Technology

Science

Information

Networks

Computing

Lecturer: Ting Wang (王挺)

利物浦大学计算机博士

清华大学计算机博士后

电子信息技术高级工程师

上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任

Homework

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1. What is the mean of the data? What is the median?
2. What is the mode of the data? Comment on the data’ s modality (i.e., bimodal, trimodal, etc.).
3. What is the midrange of the data?
4. Can you find (roughly) the first quartile (Q1) and the third quartile(Q3) of the data?
5. Give the five-number summary of the data.
6. Show a boxplot of the data.
7. How is a quantile-quantile plot different from a quantile plot?

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1. What is the mean of the data? What is the median?

Mean：

(13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70)/27=**29.96**

Excel: =AVERAGE(A1:A27)

Median： Q2=**25**

Excel: =MEDIAN(A1:A27)

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

2. What is the mode of the data? Comment on the data’ s modality (i.e., bimodal, trimodal, etc.).

25 and **35**, **bimodal**

Excel: =MODE(A1:A27)

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

3. What is the midrange of the data?

midrange=(70+13)/2=**41.5**

Excel: =AVERAGE(A27,A1)

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

4. Can you find (roughly) the first quartile (Q1) and the third quartile(Q3) of the data?

Q1: 25% (N+1)/4=(27+1)/4=7, thus, Q1= **20** **Excel: =QUARTILE(A1:A27,1)**

Q3: 75% 3×(N+1)/4=21 , thus, Q3=**35** **Excel: =QUARTILE(A1:A27,3)**

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

5. Give the five-number summary of the data.

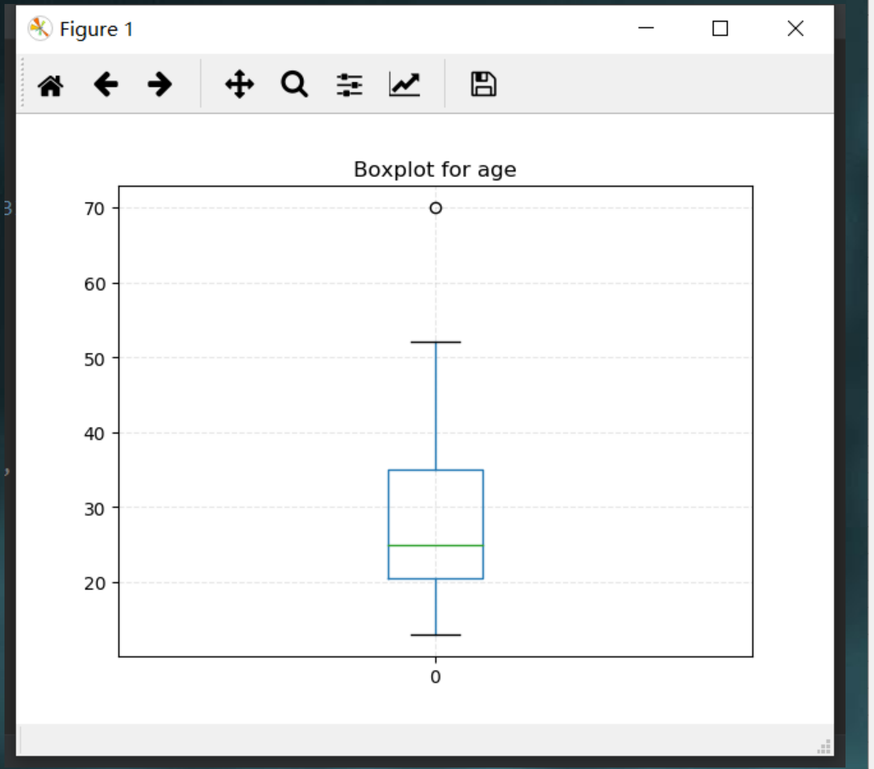
Min, Q1, Median, Q3, Max

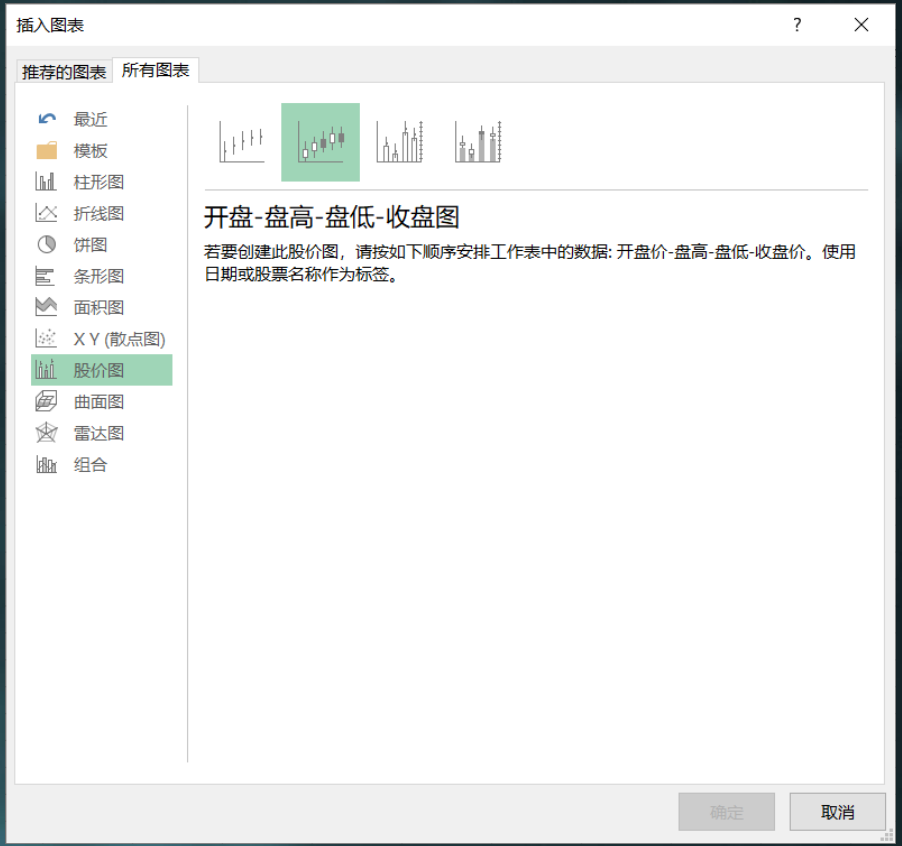
13、20、25、35、70

Problem 1

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

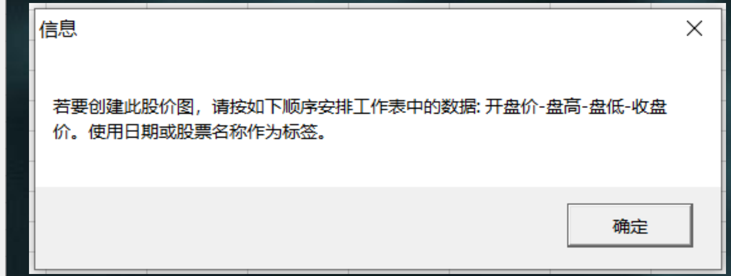
6. Show a boxplot of the data.



