each node

6.2 Decision tree model

❖ Decision tree algorithm

❖ E = {data 집합} F = {변수집합}



```
TreeGrowth ($E, F$)
Step 1:   if stopping_condition($E, F$) = true then
Step 2:       leaf = creatNode().
Step 3:       leaf.label = Classify($E$)
Step 4:       return leaf
Step 5:   else
Step 6:       root = creatNode()
Step 7:       root.test_condition = find_best_split($E, F$)
Step 8:       let $V$ = {$v$ : $v$ is a possible outcome of root.test_condition}
Step 9:       for each $v \in V$ do
Step 10:          $E_v$ ={ {root.test_condition($e$) = } ∩ {$e \in E$} }
Step 11:          child = TreeGrowth($E_v, F$)
Step 12:          add child as descendent of root and label the edge(root $\rightarrow$ child) as $v$
Step 13:       end for
Step 14:  end if
Step 15:  return root
```

6.2 Decision tree model

❖ Selection of a variable for branching

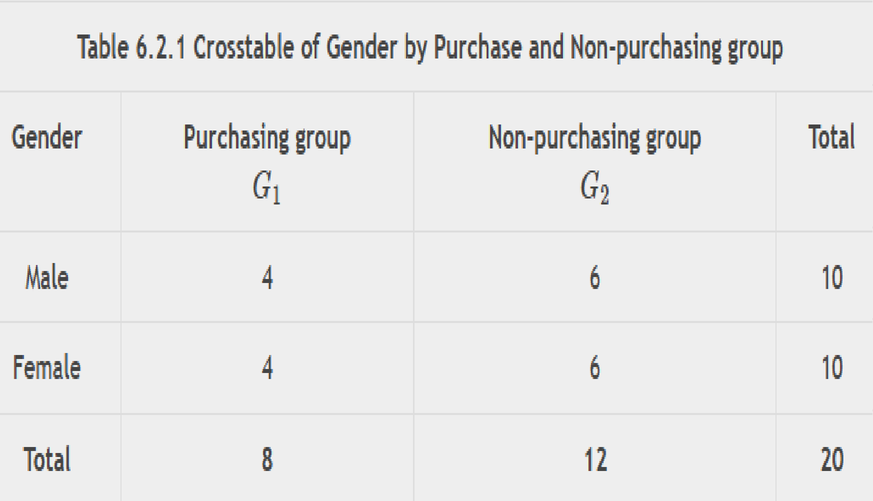


Table 6.2.1 Crosstable of Gender by Purchase and Non-purchasing group

Gender	Purchasing group G_1	Non-purchasing group G_2	Total
Male	4	6	10
Female	4	6	10
Total	8	12	20

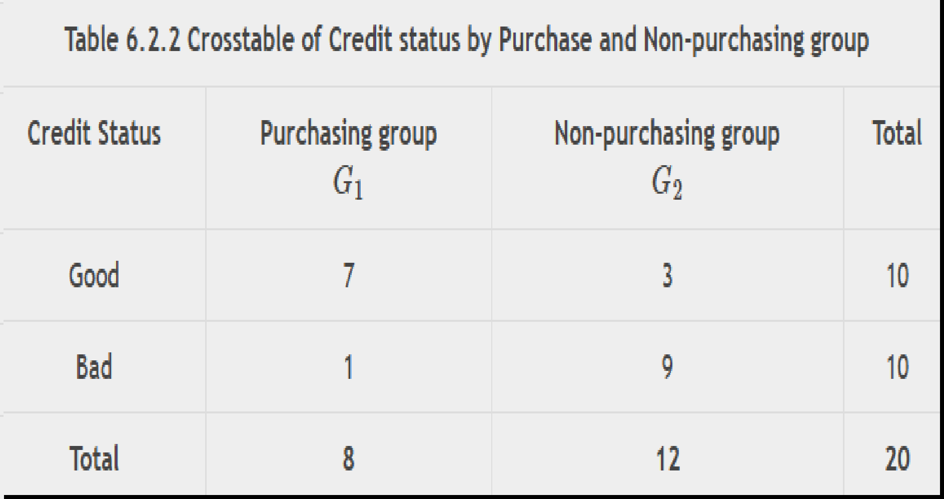


Table 6.2.2 Crosstable of Credit status by Purchase and Non-purchasing group

Credit Status	Purchasing group G_1	Non-purchasing group G_2	Total
Good	7	3	10
Bad	1	9	10
Total	8	12	20

- Gender: ratio of (Purchasing:Nonpurchasing) are the same (4:6) in males and females
- Credit status: Good (7:3), Bad (1:9)
- Selecting Credit Status will increase the classification accuracy.

6.2 Decision tree model

❖ Selection of a variable for branching

- At each node, select a variable using

  - chi square independence test

  - entropy coefficient

  - Gini coefficient

  - classification error rate

6.2 Decision tree model

❖ Selection of a variable for branching

A. Chi-square independence test

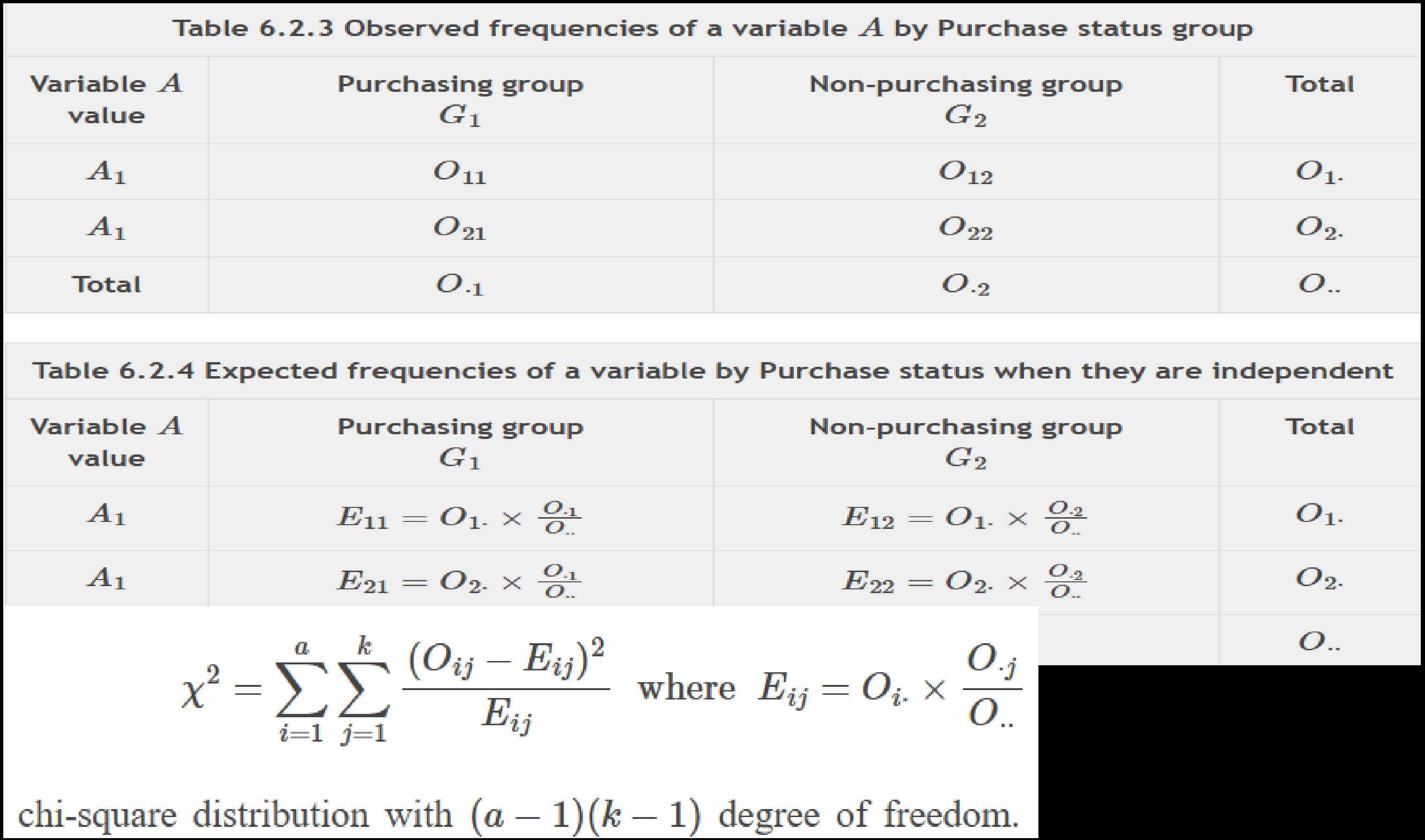


Table 6.2.3 Observed frequencies of a variable A by Purchase status group

Variable A value	Purchasing group G_1	Non-purchasing group G_2	Total
A_1	O_{11}	O_{12}	$O_{1\cdot}$
A_1	O_{21}	O_{22}	$O_{2\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot\cdot}$

Table 6.2.4 Expected frequencies of a variable by Purchase status when they are independent

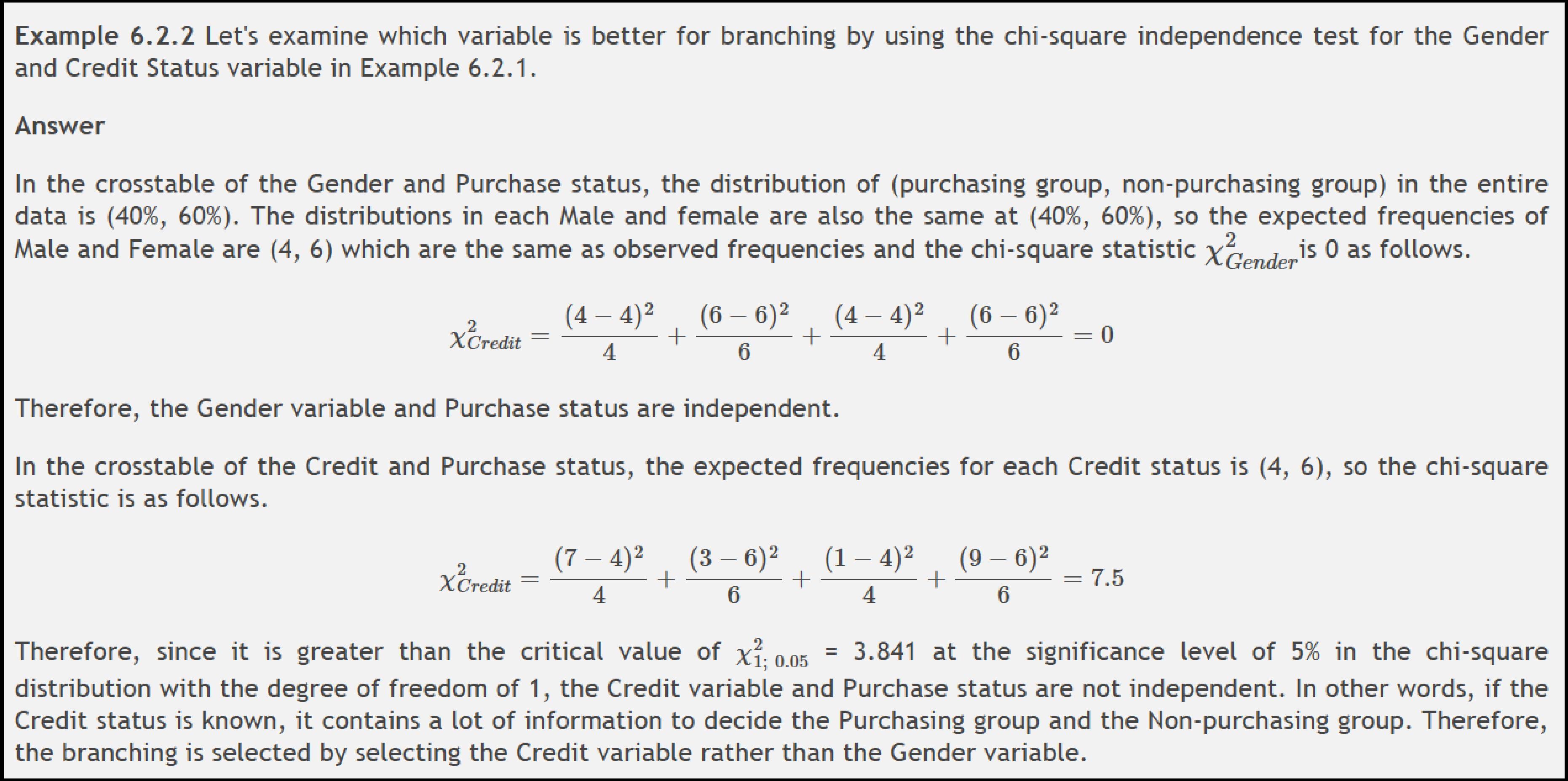
Variable A value	Purchasing group G_1	Non-purchasing group G_2	Total
A_1	$E_{11} = O_{1\cdot} \times \frac{O_{\cdot 1}}{O_{\cdot\cdot}}$	$E_{12} = O_{1\cdot} \times \frac{O_{\cdot 2}}{O_{\cdot\cdot}}$	$O_{1\cdot}$
A_1	$E_{21} = O_{2\cdot} \times \frac{O_{\cdot 1}}{O_{\cdot\cdot}}$	$E_{22} = O_{2\cdot} \times \frac{O_{\cdot 2}}{O_{\cdot\cdot}}$	$O_{2\cdot}$
			$O_{\cdot\cdot}$

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = O_{i\cdot} \times \frac{O_{\cdot j}}{O_{\cdot\cdot}}$$

chi-square distribution with $(a-1)(k-1)$ degree of freedom.

6.2 Decision tree model

❖ Selection of a variable for branching



Example 6.2.2 Let's examine which variable is better for branching by using the chi-square independence test for the Gender and Credit Status variable in Example 6.2.1.

Answer

In the crosstable of the Gender and Purchase status, the distribution of (purchasing group, non-purchasing group) in the entire data is (40%, 60%). The distributions in each Male and female are also the same at (40%, 60%), so the expected frequencies of Male and Female are (4, 6) which are the same as observed frequencies and the chi-square statistic χ^2_{Gender} is 0 as follows.

$$\chi^2_{Credit} = \frac{(4-4)^2}{4} + \frac{(6-6)^2}{6} + \frac{(4-4)^2}{4} + \frac{(6-6)^2}{6} = 0$$

Therefore, the Gender variable and Purchase status are independent.

In the crosstable of the Credit and Purchase status, the expected frequencies for each Credit status is (4, 6), so the chi-square statistic is as follows.

$$\chi^2_{Credit} = \frac{(7-4)^2}{4} + \frac{(3-6)^2}{6} + \frac{(1-4)^2}{4} + \frac{(9-6)^2}{6} = 7.5$$

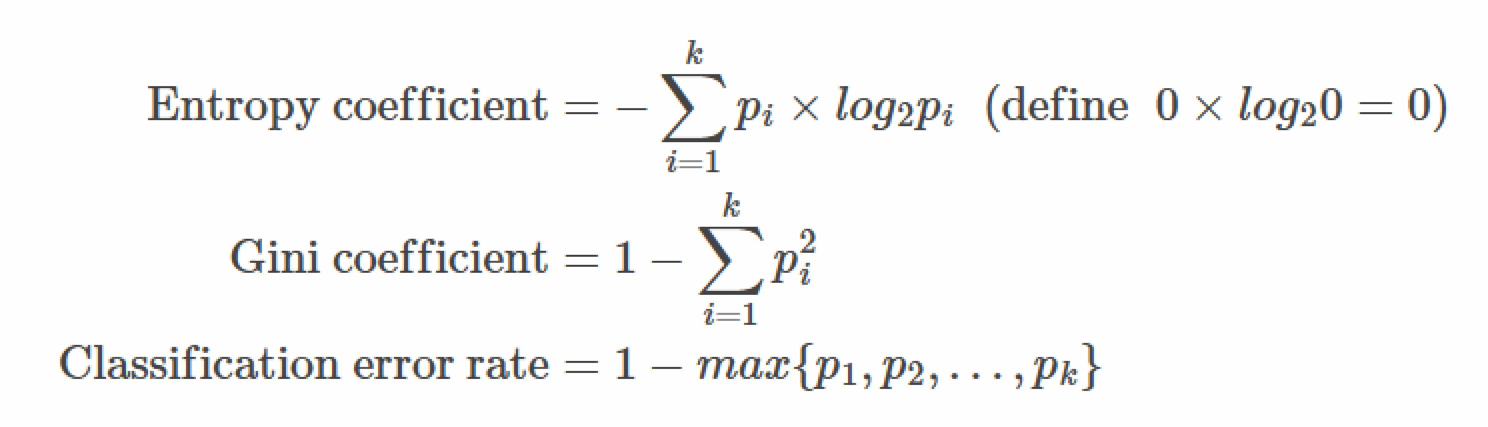
Therefore, since it is greater than the critical value of $\chi^2_{1;\ 0.05}$ = 3.841 at the significance level of 5% in the chi-square distribution with the degree of freedom of 1, the Credit variable and Purchase status are not independent. In other words, if the Credit status is known, it contains a lot of information to decide the Purchasing group and the Non-purchasing group. Therefore, the branching is selected by selecting the Credit variable rather than the Gender variable.

