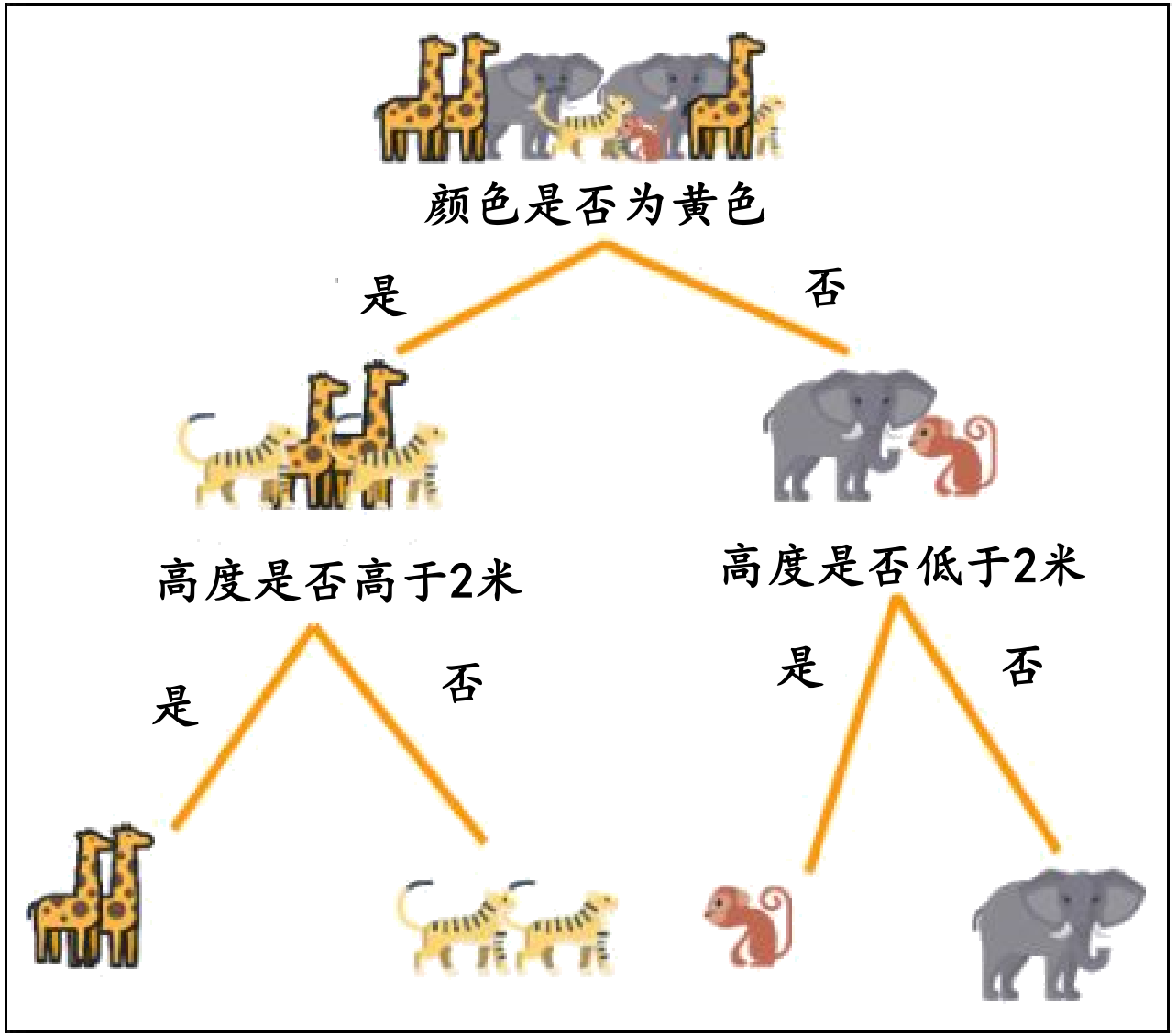
决策树

李 波

决策树（Decision Tree）

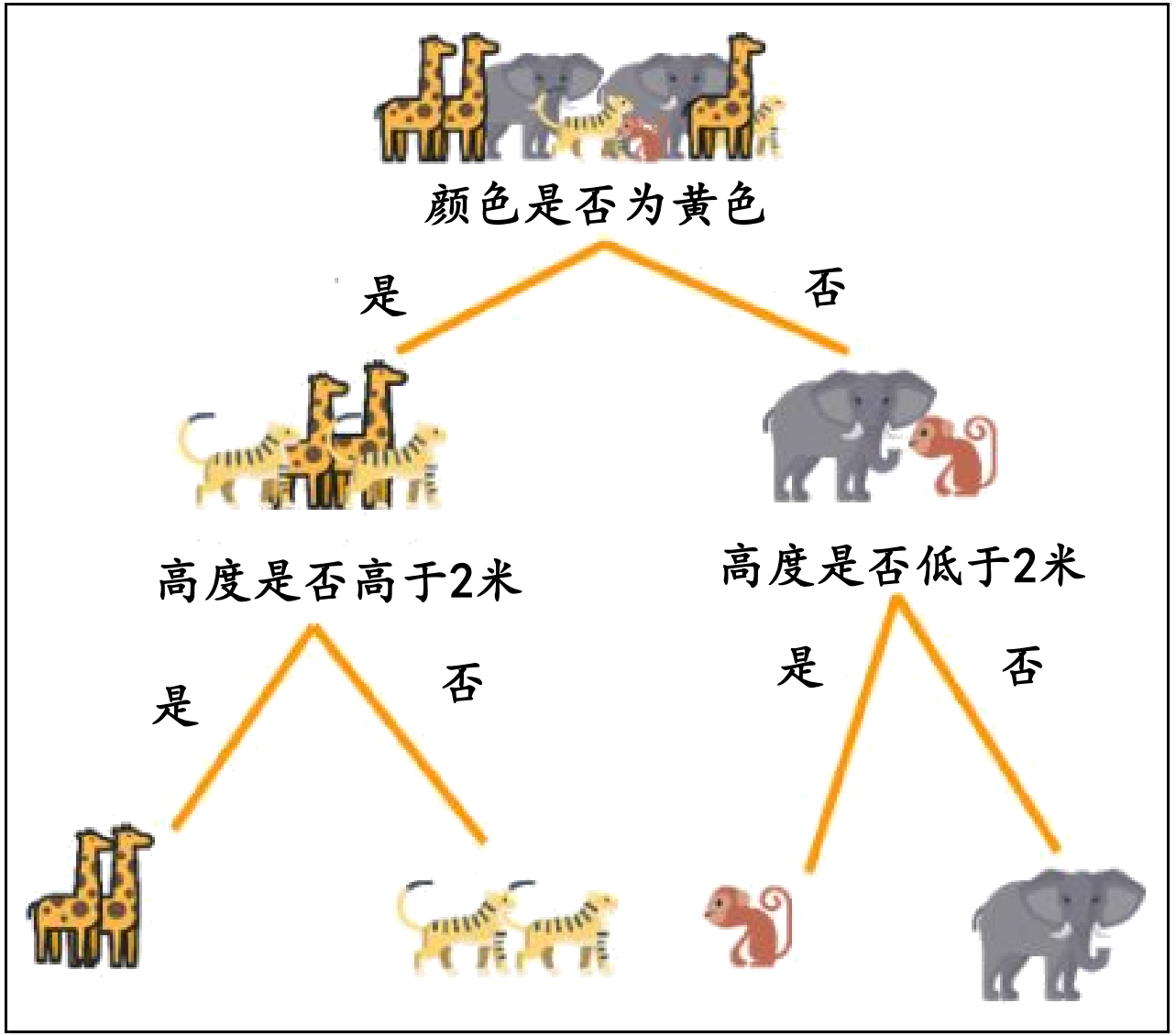
- 决策树是一个有监督学习算法.
- 与人类做出判断的流程类似.
- 决策树基于特征做出判断.
- “重要的”特征首先考虑，然后再考虑次要特征，直至得到结论.
- 如何找出“重要”特征?