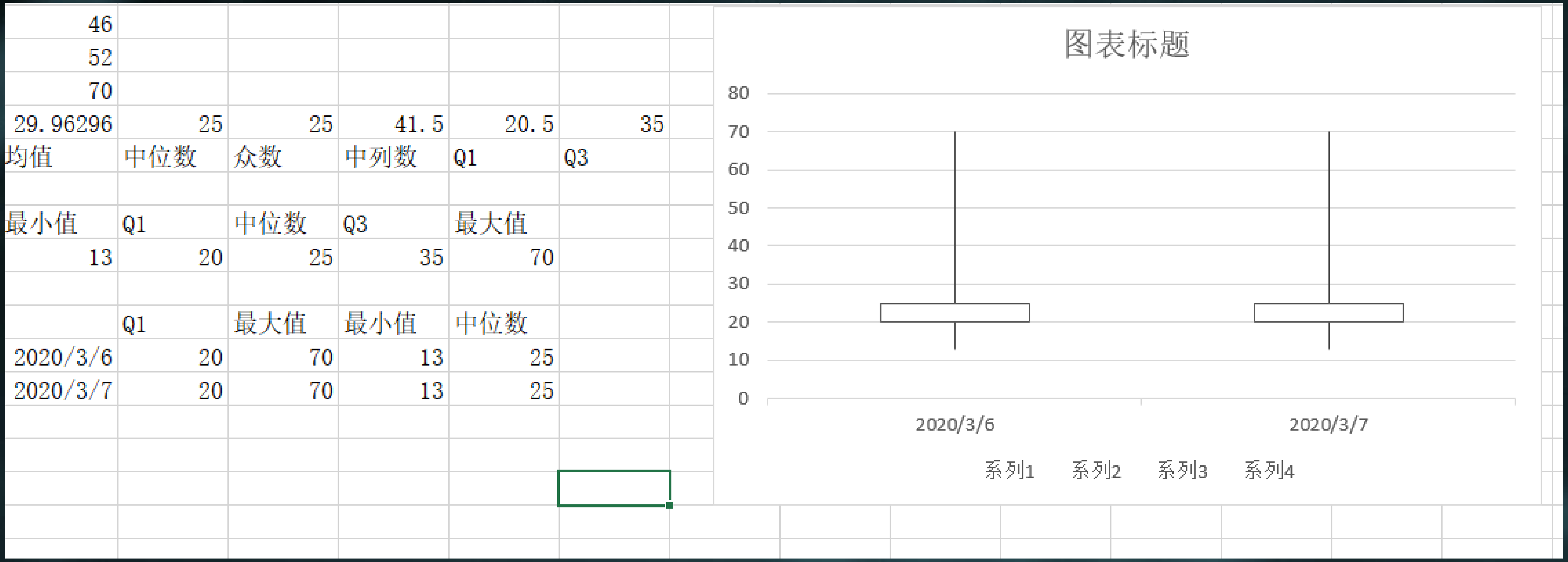


开盘-盘高-盘低-收盘图

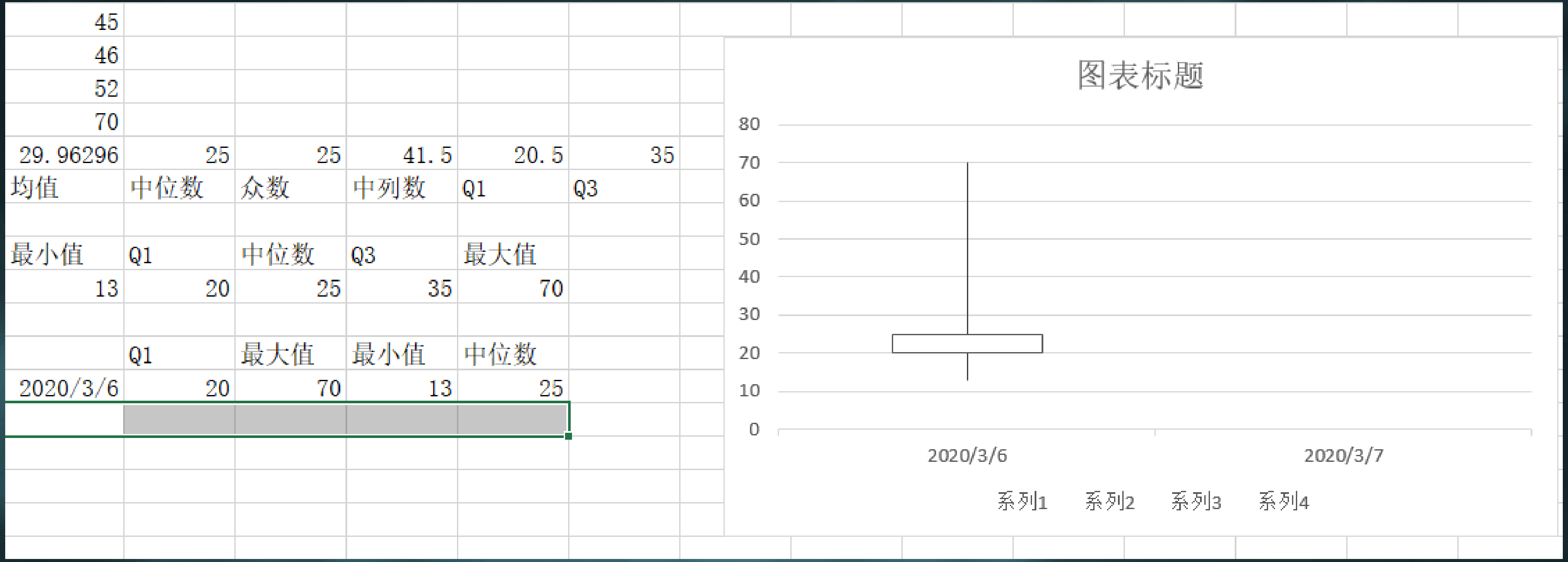
若要创建此股价图，请按如下顺序安排工作表中的数据: 开盘价-盘高-盘低-收盘价。使用日期或股票名称作为标签。



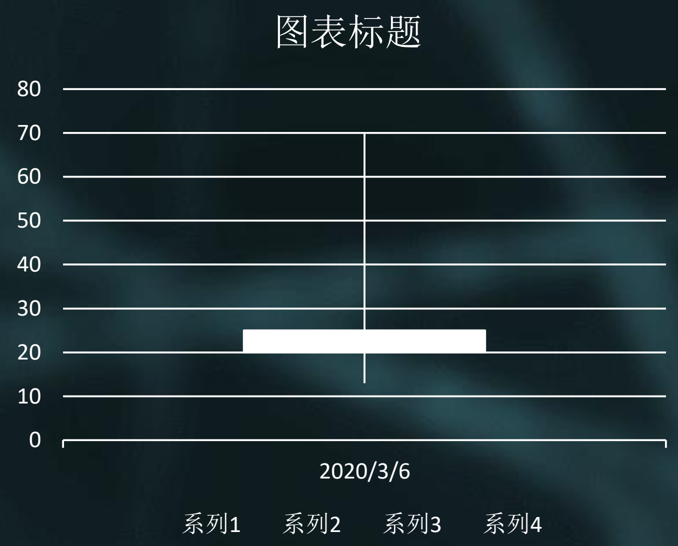
若要创建此股价图，请按如下顺序安排工作表中的数据: 开盘价-盘高-盘低-收盘价。使用日期或股票名称作为标签。



46					
52					
70					
29.96296	25	25	41.5	20.5	35
均值	中位数	众数	中列数	Q1	Q3
最小值	Q1	中位数	Q3	最大值	
13	20	25	35	70	
	Q1	最大值	最小值	中位数	
2020/3/6	20	70	13	25	
2020/3/7	20	70	13	25	



45					
46					
52					
70					
29.96296	25	25	41.5	20.5	35
均值	中位数	众数	中列数	Q1	Q3
最小值	Q1	中位数	Q3	最大值	
13	20	25	35	70	
	Q1	最大值	最小值	中位数	
2020/3/6	20	70	13	25	



Problem 1

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