6.2 Decision tree model

❖ Selection of a variable for branching

B. Entropy, Gini coefficient, and classification error

- measure of uncertainty or purity about the distribution of each variable value by group.



$$\text{Entropy coefficient} = -\sum_{i=1}^{k} p_i \times log_2 p_i \quad (\text{define} \;\; 0 \times log_2 0 = 0)$$

$$\text{Gini coefficient} = 1 - \sum_{i=1}^{k} p_i^2$$

$$\text{Classification error rate} = 1 - max\{p_1, p_2, \ldots, p_k\}$$

6.2 Decision tree model

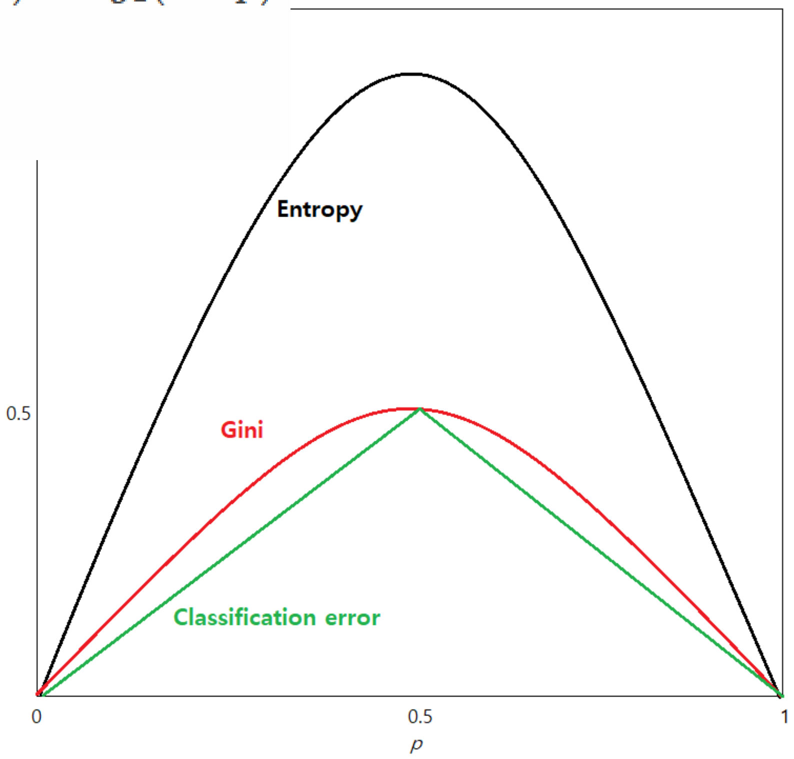
❖ Selection of a variable for branching

B. Entropy, Gini coefficient, and classification error: two groups

$$\text{Entropy coefficient} = -p \times log_2 p - (1-p) \times log_2(1-p)$$

$$\text{Gini coefficient} = 1 - p^2 - (1-p)^2$$

$$\text{Classification error rate} = 1 - max\{p, 1-p\}$$



6.2 Decision tree model

❖ Selection of a variable for branching

[Example 6.2.3] Find the entropy coefficient, Gini coefficient, and classification error rate for the distribution of the Purchasing group and the Non-purchasing group (40%, 60%) in the entire data. Also, find the entropy coefficient, Gini coefficient, and classification error rate for each Gender and Credit Status, and examine which variable is good for branching.

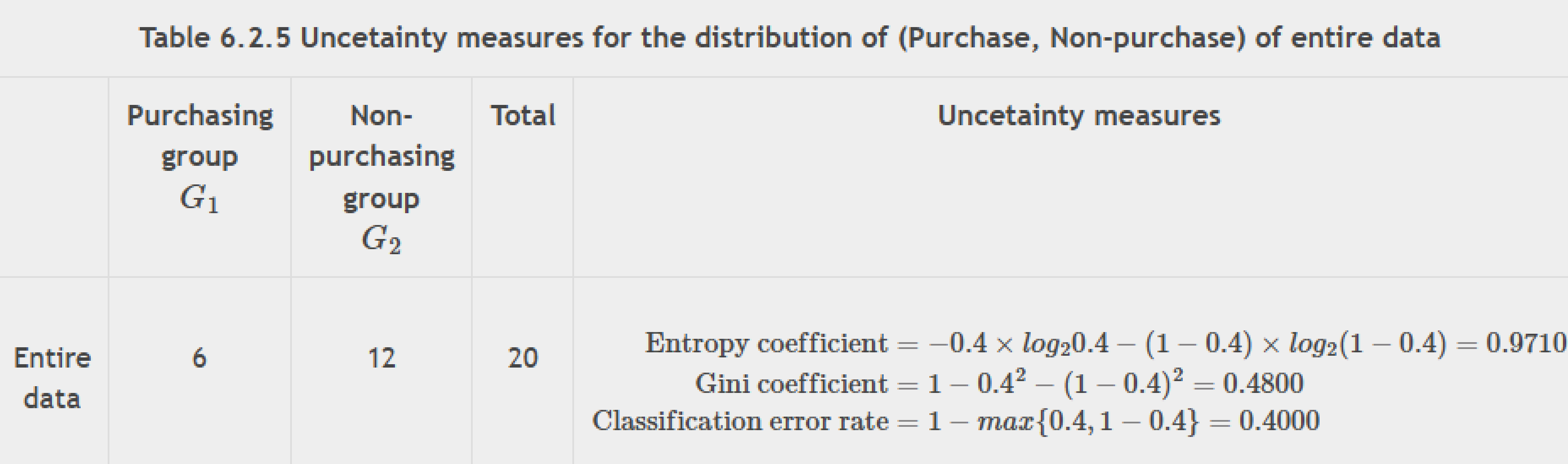


Table 6.2.5 Uncetainty measures for the distribution of (Purchase, Non-purchase) of entire data

	Purchasing group G_1	Non-purchasing group G_2	Total	Uncetainty measures
Entire data	6	12	20	$\text{Entropy coefficient} = -0.4 \times log_2 0.4 - (1-0.4) \times log_2(1-0.4) = 0.9710$ $\text{Gini coefficient} = 1 - 0.4^2 - (1-0.4)^2 = 0.4800$ $\text{Classification error rate} = 1 - max\{0.4, 1-0.4\} = 0.4000$

6.2 Decision tree model

❖ Selection of a variable for branching

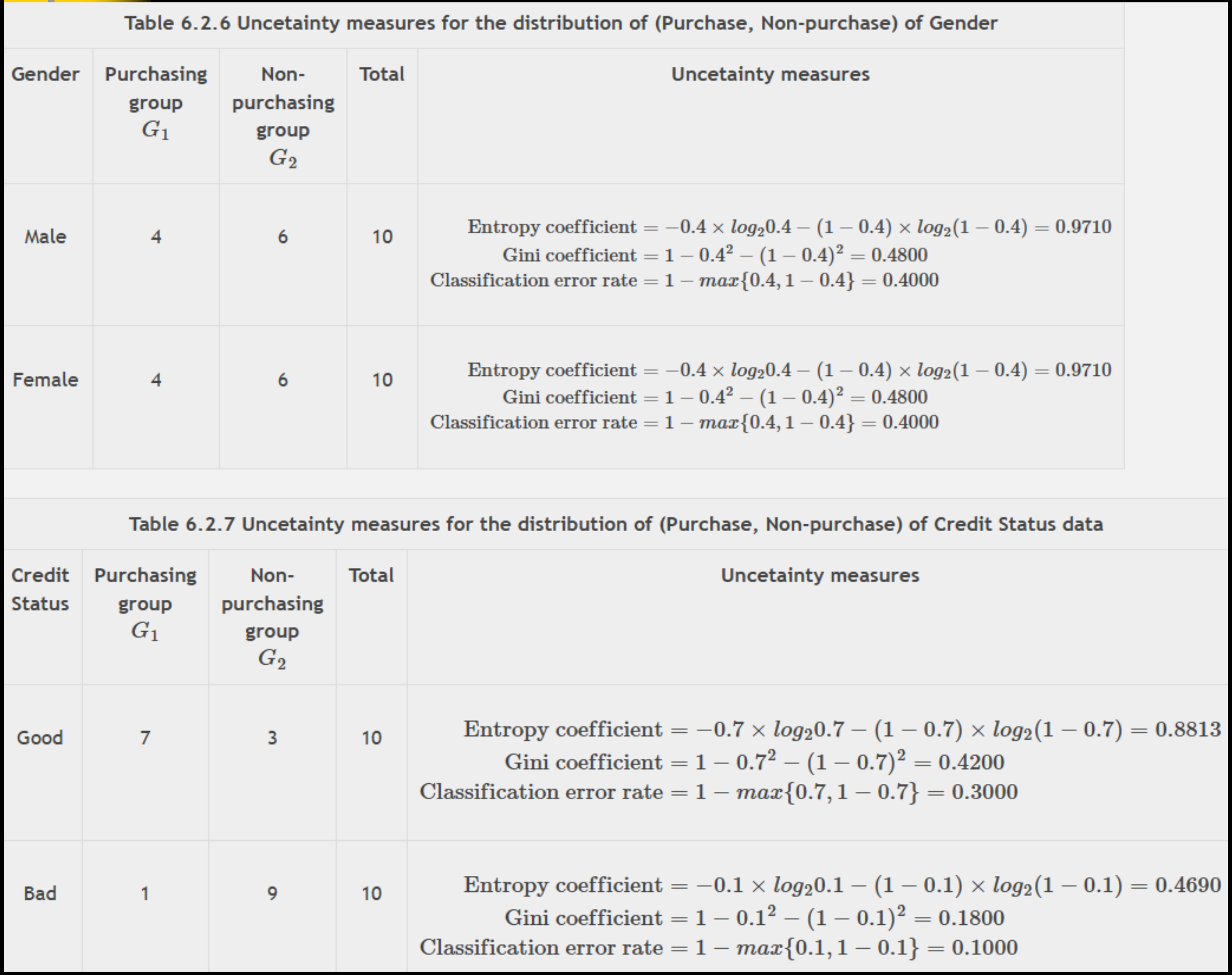


Table 6.2.6 Uncetainty measures for the distribution of (Purchase, Non-purchase) of Gender

Gender	Purchasing group G_1	Non-purchasing group G_2	Total	Uncetainty measures
Male	4	6	10	$\text{Entropy coefficient} = -0.4 \times log_2 0.4 - (1-0.4) \times log_2(1-0.4) = 0.9710$ $\text{Gini coefficient} = 1 - 0.4^2 - (1-0.4)^2 = 0.4800$ $\text{Classification error rate} = 1 - max\{0.4, 1-0.4\} = 0.4000$
Female	4	6	10	$\text{Entropy coefficient} = -0.4 \times log_2 0.4 - (1-0.4) \times log_2(1-0.4) = 0.9710$ $\text{Gini coefficient} = 1 - 0.4^2 - (1-0.4)^2 = 0.4800$ $\text{Classification error rate} = 1 - max\{0.4, 1-0.4\} = 0.4000$

Table 6.2.7 Uncetainty measures for the distribution of (Purchase, Non-purchase) of Credit Status data

Credit Status	Purchasing group G_1	Non-purchasing group G_2	Total	Uncetainty measures
Good	7	3	10	$\text{Entropy coefficient} = -0.7 \times log_2 0.7 - (1-0.7) \times log_2(1-0.7) = 0.8813$ $\text{Gini coefficient} = 1 - 0.7^2 - (1-0.7)^2 = 0.4200$ $\text{Classification error rate} = 1 - max\{0.7, 1-0.7\} = 0.3000$
Bad	1	9	10	$\text{Entropy coefficient} = -0.1 \times log_2 0.1 - (1-0.1) \times log_2(1-0.1) = 0.4690$ $\text{Gini coefficient} = 1 - 0.1^2 - (1-0.1)^2 = 0.1800$ $\text{Classification error rate} = 1 - max\{0.1, 1-0.1\} = 0.1000$

- Uncertainty of Credit status is higher than Gender

- Combine uncertainty of each variable value
- => Expected uncertainty

6.2 Decision tree model

❖ Selection of a variable for branching

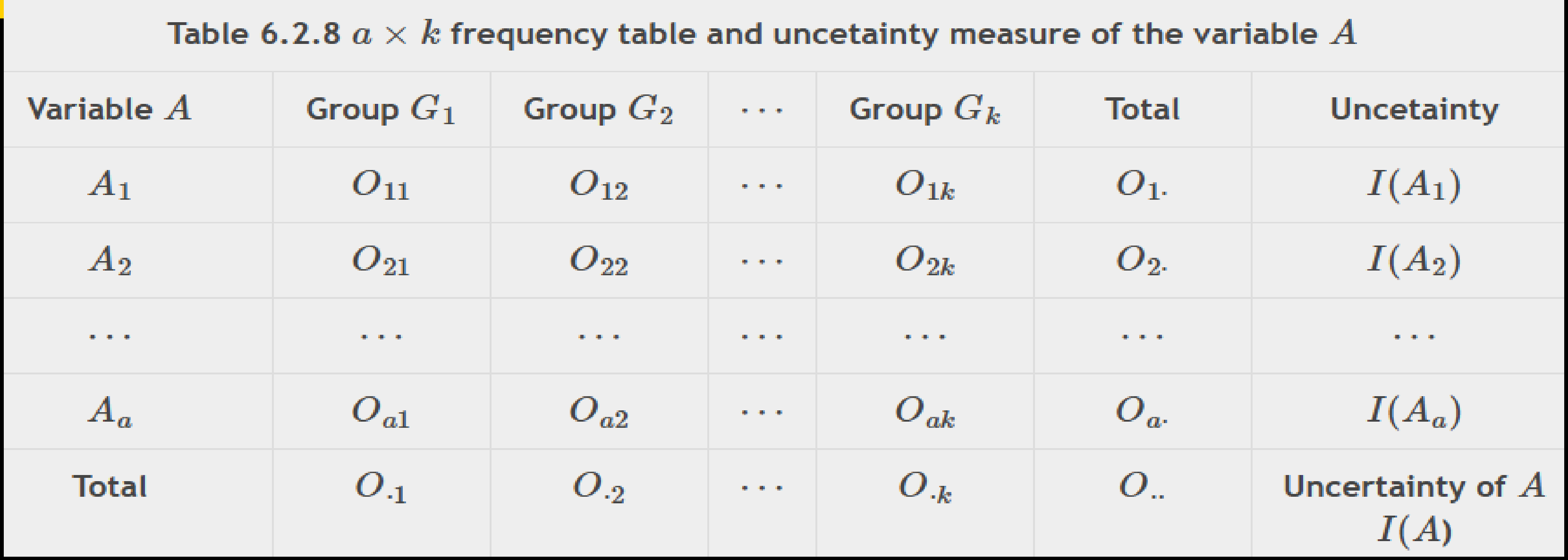


Table 6.2.8 $a \times k$ frequency table and uncetainty measure of the variable A

Variable A	Group G_1	Group G_2	$\cdots$	Group G_k	Total	Uncetainty
A_1	O_{11}	O_{12}	$\cdots$	O_{1k}	$O_{1\cdot}$	$I(A_1)$
A_2	O_{21}	O_{22}	$\cdots$	O_{2k}	$O_{2\cdot}$	$I(A_2)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_a	O_{a1}	O_{a2}	$\cdots$	O_{ak}	$O_{a\cdot}$	$I(A_a)$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot k}$	$O_{\cdot\cdot}$	Uncertainty of A $I(A)$

- Node T uncertainty

$$I(T) = -\sum_{j=1}^{k} \left(\frac{O_{\cdot j}}{O_{\cdot\cdot}} \right) \; log_2 \left(\frac{O_{\cdot j}}{O_{\cdot\cdot}} \right)$$

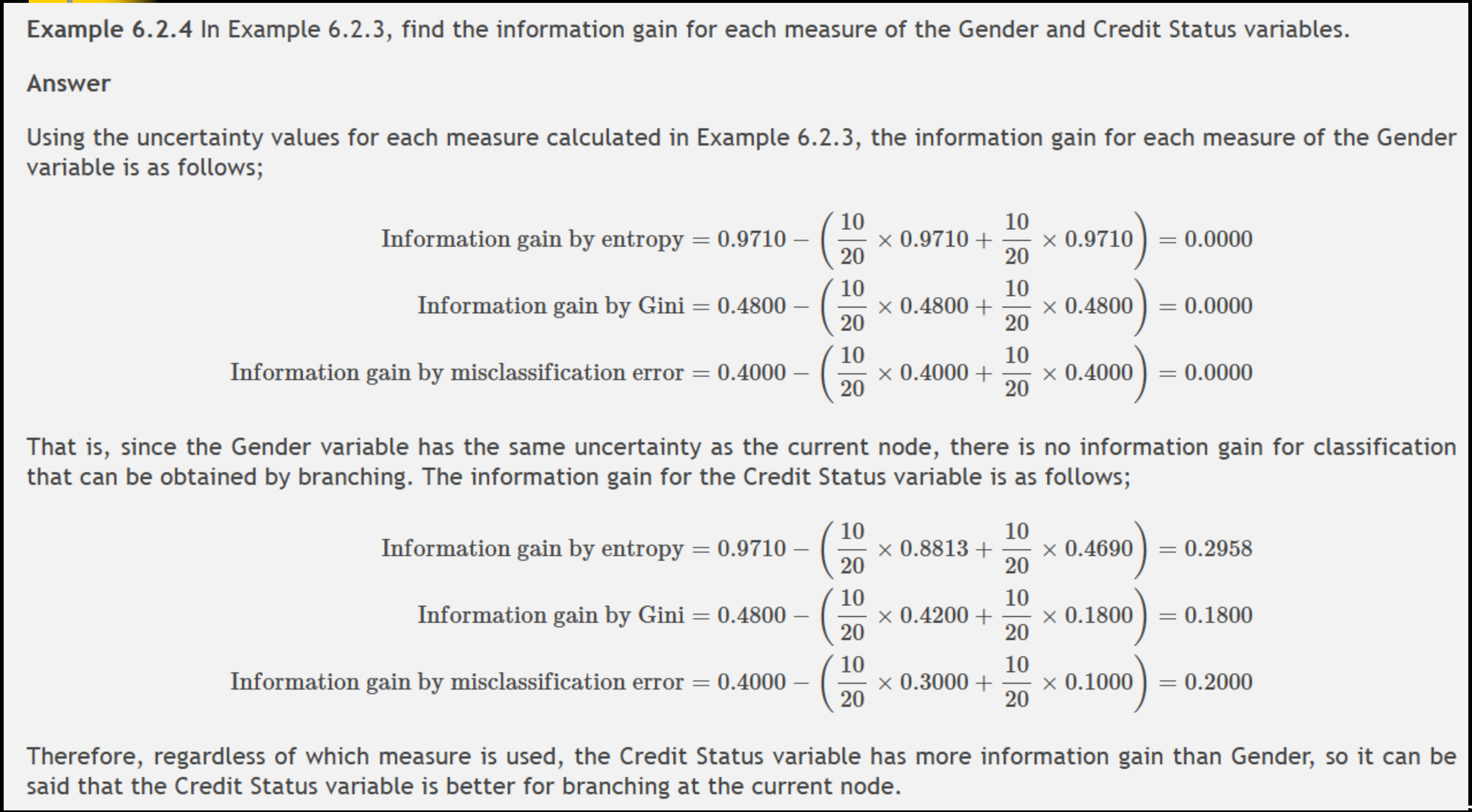
- Expected uncertainty of variable A

$$I(A) = \frac{O_{1\cdot}}{O_{\cdot\cdot}} \times I(A_1) + \frac{O_{2\cdot}}{O_{\cdot\cdot}} \times I(A_2) + \cdots + \frac{O_{a\cdot}}{O_{\cdot\cdot}} \times I(A_a)$$

- **Branch into a variable with larger information gain = *I(T) – I(A)***

6.2 Decision tree model

❖ Selection of a variable for branching



Example 6.2.4 In Example 6.2.3, find the information gain for each measure of the Gender and Credit Status variables.