决策树（Decision Tree）

- 决策树是一个有监督学习算法.
- 与人类做出判断的流程类似.
- 决策树基于特征做出判断.
- “重要的”特征首先考虑，然后再考虑次要特征，直至得到结论.
- 如何找出“重要”特征?

信息增益或熵增益

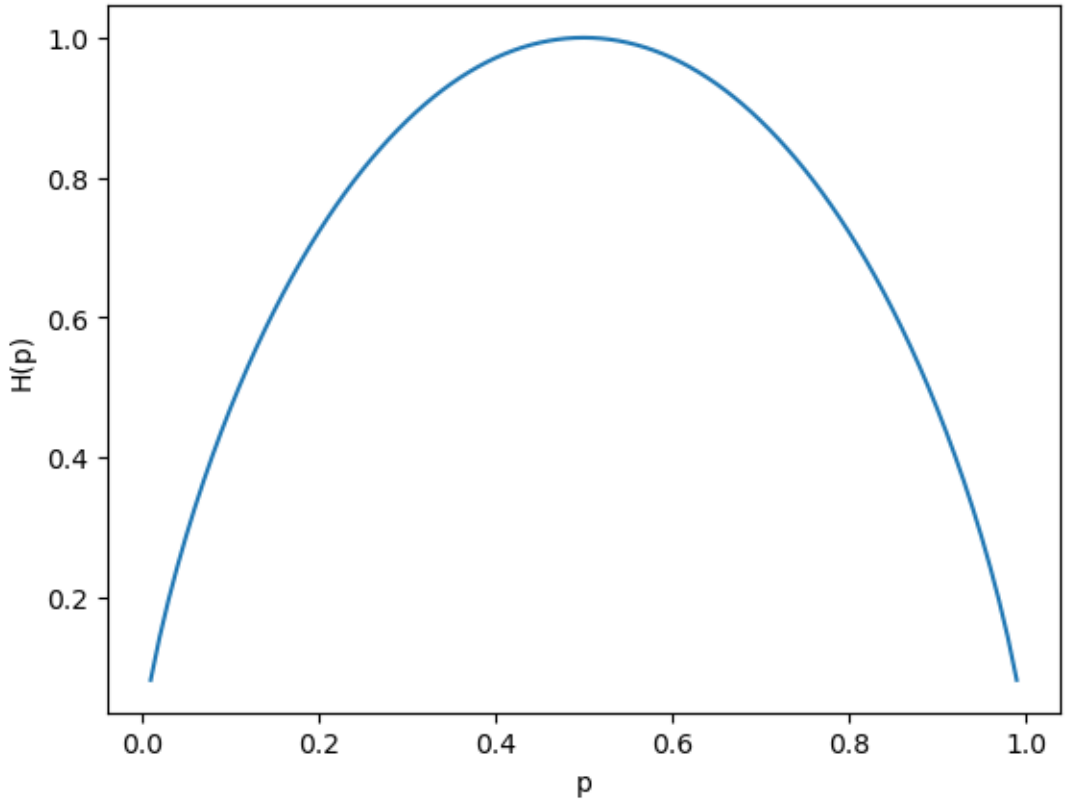


决策树（Decision Tree）

- **随机变量的熵**

$$H(Y) = -\sum_{i=1}^{k} p_i log(p_i), \quad p_i = p(Y = i)$$

$$H(Y) = \int_{-\infty}^{+\infty} f(y)dy$$



伯努利分布熵与参数关系

- 熵代表混乱程度，混乱程度越大，熵越大。
- 那种分布熵大?
  - 高斯分布?
  - 均匀分布?