7. How is a quantile-quantile plot different from a quantile plot?

A quantile plot displays a univariate data distribution.

A quantile–quantile plot shows the quantiles of two variables plotted against each other.

分位数图是一种观察单变量数据分布的简单有效方法；

而分位数-分位数图或q-q图对着另一个对应的分位数，绘制一个单变量分布的分位数。它是一种强有力的可视化工具，使得用户可以观察从一个分布到另一个分布是否有漂移；

Problem 2

A database has five transactions. Let $min_{sup} = 60\%$ and $min_{conf} = 80\%.$

TID	Items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(1) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

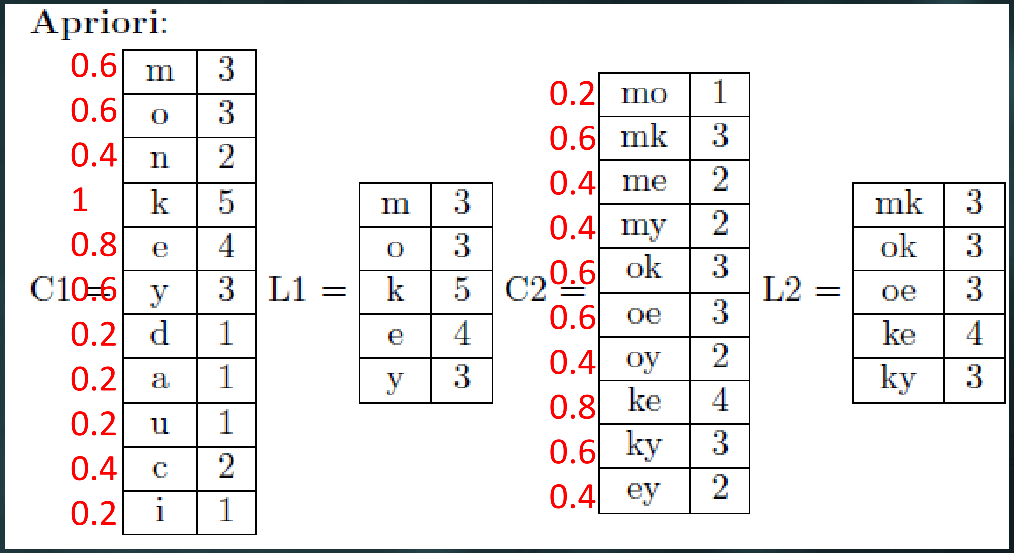
(2) List all of the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., “A” , “B” , etc.):