Answer

Using the uncertainty values for each measure calculated in Example 6.2.3, the information gain for each measure of the Gender variable is as follows;

$$\text{Information gain by entropy} = 0.9710 - \left(\frac{10}{20} \times 0.9710 + \frac{10}{20} \times 0.9710\right) = 0.0000$$

$$\text{Information gain by Gini} = 0.4800 - \left(\frac{10}{20} \times 0.4800 + \frac{10}{20} \times 0.4800\right) = 0.0000$$

$$\text{Information gain by misclassification error} = 0.4000 - \left(\frac{10}{20} \times 0.4000 + \frac{10}{20} \times 0.4000\right) = 0.0000$$

That is, since the Gender variable has the same uncertainty as the current node, there is no information gain for classification that can be obtained by branching. The information gain for the Credit Status variable is as follows;

$$\text{Information gain by entropy} = 0.9710 - \left(\frac{10}{20} \times 0.8813 + \frac{10}{20} \times 0.4690\right) = 0.2958$$

$$\text{Information gain by Gini} = 0.4800 - \left(\frac{10}{20} \times 0.4200 + \frac{10}{20} \times 0.1800\right) = 0.1800$$

$$\text{Information gain by misclassification error} = 0.4000 - \left(\frac{10}{20} \times 0.3000 + \frac{10}{20} \times 0.1000\right) = 0.2000$$

Therefore, regardless of which measure is used, the Credit Status variable has more information gain than Gender, so it can be said that the Credit Status variable is better for branching at the current node.

6.2 Decision tree model

❖ Selection of a variable for branching

[Example 6.2.5] When we surveyed 20 customers who visited a computer store, 8 customers purchased a computer (Purchasing group, and 12 customers did not purchase a computer (Non-purchasing group). The survey included variables such as gender, age, monthly income, and credit status of these 20 customers as well as Purchase status as shown in Table 6.2.9.

Note that all variables are surveyed as categorical variables such as (Female, Male) for gender, (20s, 30s) for age, (GE2000, LT2000) for income, (Bad, Fair, Good) for credit status, and (No, Yes) for Purchase status.

6.2 Decision tree model

❖ Selection of a variable for branching

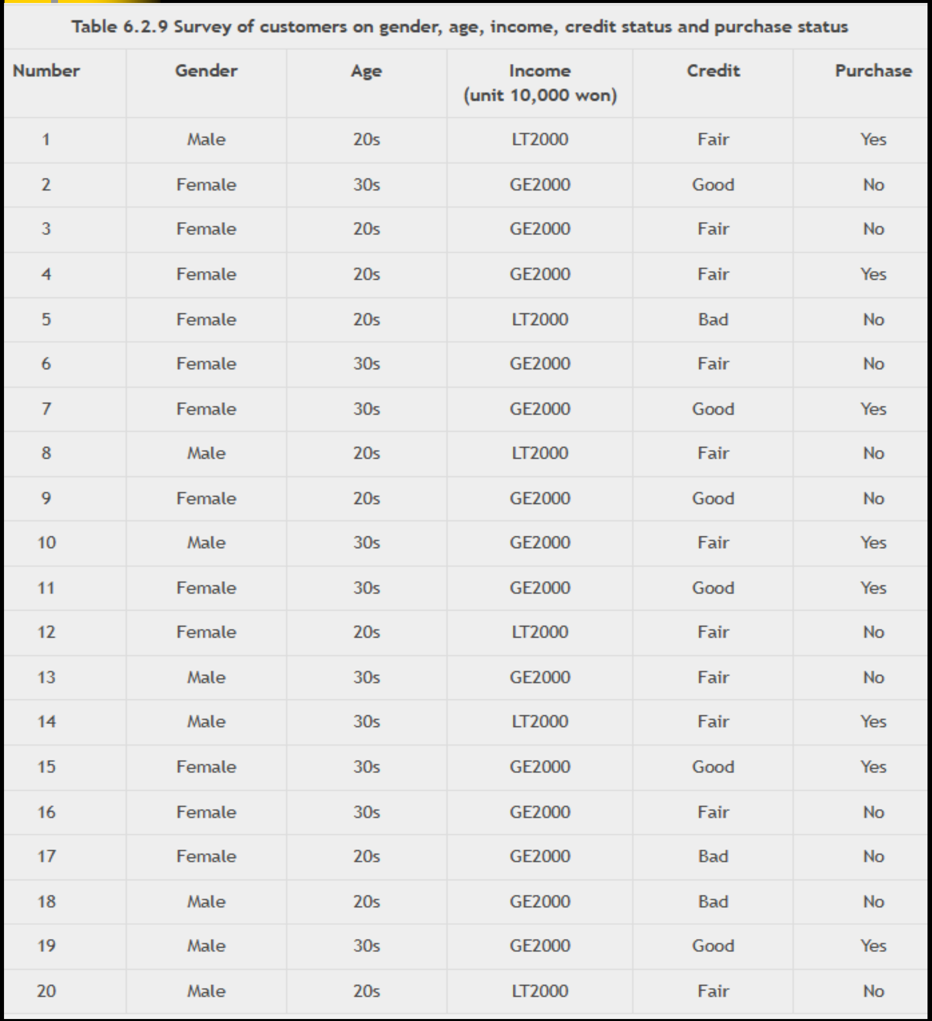


Table 6.2.9 Survey of customers on gender, age, income, credit status and purchase status

Number	Gender	Age	Income (unit 10,000 won)	Credit	Purchase
1	Male	20s	LT2000	Fair	Yes
2	Female	30s	GE2000	Good	No
3	Female	20s	GE2000	Fair	No
4	Female	20s	GE2000	Fair	Yes
5	Female	20s	LT2000	Bad	No
6	Female	30s	GE2000	Fair	No
7	Female	30s	GE2000	Good	Yes
8	Male	20s	LT2000	Fair	No
9	Female	20s	GE2000	Good	No
10	Male	30s	GE2000	Fair	Yes
11	Female	30s	GE2000	Good	Yes
12	Female	20s	LT2000	Fair	No
13	Male	30s	GE2000	Fair	No
14	Male	30s	LT2000	Fair	Yes
15	Female	30s	GE2000	Good	Yes
16	Female	30s	GE2000	Fair	No
17	Female	20s	GE2000	Bad	No
18	Male	20s	GE2000	Bad	No
19	Male	30s	GE2000	Good	Yes
20	Male	20s	LT2000	Fair	No

- E = {data set}
- F = {Gender, Age, Income, Credit}
- Target variable = Purchase
- Stopping rule
  - min number of data in leaf <= 5
  - all data belong to a group
- Uncertainty of root node T – Entropy coefficient of (8/20, 12/20)

$$I(T) = -0.4\times\log_2 0.4 - 0.6\times\log_2 0.6 = 0,9710$$

6.2 Decision tree model

❖ Selection of a variable for branching

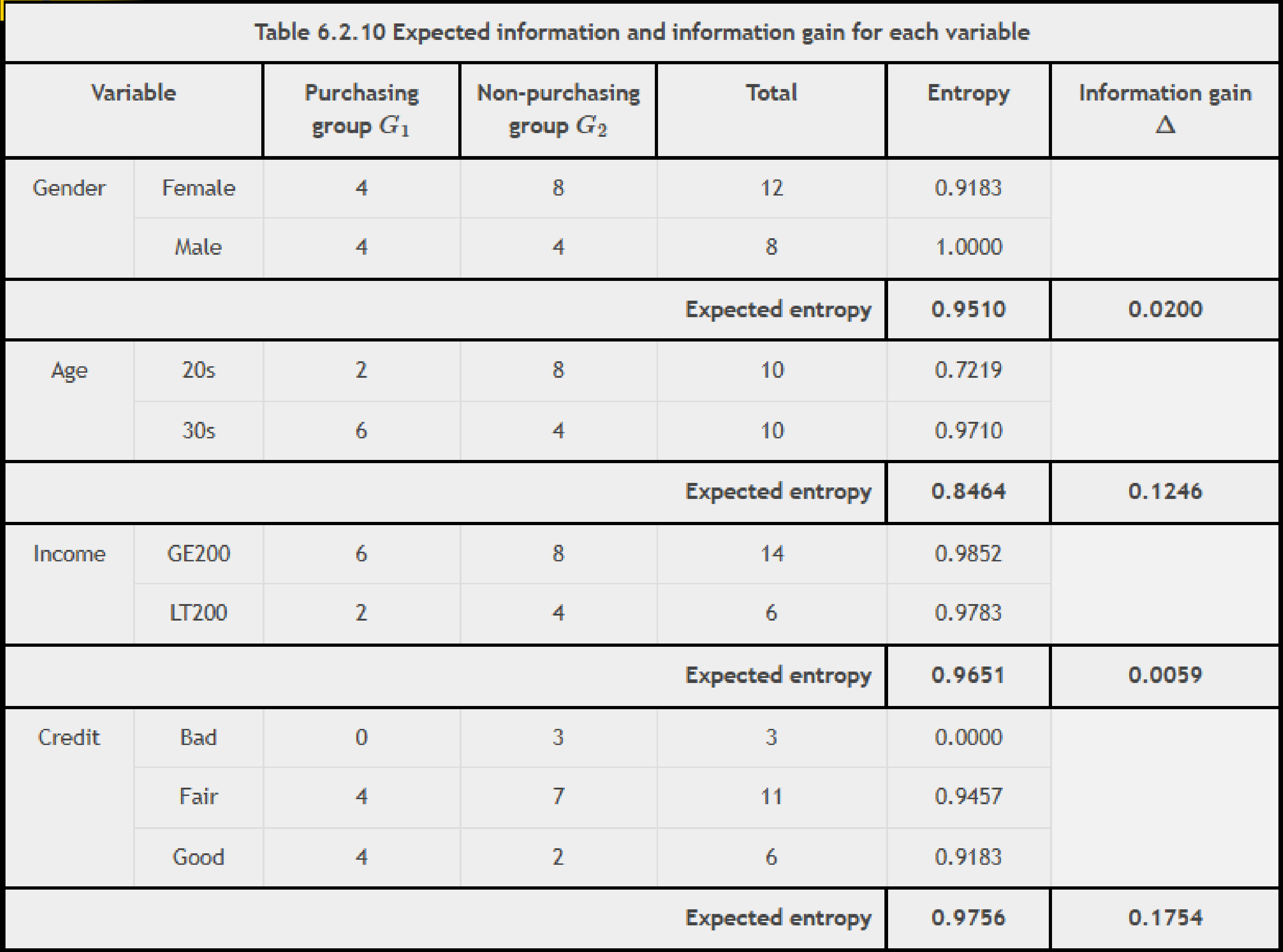


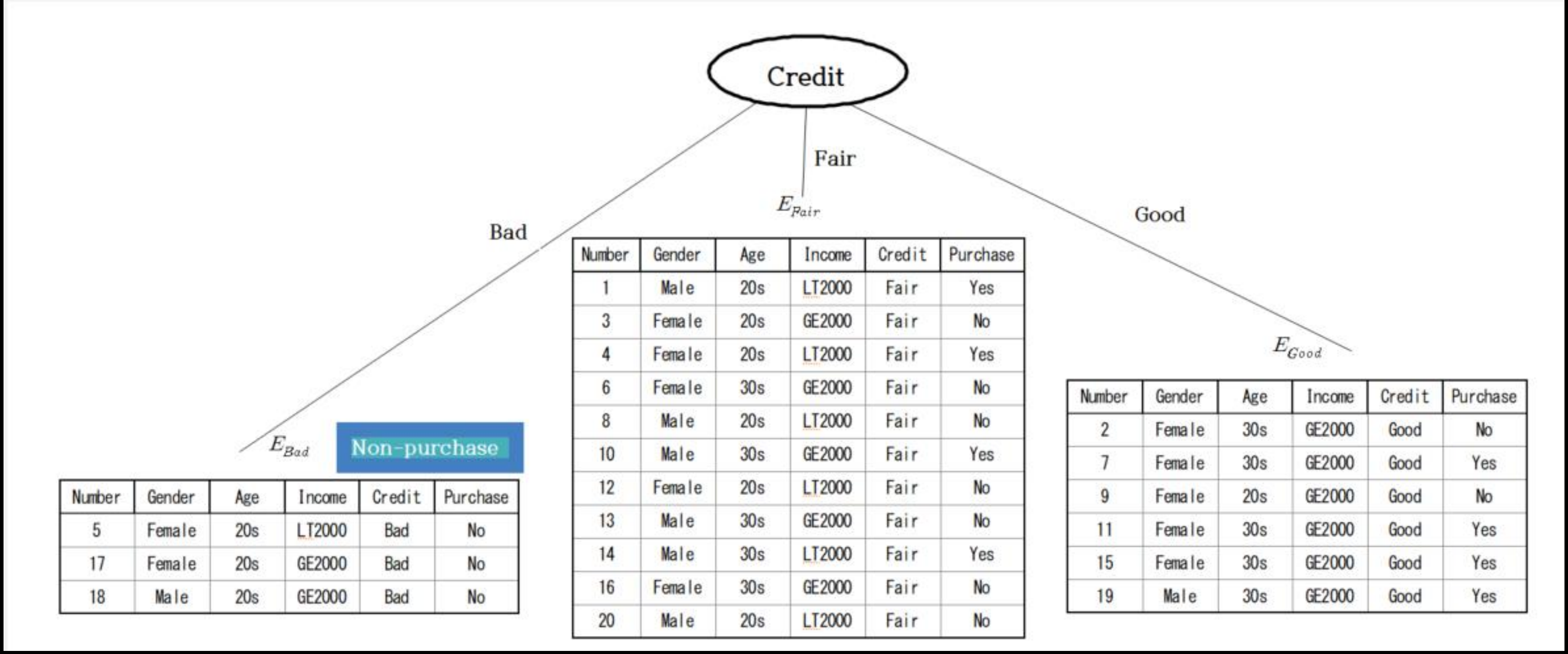
Table 6.2.10 Expected information and information gain for each variable

Variable		Purchasing group G_1	Non-purchasing group G_2	Total	Entropy	Information gain Δ
Gender	Female	4	8	12	0.9183	
	Male	4	4	8	1.0000	
Expected entropy					**0.9510**	**0.0200**
Age	20s	2	8	10	0.7219	
	30s	6	4	10	0.9710	
Expected entropy					**0.8464**	**0.1246**
Income	GE200	6	8	14	0.9852	
	LT200	2	4	6	0.9783	
Expected entropy					**0.9651**	**0.0059**
Credit	Bad	0	3	3	0.0000	
	Fair	4	7	11	0.9457	
	Good	4	2	6	0.9183	
Expected entropy					**0.9756**	**0.1754**

- Credit has the highest information gain.

6.2 Decision tree model

❖ Example of decision tree



Number	Gender	Age	Income	Credit	Purchase
5	Female	20s	LT2000	Bad	No
17	Female	20s	GE2000	Bad	No
18	Male	20s	GE2000	Bad	No

Number	Gender	Age	Income	Credit	Purchase
1	Male	20s	LT2000	Fair	Yes
3	Female	20s	GE2000	Fair	No
4	Female	20s	LT2000	Fair	Yes
6	Female	30s	GE2000	Fair	No
8	Male	20s	LT2000	Fair	No
10	Male	30s	GE2000	Fair	Yes
12	Female	20s	LT2000	Fair	No
13	Male	30s	GE2000	Fair	No
14	Male	30s	LT2000	Fair	Yes
16	Female	30s	GE2000	Fair	No
20	Male	20s	LT2000	Fair	No

Number	Gender	Age	Income	Credit	Purchase
2	Female	30s	GE2000	Good	No
7	Female	30s	GE2000	Good	Yes
9	Female	20s	GE2000	Good	No
11	Female	30s	GE2000	Good	Yes
15	Female	30s	GE2000	Good	Yes
19	Male	30s	GE2000	Good	Yes

6.2 Decision tree model

❖ Example of decision tree

- Since the stopping rule is not satisfied for the data set of 11 people with Fair credit, , this data set needs further split.
- Entropy coefficients for the distribution (4/11, 7/11): $I(E_{Fair}) = -\frac{4}{11} \times log_2 \frac{4}{11} - \frac{7}{11} \times log_2 \frac{7}{11} = 0.9457$