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

决策树（Decision Tree）

一共15个申请人，通过贷款9人，未通过贷款6人。令$Y=1$表示通过贷款，$Y=0$表示不通过贷款

$$p(Y=1)\approx\frac{9}{15}=0.6,\qquad p(Y=0)\approx\frac{6}{15}=0.4$$

$$\mathrm{H}(Y)=-0.6\times\log(0.6)-0.4\times\log(0.4)\approx 0.6730$$

决策树（Decision Tree）

- **条件熵**

$$H(Y|A) = -\sum_{i=1}^{k} p(A = \mathrm{a}_i) H(Y|A = a_i)$$，其中A为一个特征

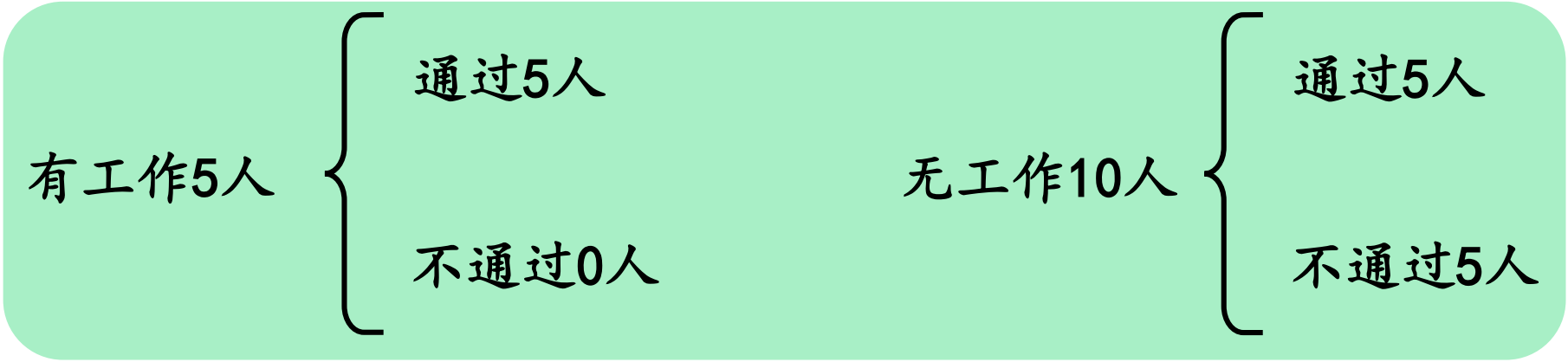
- **信息增益**

$$G(A) = H(Y) - H(Y|A)$$

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	**青年**	**是**	**否**	**好**	**是**
4	**青年**	**是**	**是**	**一般**	**是**
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	**中年**	**是**	**是**	**好**	**是**
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	**老年**	**是**	**否**	**好**	**是**
14	**老年**	**是**	**否**	**非常好**	**是**
15	老年	否	否	一般	否

有工作5人
- 通过5人
- 不通过0人

无工作10人
- 通过5人
- 不通过5人



A=工作，$p(A=\text{有工作})=\frac{5}{15}$，$p(A=\text{无工作})=10/15$

$$H(Y|A)=p(A=\text{有工作})H(Y|A=\text{有工作})+p(A=\text{无工作})H(Y|A=\text{无工作})$$

$$=\frac{5}{15}[-1\log(1)-0\log(0)]+\frac{10}{15}\left[-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right)\right]$$

$$=\frac{2}{3}\log(2)\approx 0.4621$$

信息增益：$G(A)=H(Y)-H(Y|A)=0.6730-0.4621=0.2109$

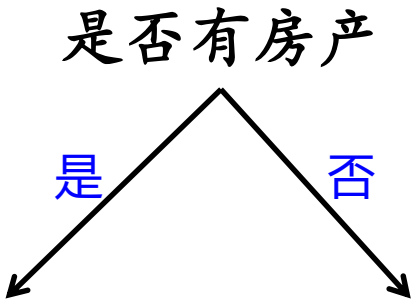
决策树（Decision Tree）

特征名称	条件熵	信息增益
年龄	0.6730	0.0000
工作	0.4621	0.2109
房产	0.3819	0.2911
信贷情况	0.4214	0.2516