$\forall x \in$ transaction, $buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$ [s, c]

TID	Items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

min$_{sup}$ *= 60%* and *min*$_{conf}$ *= 80%.*

Apriori



Apriori:

C1 =

0.6	m	3
0.6	o	3
0.4	n	2
1	k	5
0.8	e	4
0.6	y	3
0.2	d	1
0.2	a	1
0.2	u	1
0.4	c	2
0.2	i	1

L1 =

m	3
o	3
k	5
e	4
y	3

C2 =

0.2	mo	1
0.6	mk	3
0.4	me	2
0.4	my	2
0.6	ok	3
0.6	oe	3
0.4	oy	2
0.8	ke	4
0.6	ky	3
0.4	ey	2

L2 =

mk	3
ok	3
oe	3
ke	4
ky	3

C3=

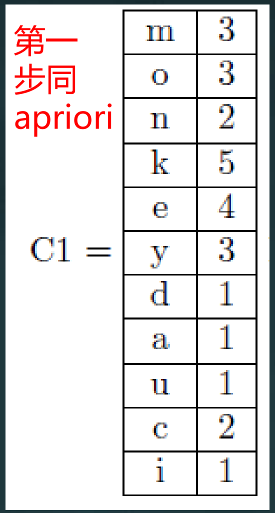
mok	1
mke	2
mky	2
oke	3
oky	2
key	2

L3=

oke	3

TID	Items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

min_{sup} = 60% and min_{conf} = 80%.



第一步同apriori

$C1 =$

m	3
o	3
n	2
k	5
e	4
y	3
d	1
a	1
u	1
c	2
i	1

k	5
e	4
m	3
o	3
y	3

Sort

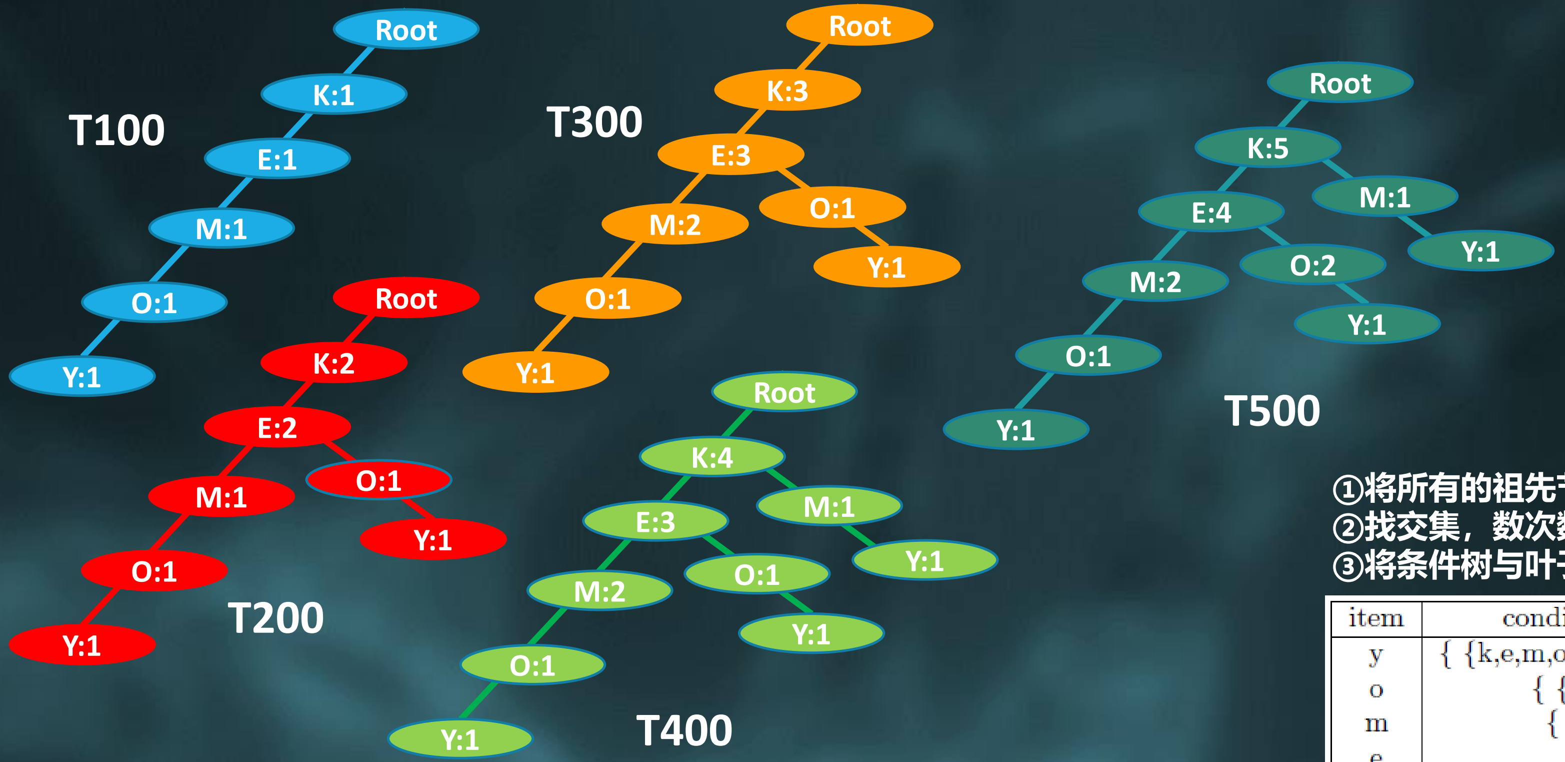
TID	Items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