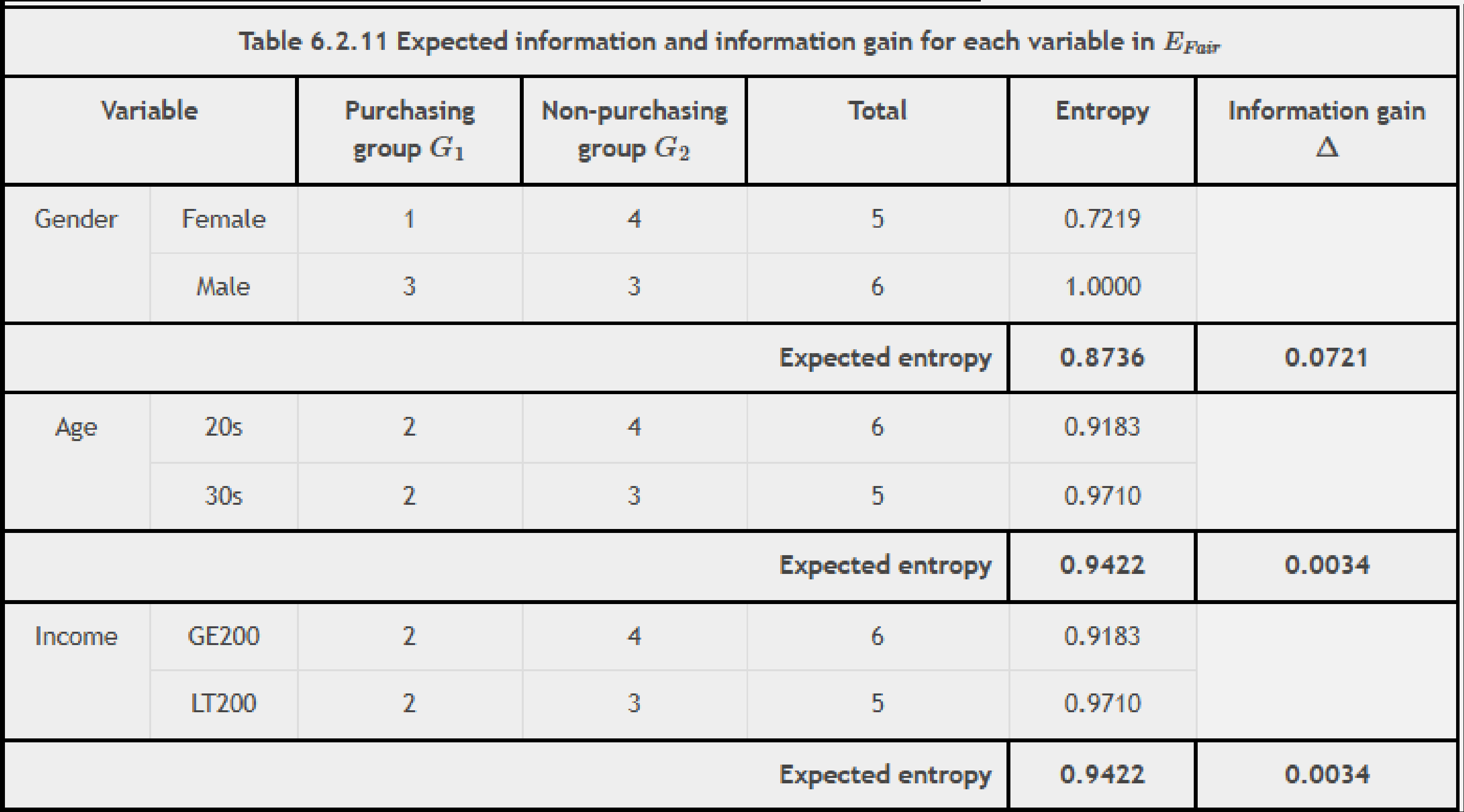
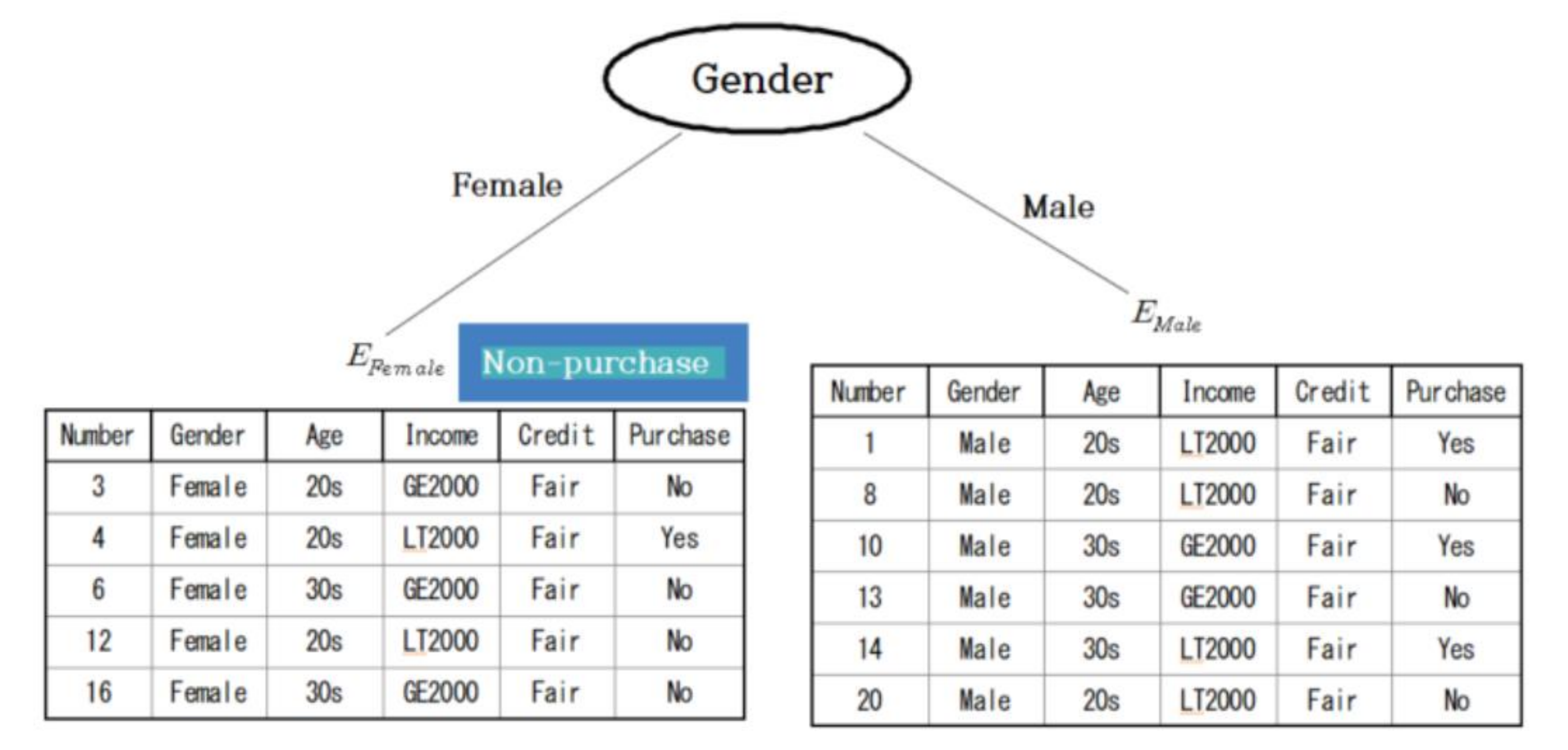


Table 6.2.11 Expected information and information gain for each variable in E_{Fair}

Variable		Purchasing group G_1	Non-purchasing group G_2	Total	Entropy	Information gain Δ
Gender	Female	1	4	5	0.7219	
	Male	3	3	6	1.0000	
				Expected entropy	**0.8736**	**0.0721**
Age	20s	2	4	6	0.9183	
	30s	2	3	5	0.9710	
				Expected entropy	**0.9422**	**0.0034**
Income	GE200	2	4	6	0.9183	
	LT200	2	3	5	0.9710	
				Expected entropy	**0.9422**	**0.0034**

6.2 Decision tree model

❖ Example of decision tree



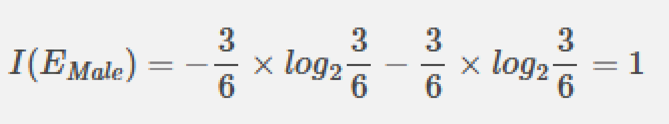
Number	Gender	Age	Income	Credit	Purchase
3	Female	20s	GE2000	Fair	No
4	Female	20s	LT2000	Fair	Yes
6	Female	30s	GE2000	Fair	No
12	Female	20s	LT2000	Fair	No
16	Female	30s	GE2000	Fair	No

Number	Gender	Age	Income	Credit	Purchase
1	Male	20s	LT2000	Fair	Yes
8	Male	20s	LT2000	Fair	No
10	Male	30s	GE2000	Fair	Yes
13	Male	30s	GE2000	Fair	No
14	Male	30s	LT2000	Fair	Yes
20	Male	20s	LT2000	Fair	No

6.2 Decision tree model

❖ Example of decision tree

- Since there are 6 data sets of Male with Fair credit, the stopping rule is not satisfied and this data set needs further split.
- Entropy coefficients for the distribution (3/6, 3/6):



$$I(E_{Male}) = -\frac{3}{6} \times log_2 \frac{3}{6} - \frac{3}{6} \times log_2 \frac{3}{6} = 1$$

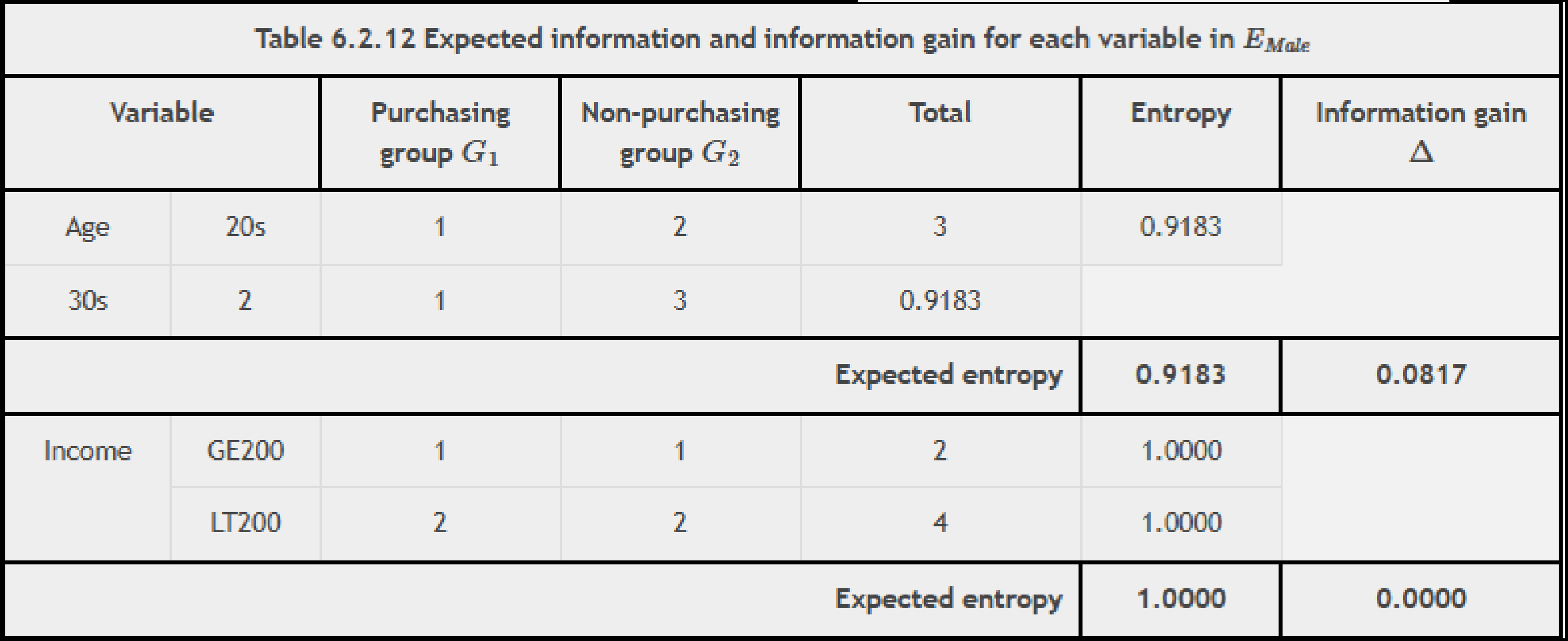
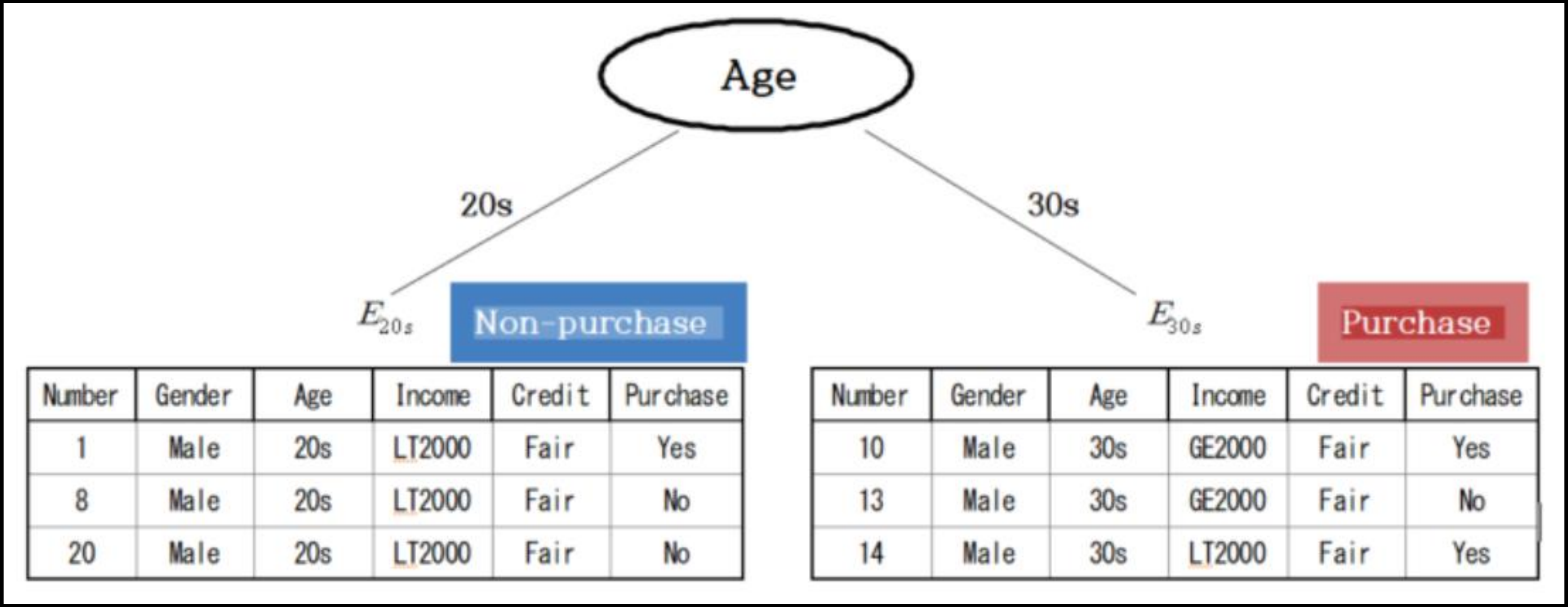


Table 6.2.12 Expected information and information gain for each variable in E_{Male}

Variable		Purchasing group G_1	Non-purchasing group G_2	Total	Entropy	Information gain Δ
Age	20s	1	2	3	0.9183	
30s	2	1	3	0.9183		
				Expected entropy	0.9183	0.0817
Income	GE200	1	1	2	1.0000	
	LT200	2	2	4	1.0000	
				Expected entropy	1.0000	0.0000

6.2 Decision tree model

❖ Example of decision tree



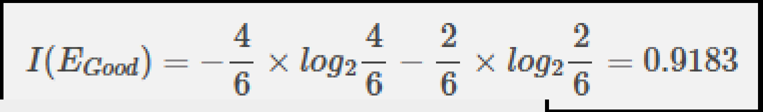
Number	Gender	Age	Income	Credit	Purchase
1	Male	20s	LT2000	Fair	Yes
8	Male	20s	LT2000	Fair	No
20	Male	20s	LT2000	Fair	No

Number	Gender	Age	Income	Credit	Purchase
10	Male	30s	GE2000	Fair	Yes
13	Male	30s	GE2000	Fair	No
14	Male	30s	LT2000	Fair	Yes

6.2 Decision tree model

❖ Example of decision tree

- At root node, since there are 6 data with good credit, the stopping rule is not satisfied and this data set needs further split.
- Entropy coefficients for the distribution (4/6, 2/6):



$$I(E_{Good}) = -\frac{4}{6} \times log_2\frac{4}{6} - \frac{2}{6} \times log_2\frac{2}{6} = 0.9183$$

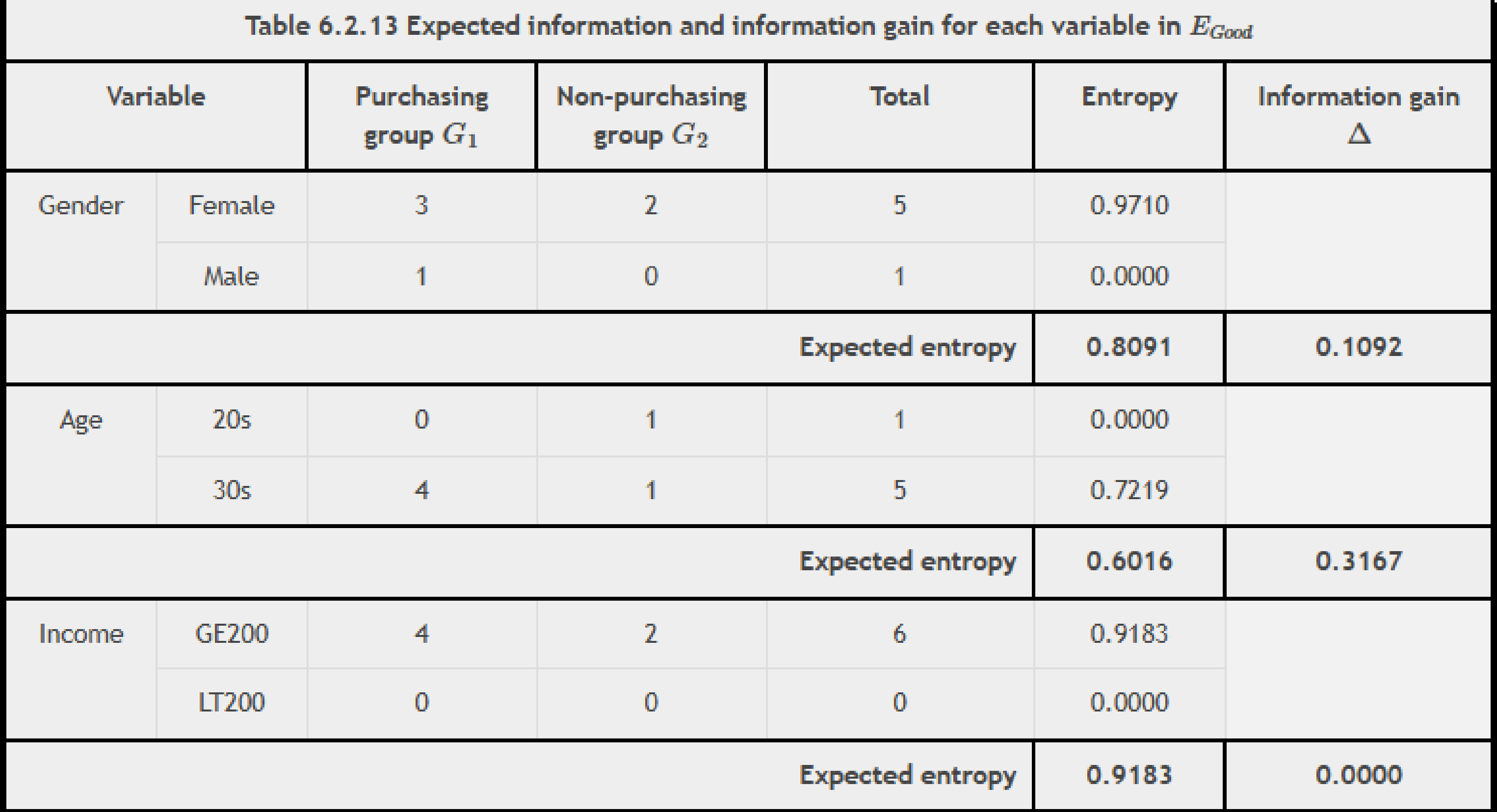
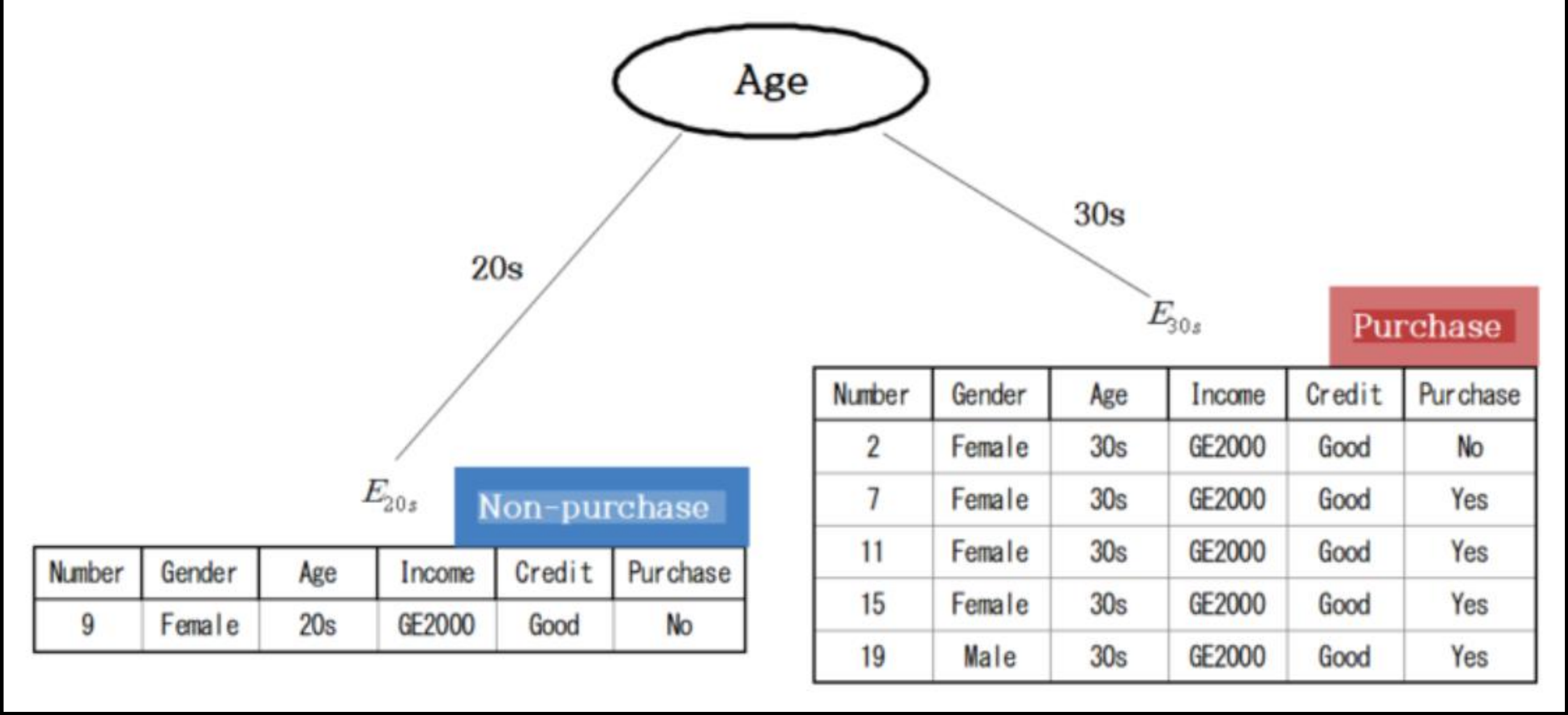


