作业 7

朱志儒 SA20225085

8.7（a）

（1）将每个元组的 count 作为计算属性选择方法中的一部分

（2）在决定元组多数表决的时候考虑元组的 count

（b）

选择根节点：

$$\text{senior} = 30 + 5 + 3 + 10 + 4 = 52$$

$$\text{junior} = 40 + 40 + 20 + 3 + 4 + 6 = 113$$

$$\text{total} = 52 + 113 = 165$$

$$\text{H(status)} = -\frac{52}{165}\log\frac{52}{165} - \frac{113}{165}\log\frac{113}{165} = 0.899$$

Department:

Sales	Senior	30	110
	Junior	80	
Systems	Senior	8	31
	Junior	23	
Marketing	Senior	10	14
	Junior	4	
Secretary	Senior	4	10
	Junior	6	

$$\begin{aligned}\text{H(status|department)} &= \frac{110}{165} \times \left(-\frac{30}{110}\log\frac{30}{110} - \frac{80}{110}\log\frac{80}{110}\right) \\ &+ \frac{38}{165} \times \left(-\frac{8}{31}\log\frac{8}{31} - \frac{23}{38}\log\frac{23}{38}\right) \\ &+ \frac{14}{165} \times \left(-\frac{10}{14}\log\frac{10}{14} - \frac{4}{14}\log\frac{4}{14}\right) \\ &+ \frac{10}{165} \times \left(-\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10}\right) = 0.828\end{aligned}$$

$$\text{g(status, department)} = 0.899 - 0.828 = 0.071$$

Age

21…25	Senior	0	20
	Junior	20	
26…30	Senior	0	49
	Junior	49	
31…35	Senior	35	79
	Junior	44	
36…40	Senior	10	10
	Junior	0	
41…45	Senior	3	3
	Junior	0	
46…50	Senior	4	4
	Junior	0	

$$\mathrm{H}(status|Age) = \frac{79}{165} \times \left(-\frac{35}{79}\log\frac{35}{79} - \frac{44}{79}\log\frac{44}{79}\right) = 0.474$$

$$\mathrm{g}(status, Age) = 0.899 - 0.474 = 0.425$$

Salary

26K…30K	Senior	0	46
	Junior	46	
31K…35K	Senior	0	40
	Junior	40	
36K…40K	Senior	4	4
	Junior	0	
41K…45K	Senior	0	4
	Junior	4	
46K…50K	Senior	40	63
	Junior	23	
66K…70K	Senior	8	8
	Junior	0	

$$\mathrm{H}(status|Salary) = \frac{64}{165} \times \left(-\frac{40}{63}\log\frac{40}{63} - \frac{23}{63}\log\frac{23}{63}\right) = 0.362$$

$$\mathrm{g}(status, Salary) = 0.899 - 0.362 = 0.537$$

显然 Salary 的信息增益最大，选择 Salary 作为根节点。

第二次划分：

Department:

Sales	Senior	30	30
	Junior	0	
Systems	Senior	0	23
	Junior	23	
Marketing	Senior	10	10
	Junior	0	
Secretary	Senior	0	0
	Junior	0	

$$\mathrm{H}(status, salary|department) = 0$$

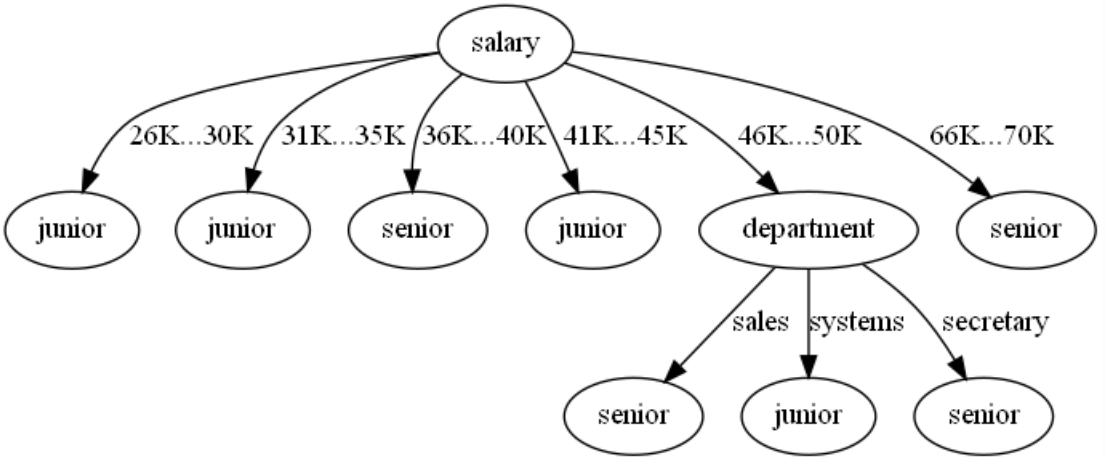
Age:

21…25	Senior	0	20
	Junior	20	
26…30	Senior	0	3
	Junior	3	
31…35	Senior	30	30
	Junior	0	
36…40	Senior	10	10
	Junior	0	
41…45	Senior	0	0
	Junior	0	
46…50	Senior	0	0
	Junior	0	

$$\mathrm{H}(status, salary|age) = 0$$

显然，department 和 age 的信息增益相同，选择任意一个进行划分。

最后的决策树如下：


salary
26K...30K
31K...35K
36K...40K
41K...45K
46K...50K
66K...70K
junior
junior
senior
junior
department
senior
sales
systems
secretary
senior
junior
senior