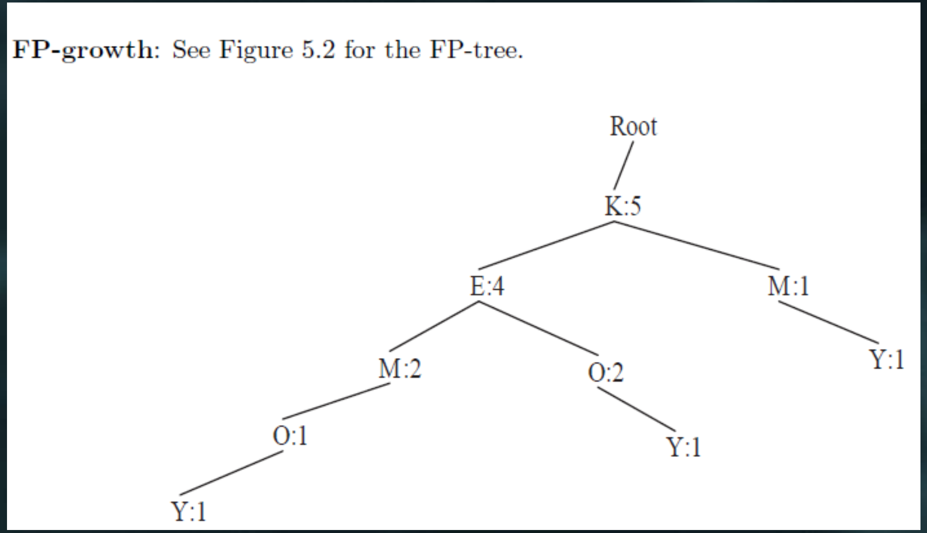
Discard

TID	Items_bought
T100	{K, E, M, O, Y}
T200	{K, E, O, Y }
T300	{K, E,M}
T400	{K, M, Y}
T500	{K, E, O}

Sort

FP-growth





FP-growth: See Figure 5.2 for the FP-tree.

①将所有的祖先节点计数设置为叶子节点的计数，建立条件模式基
②找交集，数次数，构建FP条件树
③将条件树与叶子item合起来，就是频繁模式

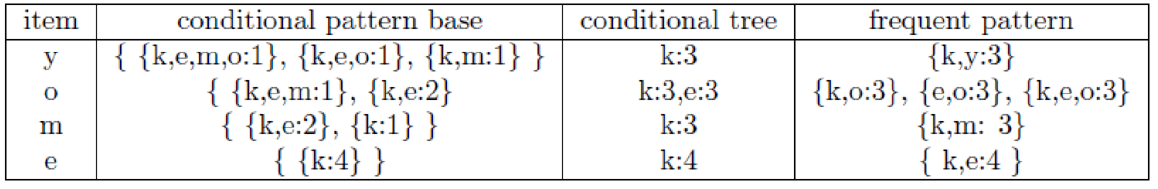
item	conditional pattern base	conditional tree	frequent pattern
y	{ {k,e,m,o:1}, {k,e,o:1}, {k,m:1} }	k:3	{k,y:3}
o	{ {k,e,m:1}, {k,e:2} }	k:3,e:3	{k,o:3}, {e,o:3}, {k,e,o:3}
m	{ {k,e:2}, {k:1} }	k:3	{k,m: 3}
e	{ {k:4} }	k:4	{ k,e:4 }

TID	Items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

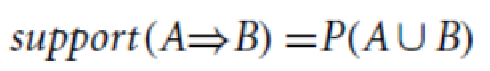
min$_{sup}$ = *60%* and *min*$_{conf}$ = *80%.*

(2) List all of the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and item$_i$ denotes variables representing items (e.g., “A”, “B”, etc.):

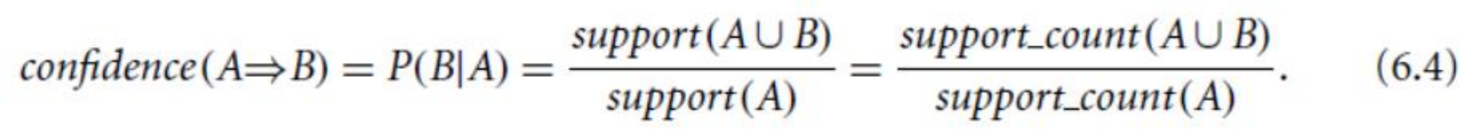
$\forall$x $\in$ transaction, buys(X, item$_1$)$\wedge$buys(X, item$_2$) $\Rightarrow$ buys(X, item$_3$) [s, c]



item	conditional pattern base	conditional tree	frequent pattern
y	{ {k,e,m,o:1}, {k,e,o:1}, {k,m:1} }	k:3	{k,y:3}
o	{ {k,e,m:1}, {k,e:2}	k:3,e:3	{k,o:3}, {e,o:3}, {k,e,o:3}
m	{ {k,e:2}, {k:1} }	k:3	{k,m: 3}
e	{ {k:4} }	k:4	{ k,e:4 }



$$support(A \Rightarrow B) = P(A \cup B)$$



$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}. \qquad (6.4)$$

1. K, E->O sup=(k,e)/5=4/5=0.8 con=(k,e,o)/(k,e)=3/4=0.75 ✗
2. O,E-> K sup=(o,e)/5=3/5=0.6 con=(k,e,o)/(o,e)=3/3=1 ☺
3. K,O->E sup=(k,o)/5=3/5=0.6 con=(k,e,o)/(k,o)=3/3=1 ☺