Table 6.2.13 Expected information and information gain for each variable in E_{Good}

Variable		Purchasing group G_1	Non-purchasing group G_2	Total	Entropy	Information gain Δ
Gender	Female	3	2	5	0.9710	
	Male	1	0	1	0.0000	
				Expected entropy	**0.8091**	**0.1092**
Age	20s	0	1	1	0.0000	
	30s	4	1	5	0.7219	
				Expected entropy	**0.6016**	**0.3167**
Income	GE200	4	2	6	0.9183	
	LT200	0	0	0	0.0000	
				Expected entropy	**0.9183**	**0.0000**

6.2 Decision tree model

❖ Example of decision tree

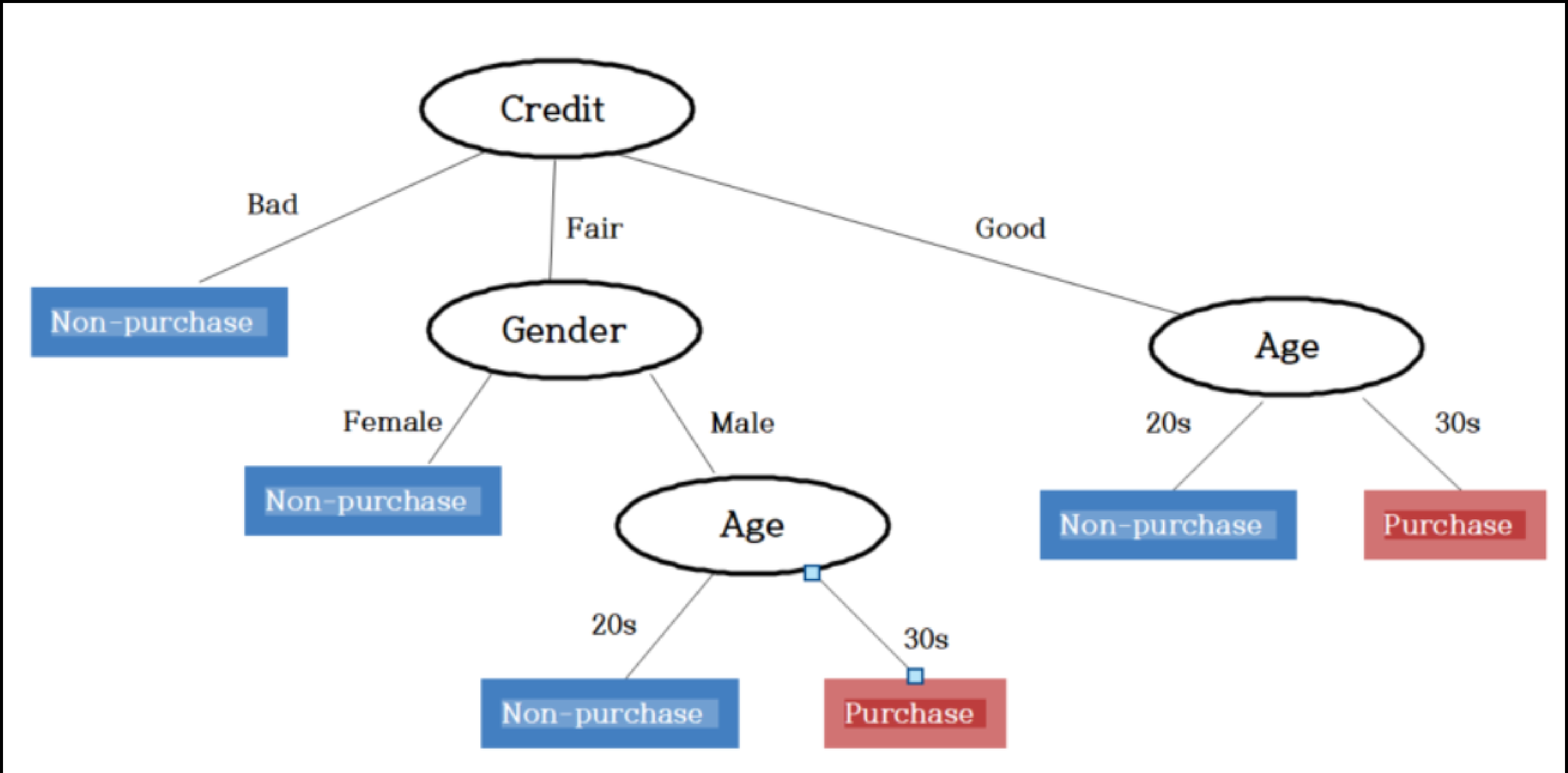


Number	Gender	Age	Income	Credit	Purchase
9	Female	20s	GE2000	Good	No

Number	Gender	Age	Income	Credit	Purchase
2	Female	30s	GE2000	Good	No
7	Female	30s	GE2000	Good	Yes
11	Female	30s	GE2000	Good	Yes
15	Female	30s	GE2000	Good	Yes
19	Male	30s	GE2000	Good	Yes

6.2 Decision tree model

❖ Example of decision tree



6.2 Decision tree model

❖ Categorization of a continuous variable

[Example 6.2.6] we want to divide the monthly income into two categories. What boundary value of the income is reasonable to divide for classification?

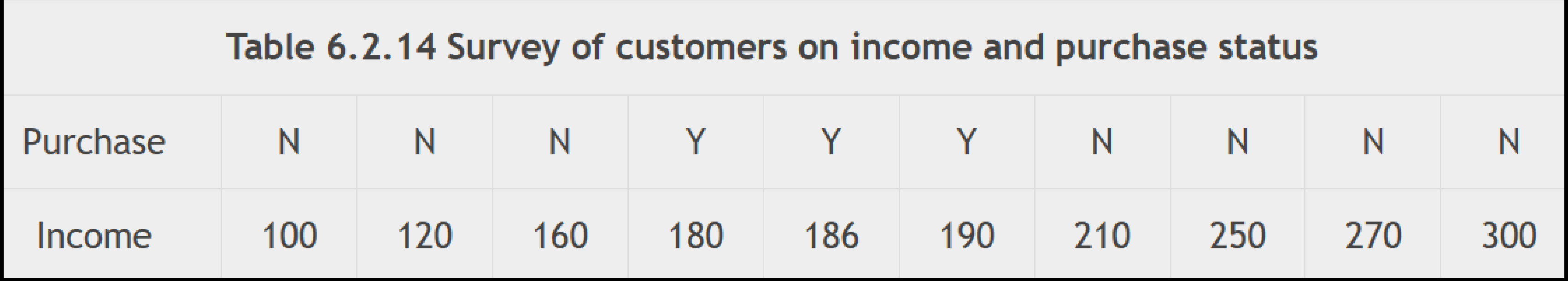
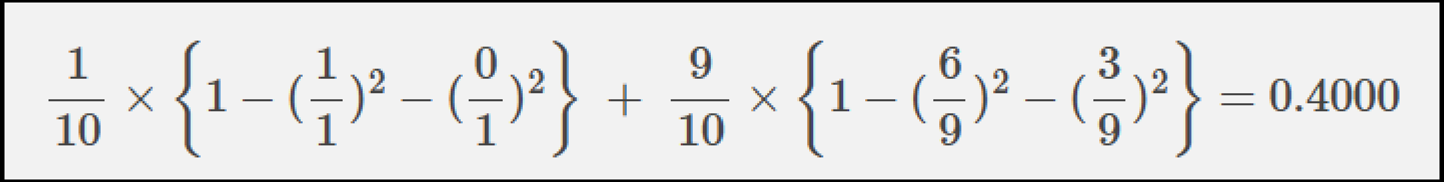


Table 6.2.14 Survey of customers on income and purchase status

Purchase	N	N	N	Y	Y	Y	N	N	N	N
Income	100	120	160	180	186	190	210	250	270	300

<Answer>

- Using the Gini coefficient as the uncertainty measure, the expected Gini coefficient when the middle value is 110 calculated as follows;



$$\frac{1}{10}\times\left\{1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2\right\}+\frac{9}{10}\times\left\{1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2\right\}=0.4000$$

6.2 Decision tree model

❖ Categorization of a continuous variable

<Answer of Example 6.2.6>

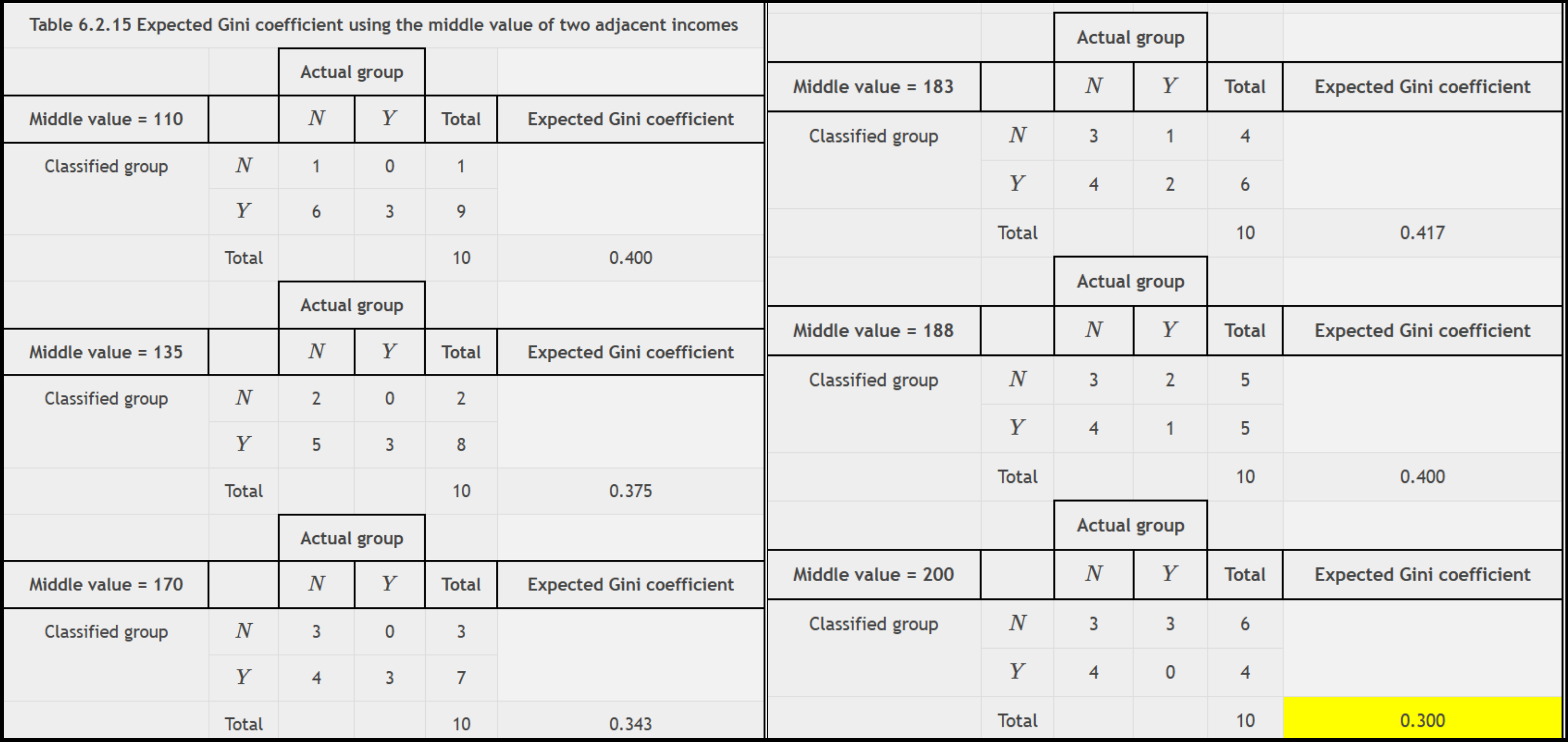


Table 6.2.15 Expected Gini coefficient using the middle value of two adjacent incomes

		Actual group			
Middle value = 110		N	Y	Total	Expected Gini coefficient
Classified group	N	1	0	1	
	Y	6	3	9	
	Total			10	0.400

		Actual group			
Middle value = 135		N	Y	Total	Expected Gini coefficient
Classified group	N	2	0	2	
	Y	5	3	8	
	Total			10	0.375

		Actual group			
Middle value = 170		N	Y	Total	Expected Gini coefficient
Classified group	N	3	0	3	
	Y	4	3	7	
	Total			10	0.343

		Actual group			
Middle value = 183		N	Y	Total	Expected Gini coefficient
Classified group	N	3	1	4	
	Y	4	2	6	
	Total			10	0.417

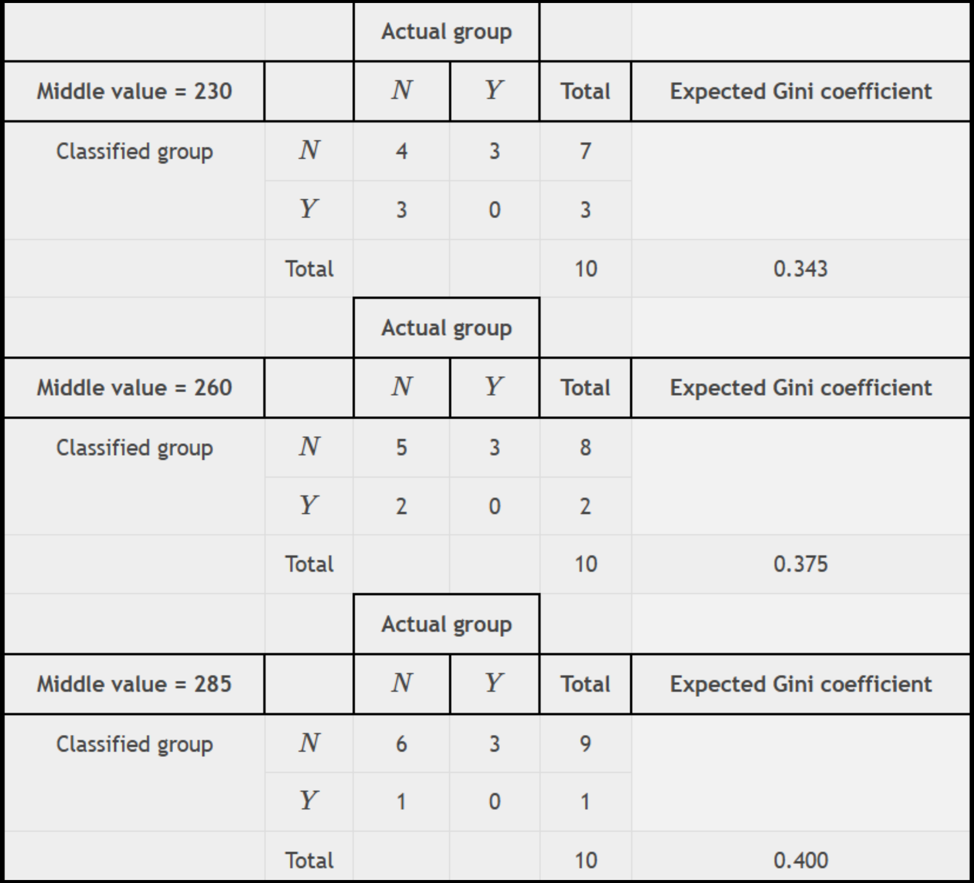
		Actual group			
Middle value = 188		N	Y	Total	Expected Gini coefficient
Classified group	N	3	2	5	
	Y	4	1	5	
	Total			10	0.400

		Actual group			
Middle value = 200		N	Y	Total	Expected Gini coefficient
Classified group	N	3	3	6	
	Y	4	0	4	
	Total			10	0.300

6.2 Decision tree model

❖ Categorization of a continuous variable

<Answer of Example 6.2.6>



		Actual group			
Middle value = 230		N	Y	Total	Expected Gini coefficient
Classified group	N	4	3	7	
	Y	3	0	3	
	Total			10	0.343
		Actual group			
Middle value = 260		N	Y	Total	Expected Gini coefficient
Classified group	N	5	3	8	
	Y	2	0	2	
	Total			10	0.375
		Actual group			
Middle value = 285		N	Y	Total	Expected Gini coefficient
Classified group	N	6	3	9	
	Y	1	0	1	
	Total			10	0.400

6.2 Decision tree model

❖ Overfitting and pruning decision tree

- Decision tree models can have an overfitting problem;
  - Classifying training data well but not good for testing data.
- Pruning is one way to solve the problem of overfitting.
  - Pre-pruning is to examine the appropriateness of the division using chi-square tests and information gain
  - A threshold value must be set, determined by researcher.
  - Post-pruning is a method of removing branches from a completed tree.
  - Pre-pruning and post-pruning are sometimes used in combination.