Problem 3

The following table consists of training data from an employee database. The data have been generalized. For example, "31 . . . 35" for age represents the age range of 31 to 35. For a given row entry, *count* represents the number of data tuples having the values for *department, status, age*, and *salary* given in that row.

department	status	age	salary	count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20
systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6

Let *status* be the class label attribute.

a) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

b) Use your algorithm to construct a decision tree from the given data.

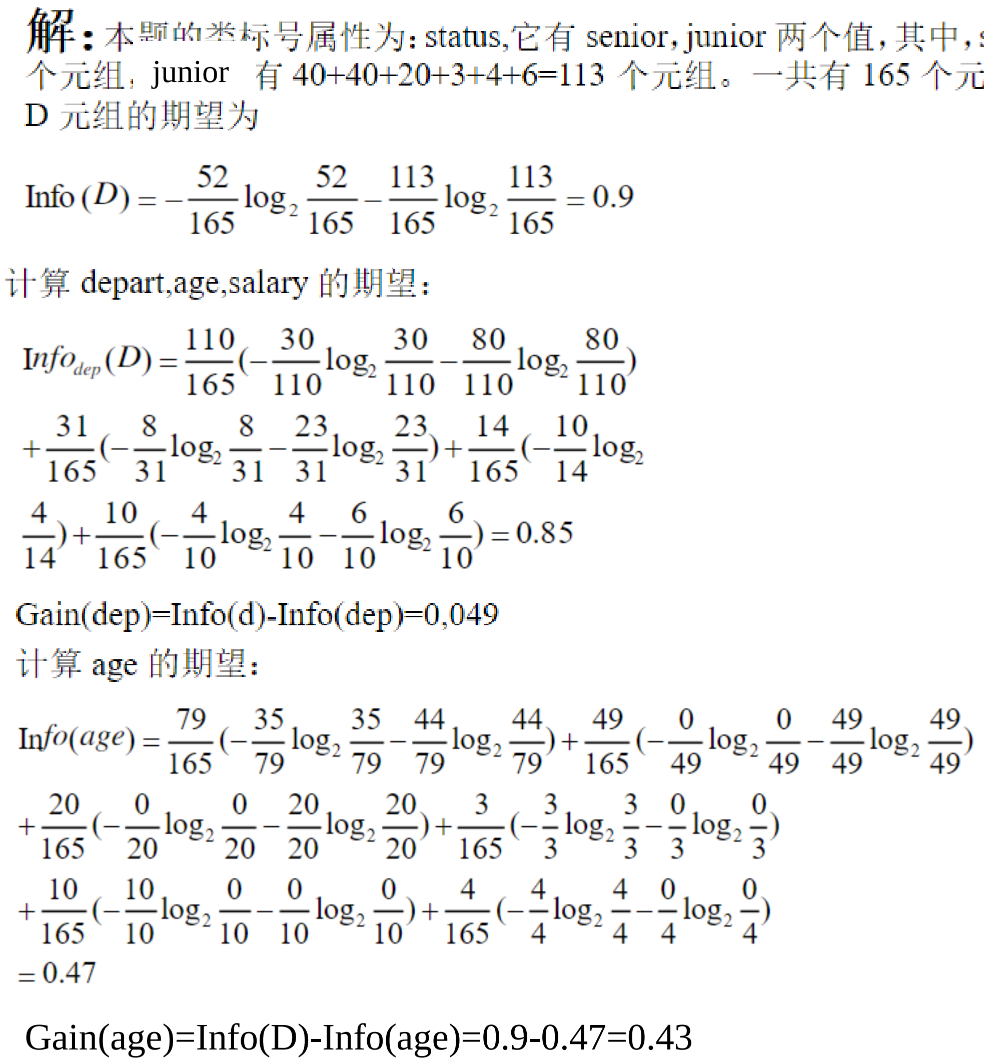
c) Given a data tuple having the values "systems," "26-30," and "46K-50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

d) Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers.

e) Using the multilayer feed-forward neural network obtained above, show the weight values after one iteration of the backpropagation algorithm, given the training instance "(sales, senior, 31-35, 46K-50K)." Indicate your initial weight values and biases, and the learning rate used.

How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

department	status	age	salary	count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20
systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6



解：本题的类标号属性为：status,它有 senior，junior 两个值，其中，senior 有 30+5+3+10+4=52 个元组，junior 有 40+40+20+3+4+6=113 个元组。一共有 165 个元组。

D 元组的期望为

$$\text{Info}\,(D) = -\frac{52}{165}\log_2\frac{52}{165} - \frac{113}{165}\log_2\frac{113}{165} = 0.9$$