6.2 Decision tree model

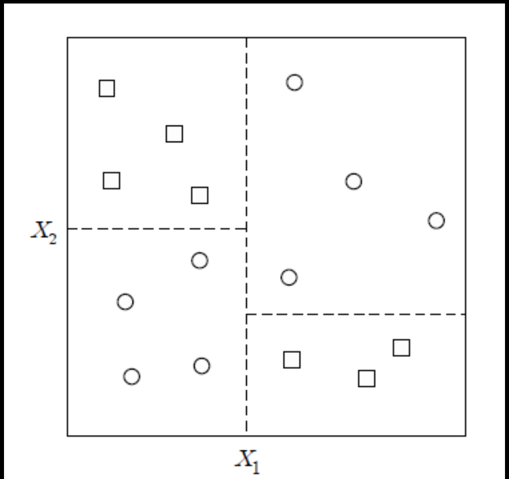
❖ Characteristics of decision tree

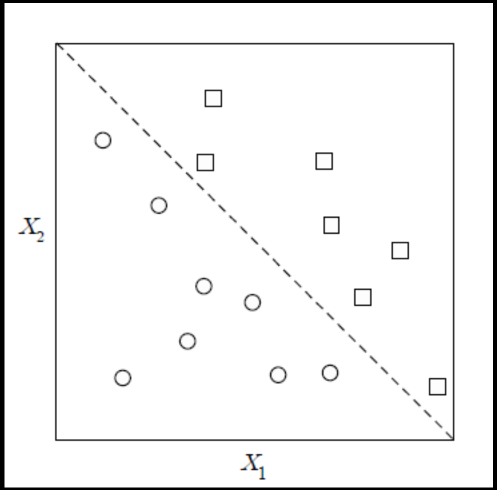
- The decision tree model is a nonparametric method that does not assume the distribution function of each group.
- Decision trees are easy to explain to anyone.
- Since the calculations are not complicated, they can be applied to large amounts of data.
- They are not very sensitive to noisy data.
- Since they use a local optimization search algorithm, they may not find the overall optimal decision tree.
- Similar small trees may appear repeatedly.

6.2 Decision tree model

❖ Characteristics of decision tree

- A decision tree examines only the conditions for one variable in one node. Therefore, the classification rule of the decision tree partitions the entire decision space into straight lines parallel to the coordinate axes





- Not easy to split by a decision tree.

6.3 Naive Bayes classification model

❖ Bayes classification model

- When the **prior probability** of being classified into each group and the **likelihood probability** of each group are known, the **Bayes classification model** is a method of classifying data into groups with high probability by calculating the posterior probability using Bayes theorem,

6.3 Naive Bayes classification model

❖ Bayes classification model

[Example 6.3.1] (Classification by prior probability)
When 20 customers who visited a computer store were surveyed, 8 customers purchased a computer, and 12 customers did not purchase a computer. Based on this information, classify a customer who visited this store on a day whether he would purchase a computer or not.

<**Answer**>

- Among the 20 customers who visited the store, only 8 (40%) purchased a computer, and 12 (60%) did not purchase a computer.
- The probability of the purchasing group is 40% and non-purchasing group is 60% are called **prior probabilities**.
- Since the probability of the non-purchasing group is higher than the purchasing group, it is reasonable to classify a customer into the non-purchasing group.

6.3 Naive Bayes classification model

❖ Bayes classification model

[Example 6.3.2] (Classification by posterior probability)
Suppose there are 10 customers in their age 20's among 20 customers and 10 customers in their age 30s. Among the 8 purchasing groups, 2 customers in their age 20s and 6 are in their age 30's. If a customer who visited the store on a day is in his age 20's, classify the customer whether he purchases the computer or not by calculating the posterior probability.

<**Answer**>

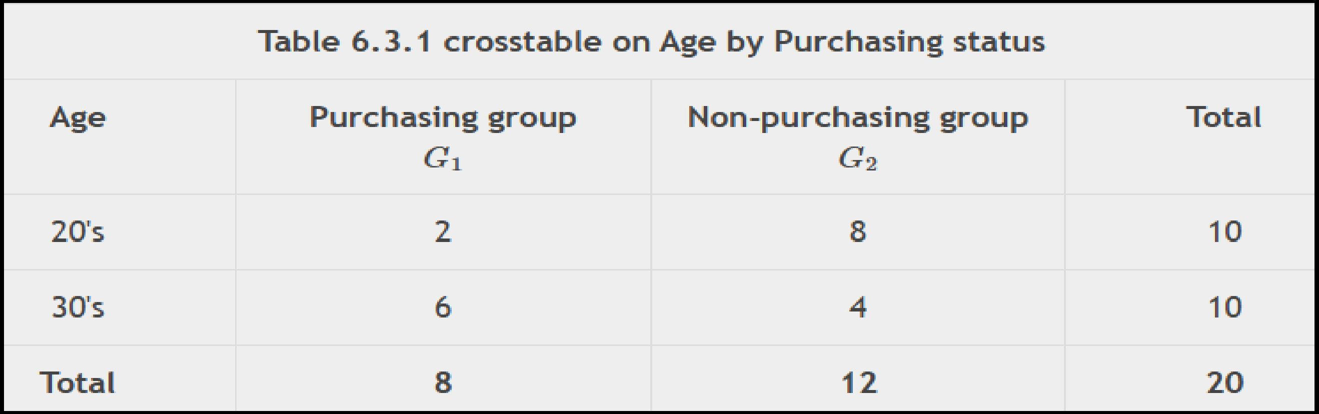


Table 6.3.1 crosstable on Age by Purchasing status

Age	Purchasing group G_1	Non-purchasing group G_2	Total
20's	2	8	10
30's	6	4	10
Total	8	12	20

6.3 Naive Bayes classification model

❖ Bayes classification model

<Answer of Example 6.3.2>

- Posterior probability;

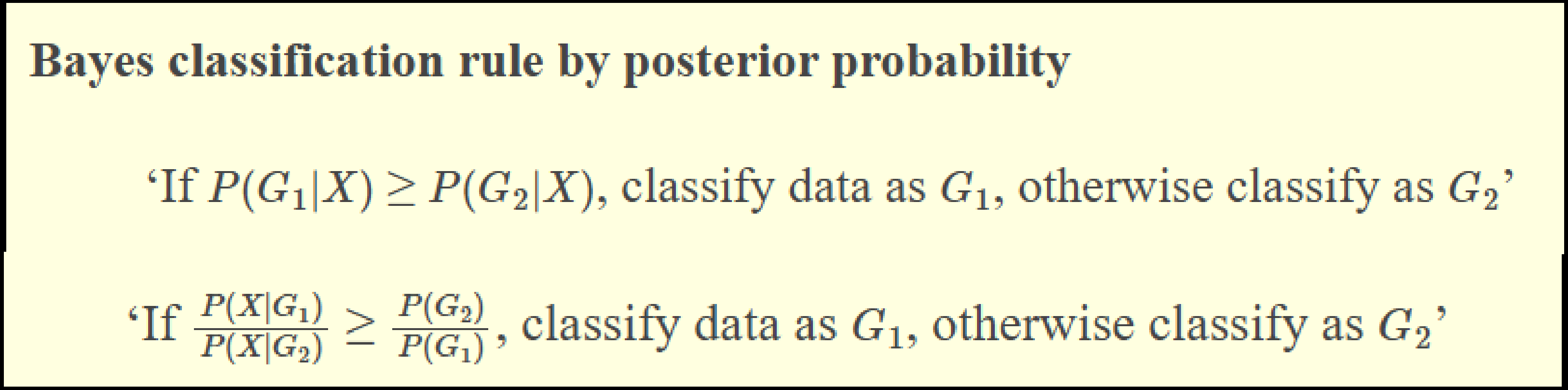
$$P(G_1|x) = \frac{P(G_1)\times P(x|G_1)}{P(G_1)\times P(x|G_1) + P(G_2)\times P(x|G_2)} = \frac{\frac{8}{20}\times\frac{2}{8}}{\frac{8}{20}\times\frac{2}{8}+\frac{12}{20}\times\frac{8}{12}} = 0.2$$

$$P(G_2|x) = \frac{P(G_2)\times P(x|G_2)}{P(G_1)\times P(x|G_1) + P(G_2)\times P(x|G_2)} = \frac{\frac{12}{20}\times\frac{8}{12}}{\frac{8}{20}\times\frac{2}{8}+\frac{12}{20}\times\frac{8}{12}} = 0.8$$

- Posterior probability that a customer belongs to non-purchasing group is 0.8, which is higher than purchasing group 0.2, this customer is classified as a non-purchasing group.

6.3 Naive Bayes classification model

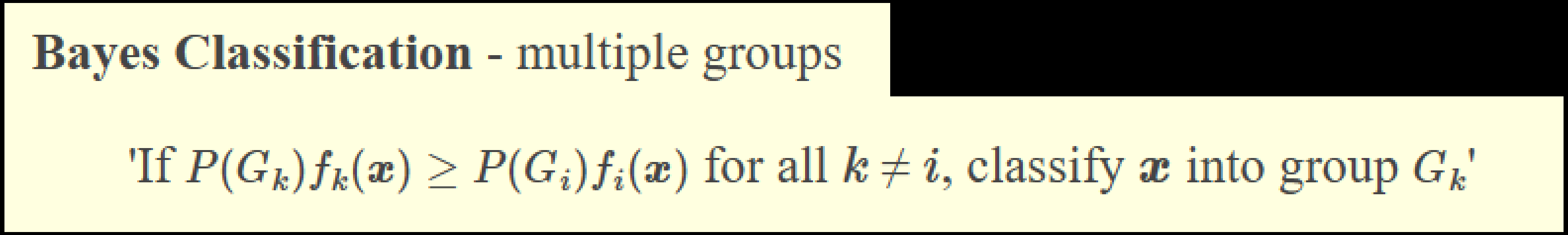
❖ Bayes classification model



Bayes classification rule by posterior probability

‘If $P(G_1|X) \geq P(G_2|X)$, classify data as G_1, otherwise classify as G_2’

‘If $\frac{P(X|G_1)}{P(X|G_2)} \geq \frac{P(G_2)}{P(G_1)}$, classify data as G_1, otherwise classify as G_2’



Bayes Classification - multiple groups

'If $P(G_k)f_k(\boldsymbol{x}) \geq P(G_i)f_i(\boldsymbol{x})$ for all $k \neq i$, classify $\boldsymbol{x}$ into group G_k'

6.3 Naive Bayes classification model

❖ Naive Bayes classification model for categorical data

[Example 6.3.3]

Consider a survey of 20 customers at a computer store on age (X1), monthly income (X2), credit status (X3) and their purchasing status as shown in Table 6.3.2. Note that the age transformed into '20s' and '30s' as categorical, income into 'LT2000' and 'GE2000', and credit status into 'Bad', 'Fair' and 'Good'.

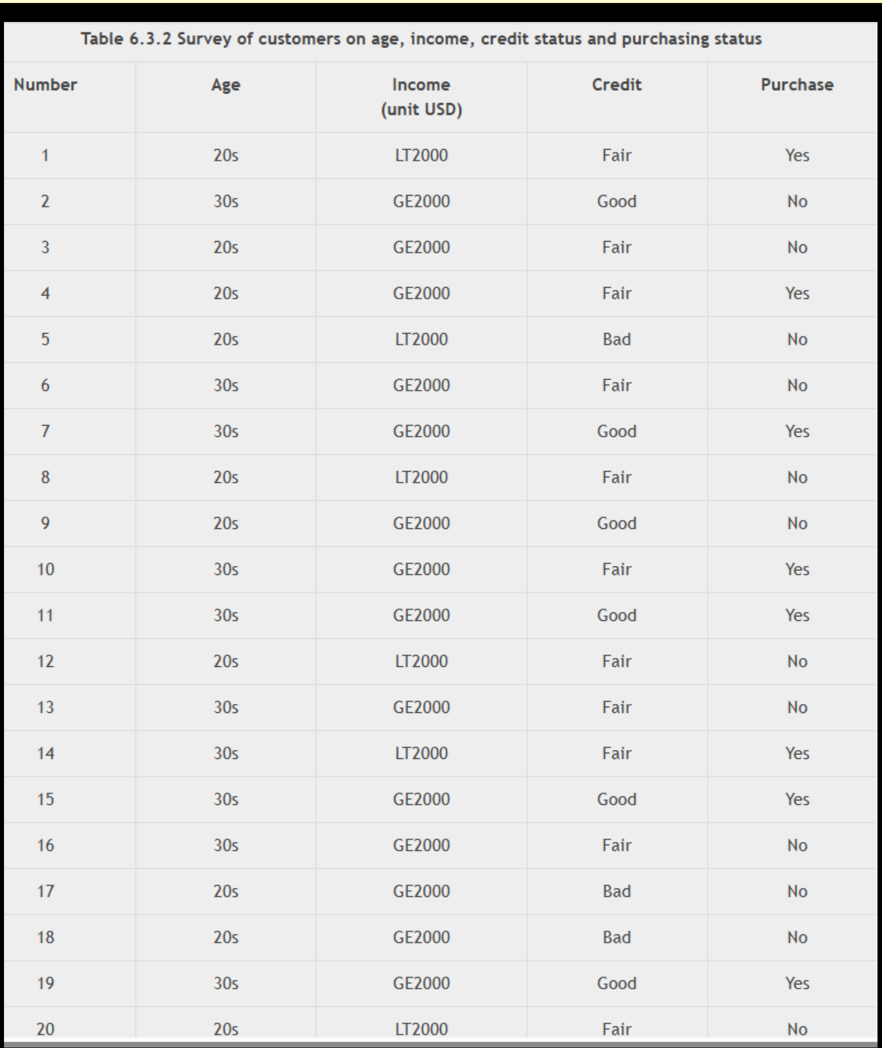


Table 6.3.2 Survey of customers on age, income, credit status and purchasing status

Number	Age	Income (unit USD)	Credit	Purchase
1	20s	LT2000	Fair	Yes
2	30s	GE2000	Good	No
3	20s	GE2000	Fair	No
4	20s	GE2000	Fair	Yes
5	20s	LT2000	Bad	No
6	30s	GE2000	Fair	No
7	30s	GE2000	Good	Yes
8	20s	LT2000	Fair	No
9	20s	GE2000	Good	No
10	30s	GE2000	Fair	Yes
11	30s	GE2000	Good	Yes
12	20s	LT2000	Fair	No
13	30s	GE2000	Fair	No
14	30s	LT2000	Fair	Yes
15	30s	GE2000	Good	Yes
16	30s	GE2000	Fair	No
17	20s	GE2000	Bad	No
18	20s	GE2000	Bad	No
19	30s	GE2000	Good	Yes
20	20s	LT2000	Fair	No

6.3 Naive Bayes classification model

❖ Naive Bayes classification model for categorical data

<Answer of Example 6.3.3>

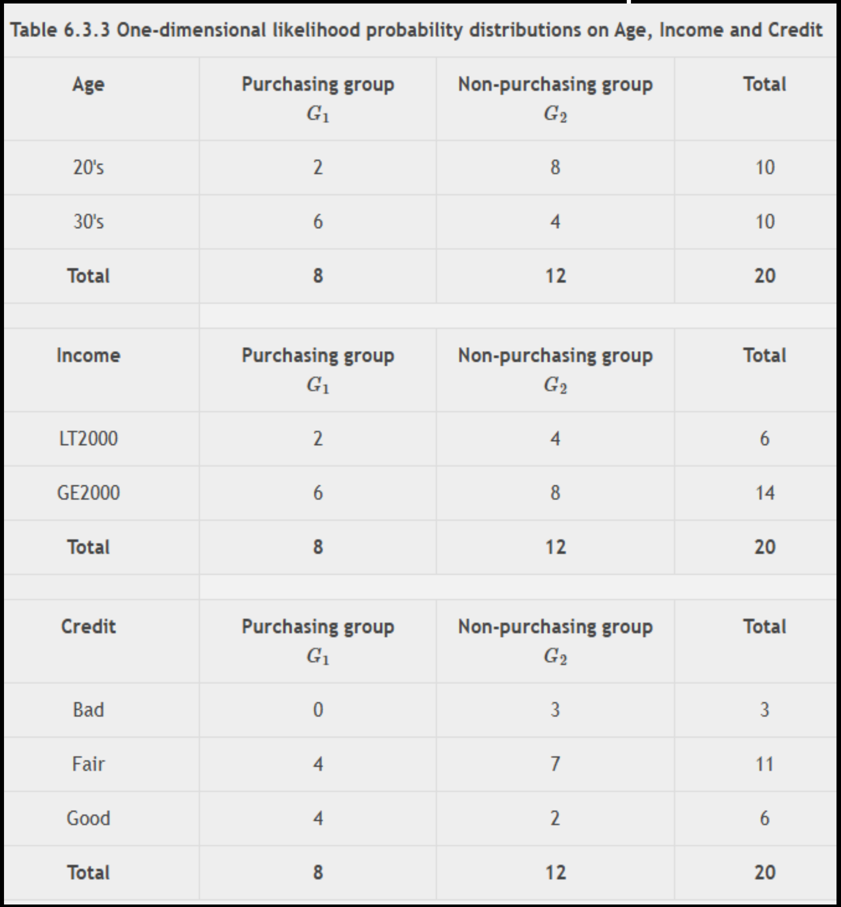


Table 6.3.3 One-dimensional likelihood probability distributions on Age, Income and Credit

Age	Purchasing group G_1	Non-purchasing group G_2	Total
20's	2	8	10
30's	6	4	10
Total	**8**	**12**	**20**

Income	Purchasing group G_1	Non-purchasing group G_2	Total
LT2000	2	4	6
GE2000	6	8	14
Total	**8**	**12**	**20**

Credit	Purchasing group G_1	Non-purchasing group G_2	Total
Bad	0	3	3
Fair	4	7	11
Good	4	2	6
Total	**8**	**12**	**20**