计算 depart,age,salary 的期望：

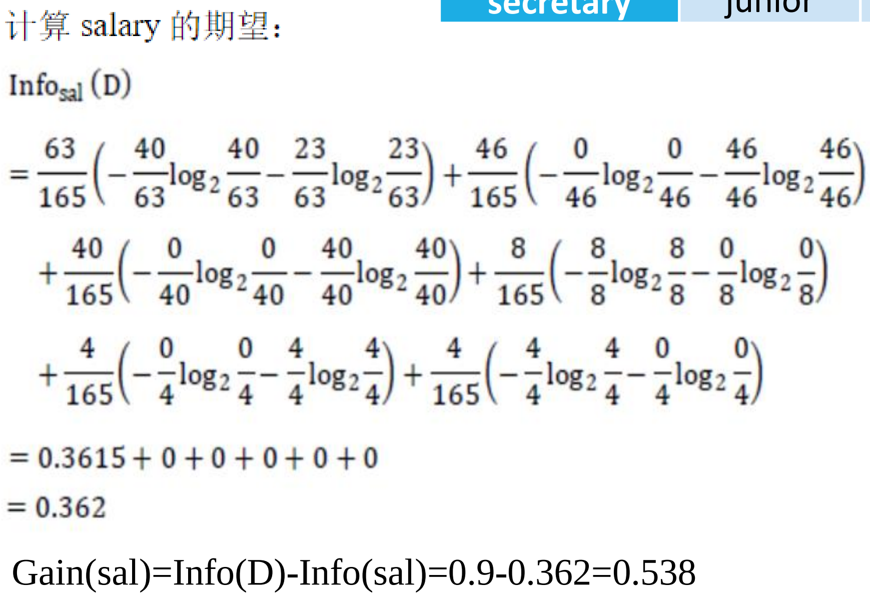
$$Info_{dep}(D) = \frac{110}{165}(-\frac{30}{110}\log_2\frac{30}{110} - \frac{80}{110}\log_2\frac{80}{110})$$
$$+\frac{31}{165}(-\frac{8}{31}\log_2\frac{8}{31} - \frac{23}{31}\log_2\frac{23}{31}) + \frac{14}{165}(-\frac{10}{14}\log_2$$
$$\frac{4}{14}) + \frac{10}{165}(-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}) = 0.85$$

Gain(dep)=Info(d)-Info(dep)=0,049

计算 age 的期望：

$$Info(age) = \frac{79}{165}(-\frac{35}{79}\log_2\frac{35}{79} - \frac{44}{79}\log_2\frac{44}{79}) + \frac{49}{165}(-\frac{0}{49}\log_2\frac{0}{49} - \frac{49}{49}\log_2\frac{49}{49})$$
$$+\frac{20}{165}(-\frac{0}{20}\log_2\frac{0}{20} - \frac{20}{20}\log_2\frac{20}{20}) + \frac{3}{165}(-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3})$$
$$+\frac{10}{165}(-\frac{10}{10}\log_2\frac{0}{10} - \frac{0}{10}\log_2\frac{0}{10}) + \frac{4}{165}(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4})$$
$$= 0.47$$

Gain(age)=Info(D)-Info(age)=0.9-0.47=0.43



计算 salary 的期望：

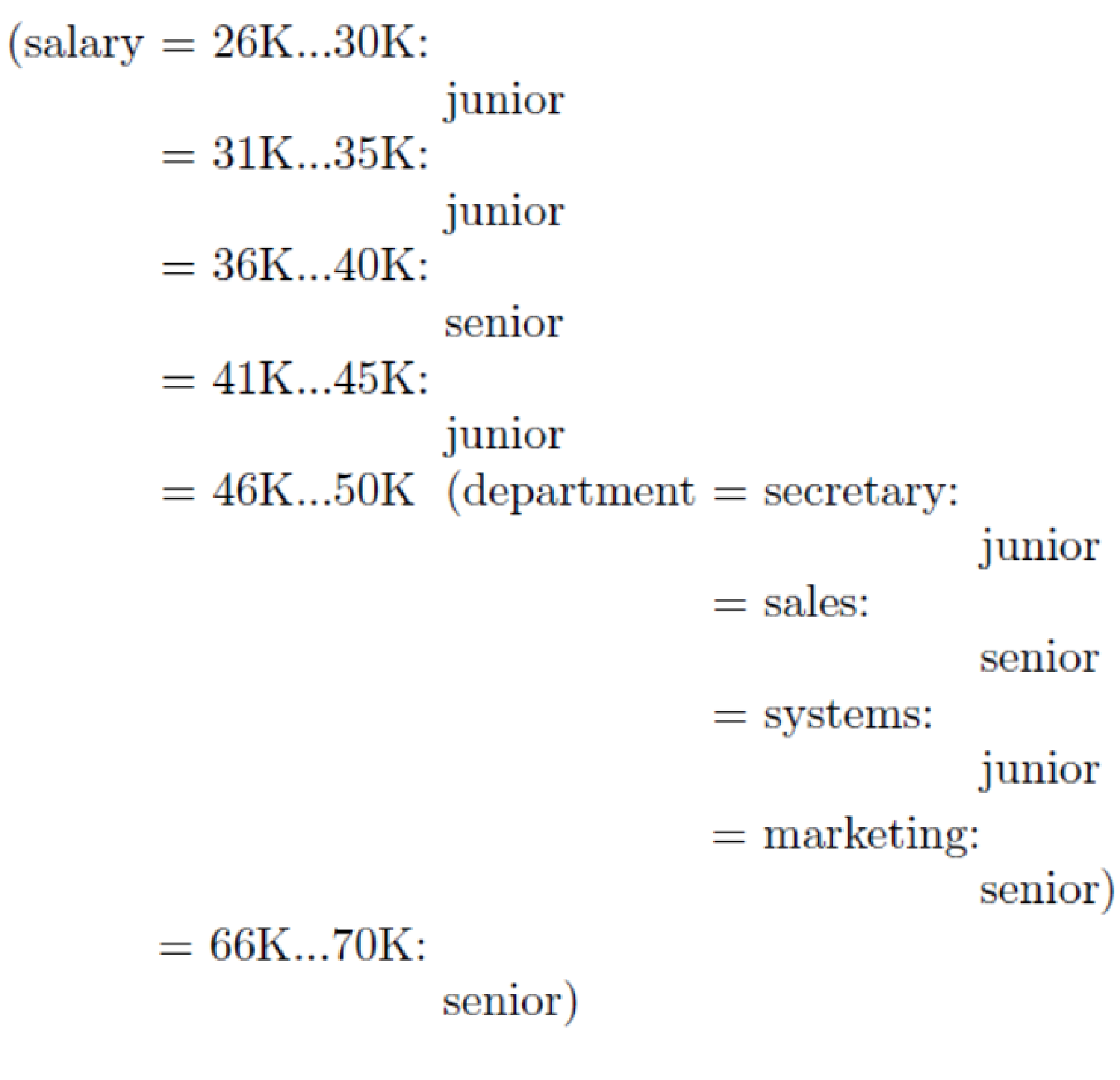
$$\text{Info}_{\text{sal}}\,(D)$$
$$= \frac{63}{165}\left(-\frac{40}{63}\log_2\frac{40}{63} - \frac{23}{63}\log_2\frac{23}{63}\right) + \frac{46}{165}\left(-\frac{0}{46}\log_2\frac{0}{46} - \frac{46}{46}\log_2\frac{46}{46}\right)$$
$$+\frac{40}{165}\left(-\frac{0}{40}\log_2\frac{0}{40} - \frac{40}{40}\log_2\frac{40}{40}\right) + \frac{8}{165}\left(-\frac{8}{8}\log_2\frac{8}{8} - \frac{0}{8}\log_2\frac{0}{8}\right)$$
$$+\frac{4}{165}\left(-\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}\right) + \frac{4}{165}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)$$
$$= 0.3615 + 0 + 0 + 0 + 0 + 0$$
$$= 0.362$$