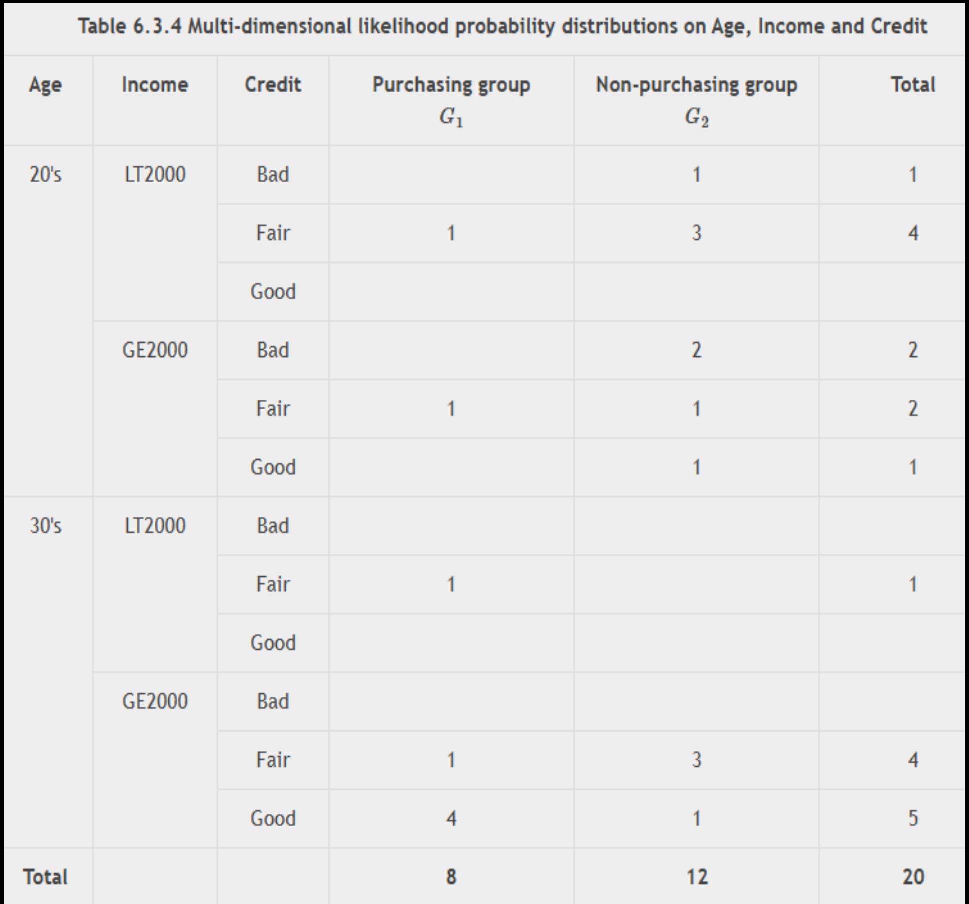


Table 6.3.4 Multi-dimensional likelihood probability distributions on Age, Income and Credit

Age	Income	Credit	Purchasing group G_1	Non-purchasing group G_2	Total
20's	LT2000	Bad		1	1
		Fair	1	3	4
		Good			
	GE2000	Bad		2	2
		Fair	1	1	2
		Good		1	1
30's	LT2000	Bad			
		Fair	1		1
		Good			
	GE2000	Bad			
		Fair	1	3	4
		Good	4	1	5
Total			**8**	**12**	**20**

6.3 Naive Bayes classification model

❖ Naive Bayes classification model for categorical data

<Answer of Example 6.3.3>

- if variables, age, income, and credit status, can be assumed to be independent, the one-dimensional likelihood probability distribution of each variable is used approximately to estimate the multidimensional likelihood probability distribution as follows.

$$P(\boldsymbol{X} = (X_1, X_2, X_3) \mid G_i) \approx P(X_1|G_i)\ P(X_2|G_i)\ P(X_3|G_i)$$

$$P(\boldsymbol{x} = (30s, LT2000, Fair) \mid G_1) \approx \tfrac{6}{8} \times \tfrac{2}{8} \times \tfrac{4}{8} = 0.0938$$

$$P(\boldsymbol{x} = (30s, LT2000, Fair) \mid G_2) \approx \tfrac{4}{12} \times \tfrac{4}{12} \times \tfrac{7}{12} = 0.0648$$

- Posterior probability

$$\begin{aligned} P(G_1|\boldsymbol{x}) &= \frac{P(G_1) \times P(\boldsymbol{x}|G_1)}{P(G_1) \times P(\boldsymbol{x}|G_1) + P(G_2) \times P(\boldsymbol{x}|G_2)} \\ &= \frac{0.4 \times 0.0938}{0.4 \times 0.0938 + 0.6 \times 0.0648} = 0.4911 \\ P(G_2|\boldsymbol{x}) &= \frac{P(G_2) \times P(\boldsymbol{x}|G_2)}{P(G_1) \times P(\boldsymbol{x}|G_1) + P(G_2) \times P(\boldsymbol{x}|G_2)} \\ &= \frac{0.6 \times 0.0648}{0.4 \times 0.0938 + 0.6 \times 0.0648} = 0.5089 \end{aligned}$$

6.3 Naive Bayes classification model

❖ Stepwise variable selection

- Naive Bayes classification can be applied by categorizing the continuous variables.
- **Forward selection** selects a variable with the highest information gain using the uncertainty measures.
  - Add another variable with next highest information gain
- **Backward elimination** includes all variables and selects a variable to remove which improve classification accuracy.
  - Continue removing a variable until there is no improvement in classification accuracy.
- A stepwise method selects variables using the forward selection method while examining whether the variables

6.4 Evaluation and comparison of the classification model

❖ Evaluation of a classification model

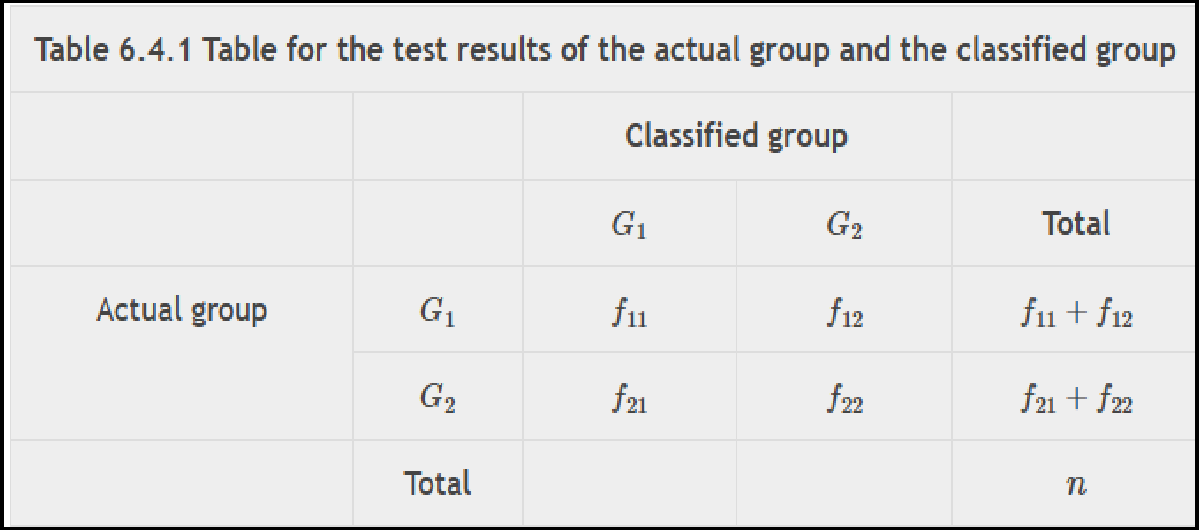
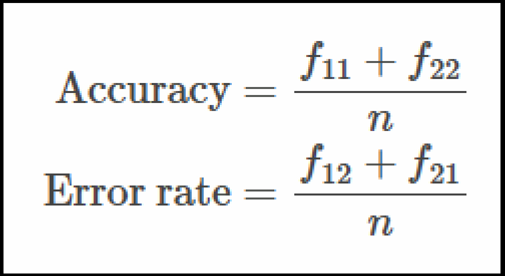


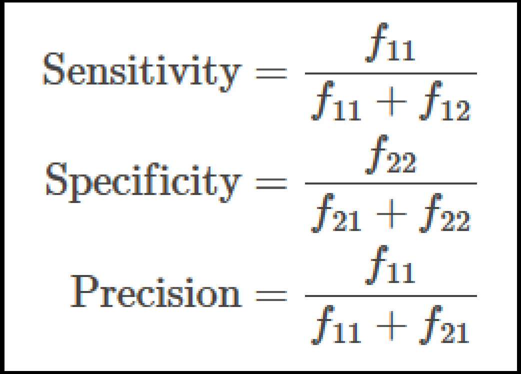
Table 6.4.1 Table for the test results of the actual group and the classified group

		Classified group		
		G_1	G_2	Total
Actual group	G_1	f_{11}	f_{12}	$f_{11}+f_{12}$
	G_2	f_{21}	f_{22}	$f_{21}+f_{22}$
	Total			n



$$\text{Accuracy} = \frac{f_{11}+f_{22}}{n}$$

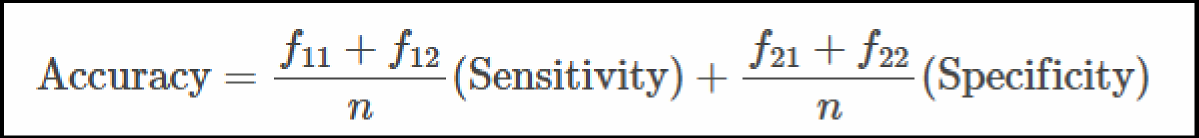
$$\text{Error rate} = \frac{f_{12}+f_{21}}{n}$$



$$\text{Sensitivity} = \frac{f_{11}}{f_{11}+f_{12}}$$

$$\text{Specificity} = \frac{f_{22}}{f_{21}+f_{22}}$$

$$\text{Precision} = \frac{f_{11}}{f_{11}+f_{21}}$$



$$\text{Accuracy} = \frac{f_{11}+f_{12}}{n}(\text{Sensitivity}) + \frac{f_{21}+f_{22}}{n}(\text{Specificity})$$

6.4 Evaluation and comparison of the classification model

❖ Lift chart

- Assume that results of a classification model are expressed as continuous values, posterior probability.
  - large values means a high probability being classified as group 1.
- Rate of group 1 among the entire data is called the **baseline response** (%).
- Arrange data in descending order of posterior probability, observe the top 10% of the data. Rate of classifying actual group 1, as group 1 will be high. Upper 10% response.
- Upper 10% response rate compared to the baseline response rate is called the **lift** or **improvement** of top 10%

6.4 Evaluation and comparison of the classification model

❖ Lift chart

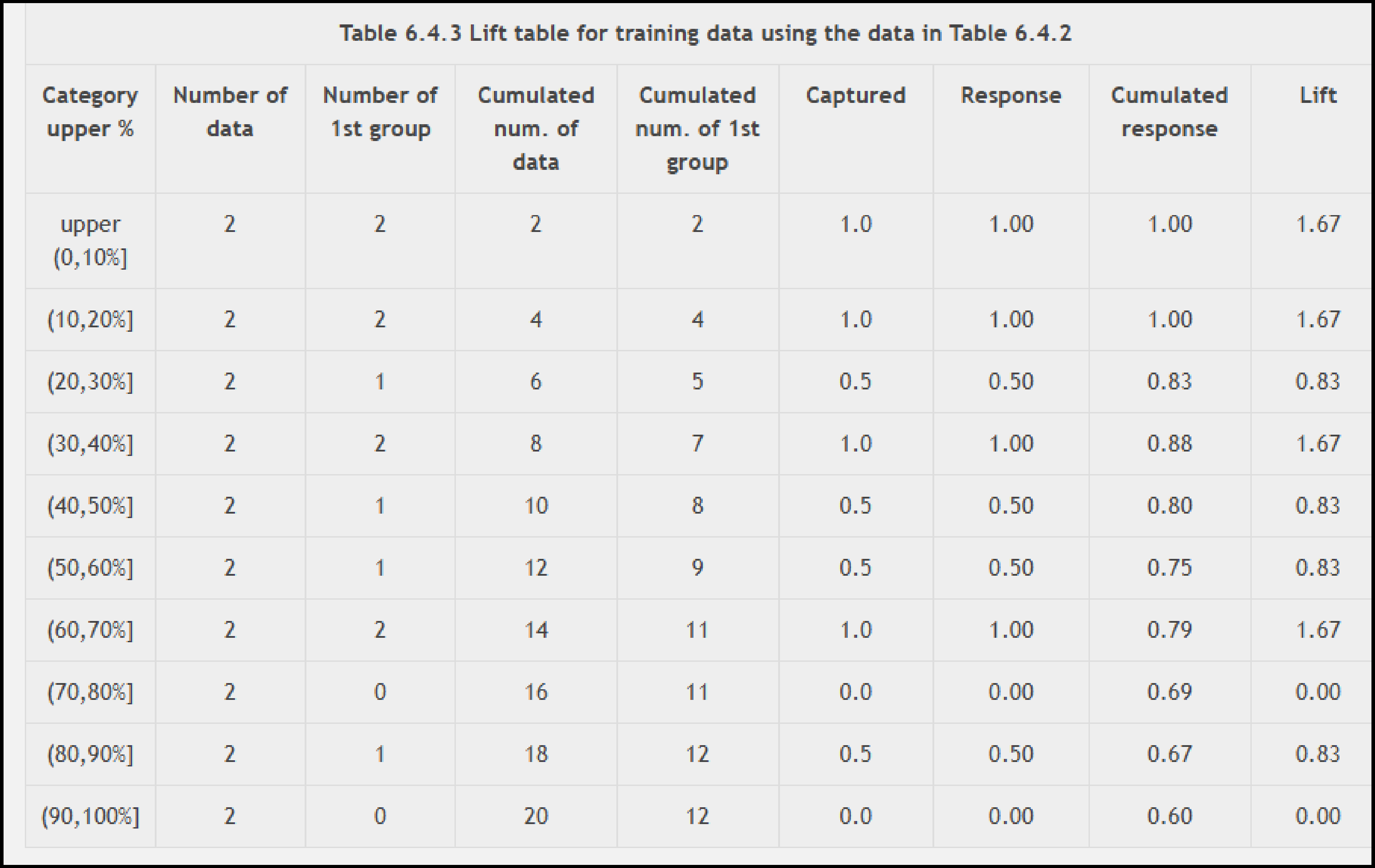
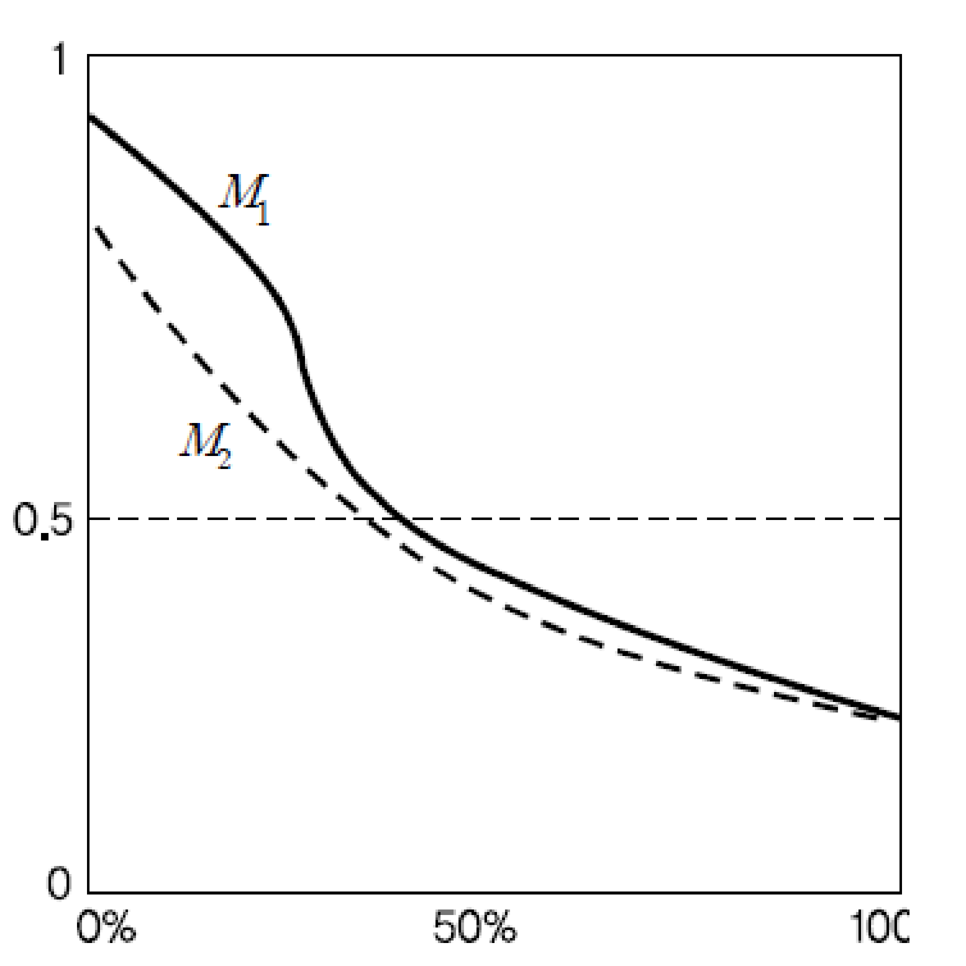


Table 6.4.3 Lift table for training data using the data in Table 6.4.2

Category upper %	Number of data	Number of 1st group	Cumulated num. of data	Cumulated num. of 1st group	Captured	Response	Cumulated response	Lift
upper (0,10%]	2	2	2	2	1.0	1.00	1.00	1.67
(10,20%]	2	2	4	4	1.0	1.00	1.00	1.67
(20,30%]	2	1	6	5	0.5	0.50	0.83	0.83
(30,40%]	2	2	8	7	1.0	1.00	0.88	1.67
(40,50%]	2	1	10	8	0.5	0.50	0.80	0.83
(50,60%]	2	1	12	9	0.5	0.50	0.75	0.83
(60,70%]	2	2	14	11	1.0	1.00	0.79	1.67
(70,80%]	2	0	16	11	0.0	0.00	0.69	0.00
(80,90%]	2	1	18	12	0.5	0.50	0.67	0.83
(90,100%]	2	0	20	12	0.0	0.00	0.60	0.00

6.4 Evaluation and comparison of the classification model

❖ Lift chart



- Since model M1 classifies group 1 more accurately than model M2 at each percentile of data, so model M1 can be said to be better than M2.

6.4 Evaluation and comparison of the classification model

❖ Confusion matrix

- Confusion matrix is often used to determine the cut-off value of the posterior probability to determine the group.
- Generally, the cut-off value for determining the group is mainly accuracy, but sensitivity and specificity should also be carefully examined.