Gain(sal)=Info(D)-Info(sal)=0.9-0.362=0.538

∵ Gain(sal)>Gain(age)>Gain(dep)，
且具有最高信息增益的属性应该作为结点N的分裂属性
∴ 决策树应从salary开始分裂

Use your algorithm to construct a decision tree from the given data.

department	status	age	salary	count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20
systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6



```
(salary = 26K...30K:
                    junior
        = 31K...35K:
                    junior
        = 36K...40K:
                    senior
        = 41K...45K:
                    junior
        = 46K...50K  (department = secretary:
                                              junior
                                 = sales:
                                              senior
                                 = systems:
                                              junior
                                 = marketing:
                                              senior)
        = 66K...70K:
                    senior)
```

Given a data tuple having the values “systems,” “26-30,” and “46K-50K” for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

department	status	age	salary	count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20
systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6

解一：设元组的各个属性之间相互独立，所以先求每个属性的类条件概率：

P(systems|junior)=(20+3)/(40+40+20+3+4+6)=23/113;

P(26-30|junior)=(40+3+6)/113=49/113；

P(46K-50K|junior)=(20+3)/113=23/113；

∵ X=(department=system,age=26…30,salary=46K…50K)；

∴ P(X|junior)=P(systems|junior)P(26-30|junior)P(46K-50K|junior)

=23×49×23/113^3=25921/1442897=0.01796；

P(systems|senior)=(5+3)/(30+5+3+10+4)=23/52;

P(26-30|senior)=(0)/53=0;

P(46K-50K|senior)=(30+10)/52=40/52；

∵ X=(department=system,age=26…30,salary=46K…50K)；

∴ P(X|senior)=P(systems|senior)P(26-30|senior)P(46K-50K|senior)=0;

∵ P(junior)=113/165=0.68；

∵ P(senior)=52/165=0.32；

∴ P(X|junior)P(junior)=0.01796×0.68=0.0122128>0=0=P(X|senior)P(senior)；

所以：朴素贝叶斯分类器将X分到junior类。

解二：设元组的各属性之间不独立，其联合概率不能写成份量相乘的形式。所以已知：
X=(department=system,age=26…30,salary=46K…50K)，元组总数为：
30+40+40+20+5+3+3+10+4+4+6=165。
先验概率：
当status=senior时，元组总数为：
30+5+3+10+4=52，P(senior)=52/165=0.32；
当status=junior时，元组总数为：
40+40+20+3+4+6=113，P(junior)=113/165=0.68；
因为status=senior状态没有对应的age=26…30区间，所以：P(X|senior)=0；
因为status=junior状态对应的partment=systems、age=26…30区间的总元组数为：3，所以：
P(X|junior)=3/113；
因为：P(X|junior)P(junior)=3/113×113/165
=0.018>0=P(X|senior)P(senior)；
所以：朴素贝叶斯分类器将X分到junior类。

- Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers.
- Using the multilayer feed-forward neural network obtained above, show the weight values after one iteration of the backpropagation algorithm, given the training instance “(sales, senior, 31-35, 46K-50K).” Indicate your initial weight values and biases, and the learning rate used.

No Standard Answer

department	status	age	salary	count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20
systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6

Project 1

Topic 1: Interview one person from a key business function, such as finance, human resources, or marketing. Concentrate your questions on the following items: How does he or she retrieve data needed to make business decisions? From what kind of system (personal database, enterprise system, or data warehouse) are the data retrieved? What data mining tasks involved in their work? What kind of tools or approaches are leveraged?

Project 2

Topic 2: Research on a local company and study how data mining approaches can help the company to make business decisions. What are the main challenges as well as contributions of their data mining systems?

Project 3

Topic 3: Based on your experiences, describe a real-world scenario where some kind of rules or knowledge needs to be discovered with data mining approaches. Introduce the background situation and the exact business requirements, and then explain why simple statistic methods cannot help. What kinds of data mining methods could be exploited in the situation?

The End

www.wangting.ac.cn