6.4 Evaluation and comparison of the classification model

❖ Confusion matrix

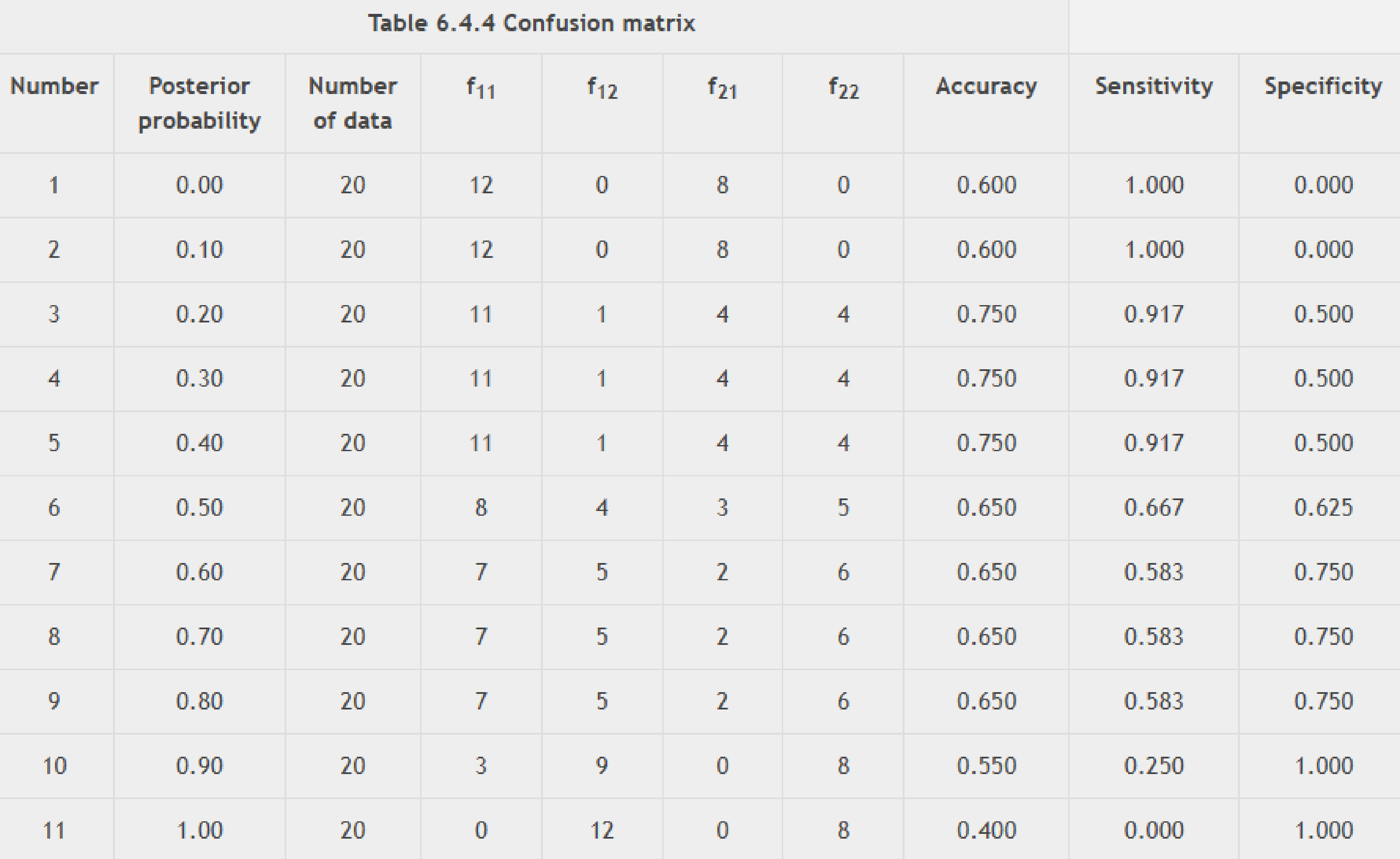


Table 6.4.4 Confusion matrix

Number	Posterior probability	Number of data	f_{11}	f_{12}	f_{21}	f_{22}	Accuracy	Sensitivity	Specificity
1	0.00	20	12	0	8	0	0.600	1.000	0.000
2	0.10	20	12	0	8	0	0.600	1.000	0.000
3	0.20	20	11	1	4	4	0.750	0.917	0.500
4	0.30	20	11	1	4	4	0.750	0.917	0.500
5	0.40	20	11	1	4	4	0.750	0.917	0.500
6	0.50	20	8	4	3	5	0.650	0.667	0.625
7	0.60	20	7	5	2	6	0.650	0.583	0.750
8	0.70	20	7	5	2	6	0.650	0.583	0.750
9	0.80	20	7	5	2	6	0.650	0.583	0.750
10	0.90	20	3	9	0	8	0.550	0.250	1.000
11	1.00	20	0	12	0	8	0.400	0.000	1.000

6.4 Evaluation and comparison of the classification model

❖ ROC graph

- ROC graph;

  (1 – specificity) on x-axis and (sensitivity) on y-axis

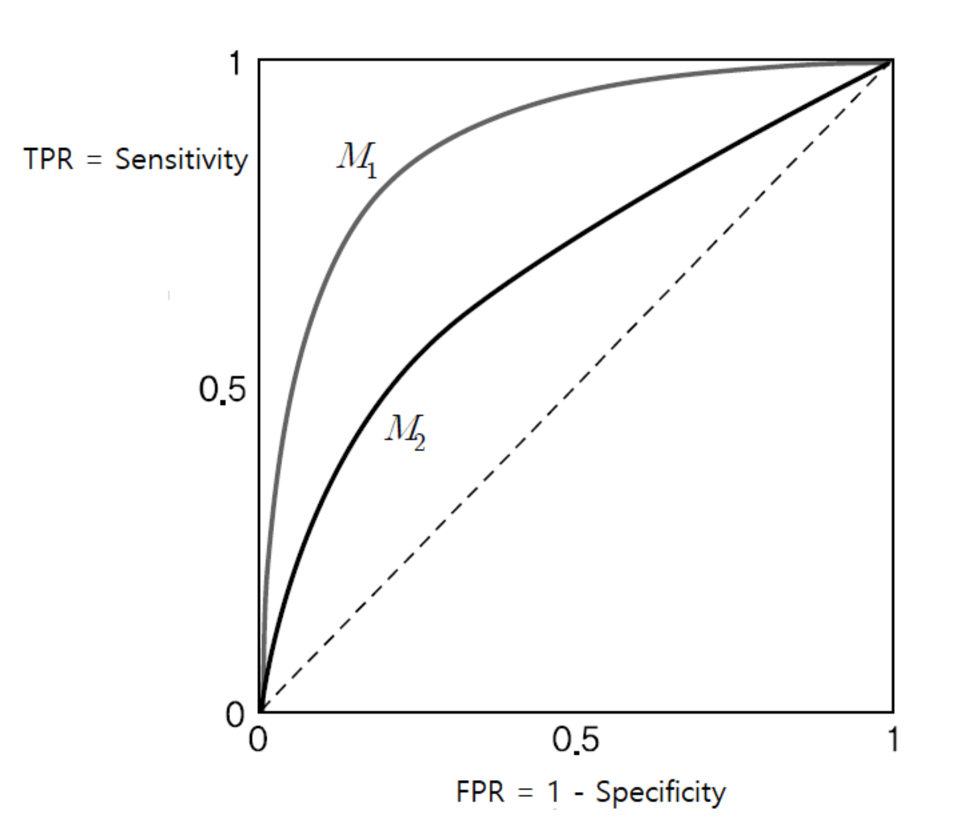
  (Sensitivity): true positive rate (TPR)

  (1 - specificity): false positive rate (FPR)

- A point on the ROC graph represents the result of a classification model when a critical value of the posterior probability is used.
- ROC graph examines the changes in FPR and TPR when we change the critical value.

6.4 Evaluation and comparison of the classification model

❖ ROC graph



- (FPR=0, TPR=0) classifies all points into group 2,
- (FPR=1, TPR=1) classifies all data into group 1
- (FPR=0, TPR=1) does not misclassify group 2 into group 1 and correctly classifies all group 1 data.
- Ggod classification model should have the classification results located in the upper left corner of the ROC graph.
- Diagonal line shows an exceptional model in which both the TPR and FPR ratios are the same, group 1 data is classified into group 1 with probability p (TPR = p), and group 2 data is also classified into group 1 with probability p (FPR = p).

- Model M1 is located upper on the left than model M2. M1 is better than model M2.

6.4 Evaluation and comparison of the classification model

❖ ROC graph

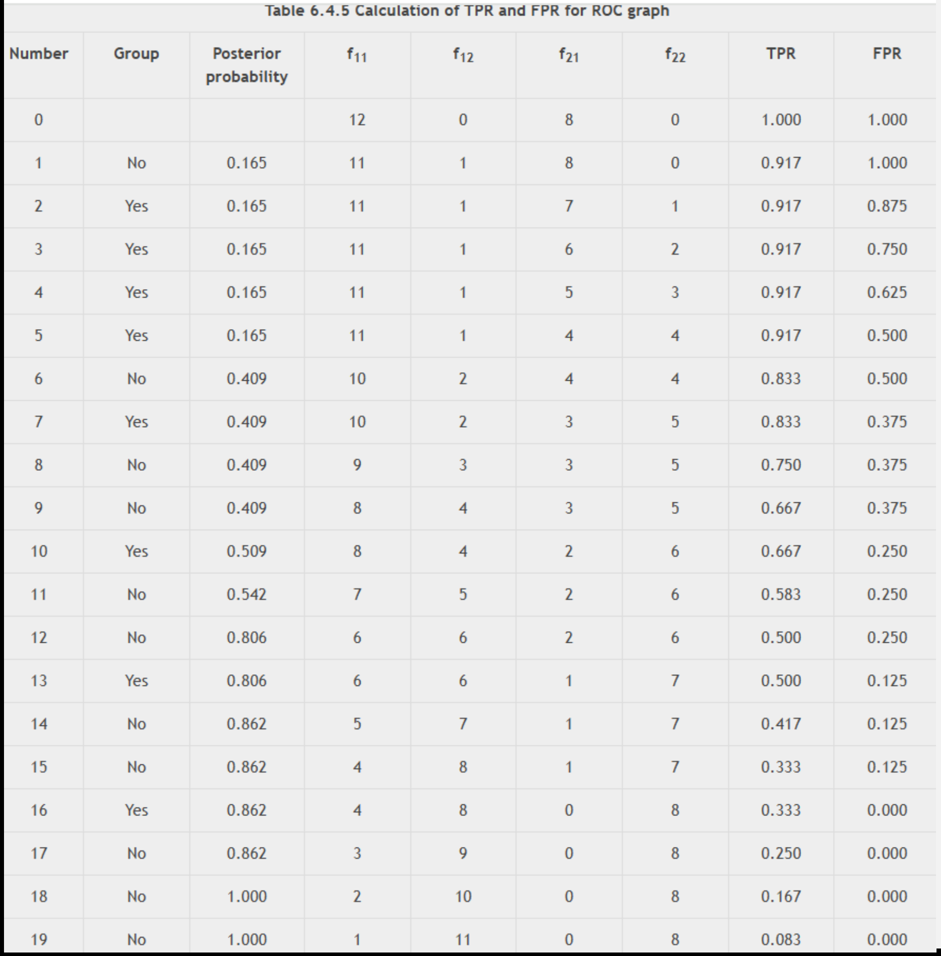
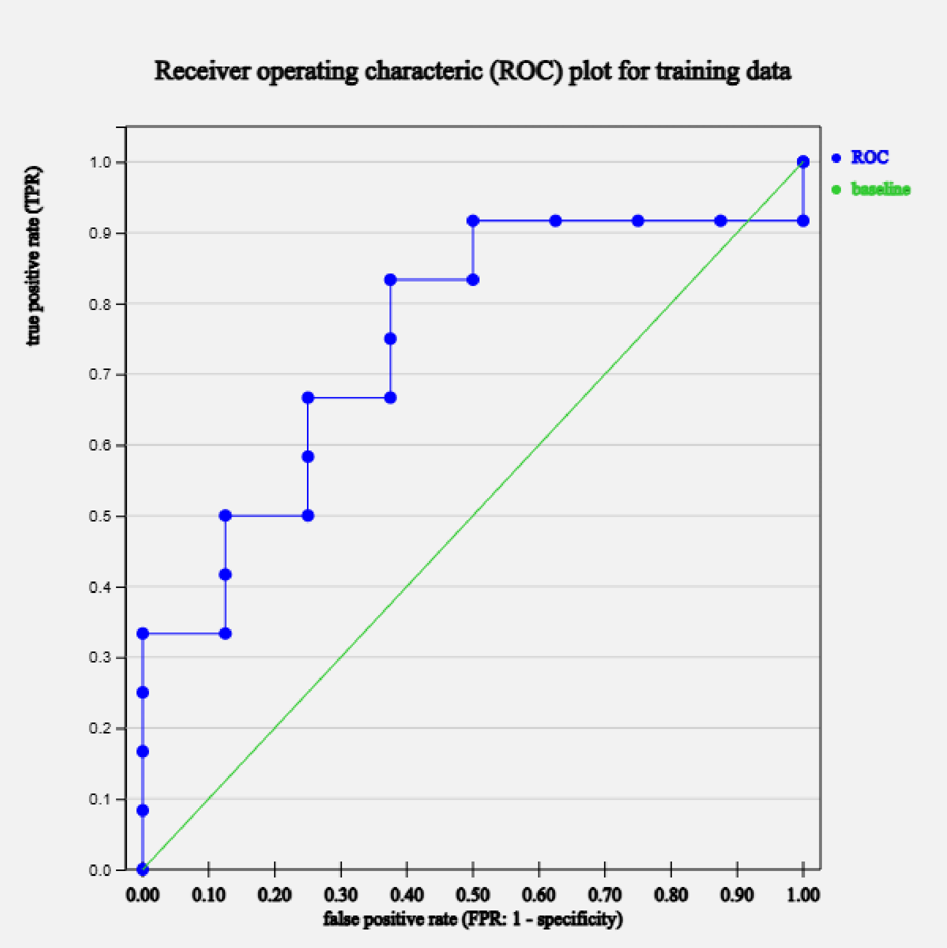


Table 6.4.5 Calculation of TPR and FPR for ROC graph

Number	Group	Posterior probability	f_{11}	f_{12}	f_{21}	f_{22}	TPR	FPR
0			12	0	8	0	1.000	1.000
1	No	0.165	11	1	8	0	0.917	1.000
2	Yes	0.165	11	1	7	1	0.917	0.875
3	Yes	0.165	11	1	6	2	0.917	0.750
4	Yes	0.165	11	1	5	3	0.917	0.625
5	Yes	0.165	11	1	4	4	0.917	0.500
6	No	0.409	10	2	4	4	0.833	0.500
7	Yes	0.409	10	2	3	5	0.833	0.375
8	No	0.409	9	3	3	5	0.750	0.375
9	No	0.409	8	4	3	5	0.667	0.375
10	Yes	0.509	8	4	2	6	0.667	0.250
11	No	0.542	7	5	2	6	0.583	0.250
12	No	0.806	6	6	2	6	0.500	0.250
13	Yes	0.806	6	6	1	7	0.500	0.125
14	No	0.862	5	7	1	7	0.417	0.125
15	No	0.862	4	8	1	7	0.333	0.125
16	Yes	0.862	4	8	0	8	0.333	0.000
17	No	0.862	3	9	0	8	0.250	0.000
18	No	1.000	2	10	0	8	0.167	0.000
19	No	1.000	1	11	0	8	0.083	0.000



Receiver operating characteric (ROC) plot for training data

6.4 Evaluation and comparison of the classification model

❖ Comparison of classification models

- Confidence interval for accuracy

$$P(-Z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{p(1-p)}} < Z_{\frac{\alpha}{2}}) = 1-\alpha$$

$$\frac{(2n \times \hat{p} + Z^2_{\frac{\alpha}{2}}) \pm Z_{\frac{\alpha}{2}}\sqrt{Z^2_{\frac{\alpha}{2}} + 4n \times \hat{p} - 4n \times \hat{p}^2}}{2n + 2Z^2_{\frac{\alpha}{2}}}$$

6.4 Evaluation and comparison of the classification model

❖ Comparison of classification models

- Comparison of two classification models

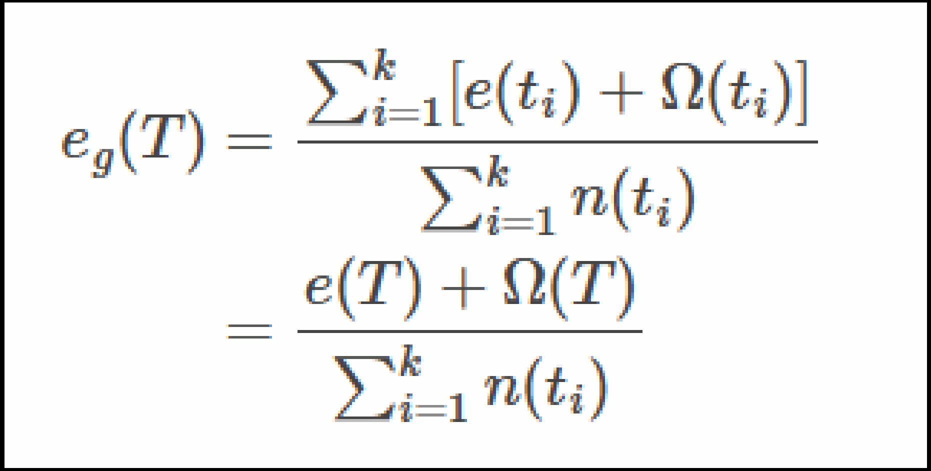
  - confidence interval of the accuracy difference

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

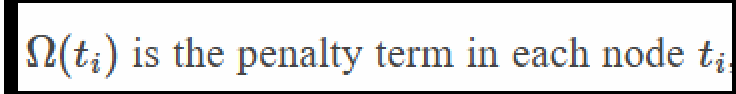
6.4 Evaluation and comparison of the classification model

❖ Comparison of classification models

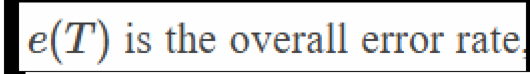
- Generalized erro considering model overfitting



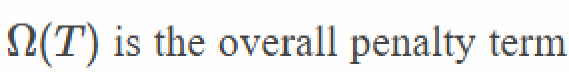
$$e_g(T) = \frac{\sum_{i=1}^{k}[e(t_i) + \Omega(t_i)]}{\sum_{i=1}^{k} n(t_i)}$$
$$= \frac{e(T) + \Omega(T)}{\sum_{i=1}^{k} n(t_i)}$$



$\Omega(t_i)$ is the penalty term in each node t_i,



$e(T)$ is the overall error rate,



$\Omega(T)$ is the overall penalty term

Summary

- Probability:
  - Classical and relative frequency definition
  - Addition and multiplication rules
  - Bayes theorem
- Random variable and probability distribution:
  - Random variable is a function from a sample space to real number
  - Binomial distribution, normal distribution
  - Multiivariate normal distribution
- Estimation of a distribution:
  - Maximum likelihood estimation

Thank you !!!