决策树（Decision Tree）

特征名称	条件熵	信息增益
年龄	0.6730	0.0000
工作	0.4621	0.2109
房产	**0.3819**	**0.2911**
信贷情况	0.4214	0.2516

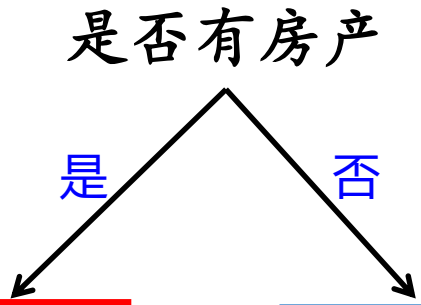
决策树（Decision Tree）



ID	年龄	有工作	有房产	信贷情况	是否通过
4	青年	是	是	一般	是
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
15	老年	否	否	一般	否

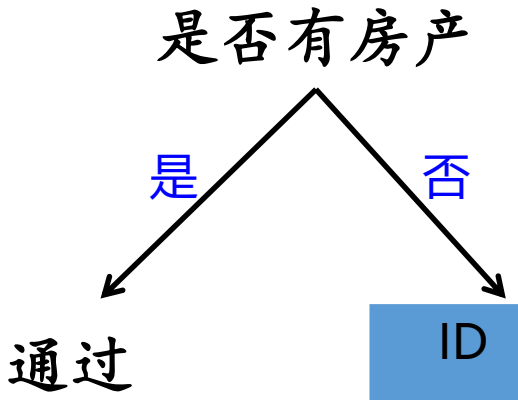
决策树（Decision Tree）



ID	年龄	有工作	有房产	信贷情况	是否通过
4	青年	是	是	一般	是
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
15	老年	否	否	一般	否

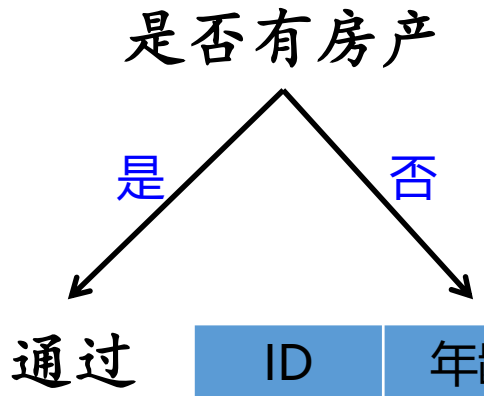
决策树（Decision Tree）



ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
15	老年	否	否	一般	否

决策树（Decision Tree）

有房产的申请人全部通过贷款，因此



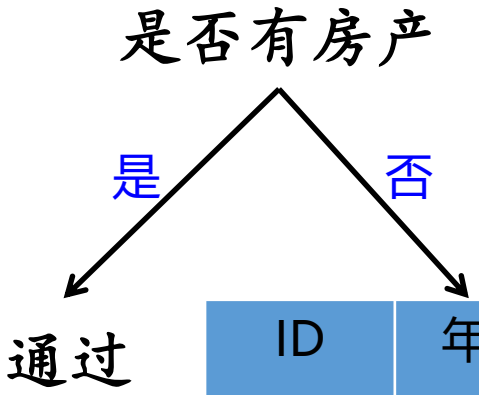
共7人，通过1人，不通过6人

$$H(Y|\text{无房产}) = -\frac{1}{7}\log\left(\frac{1}{7}\right) - \frac{6}{7}\log\left(\frac{6}{7}\right) \approx 0.4101$$

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
15	老年	否	否	一般	否

决策树（Decision Tree）

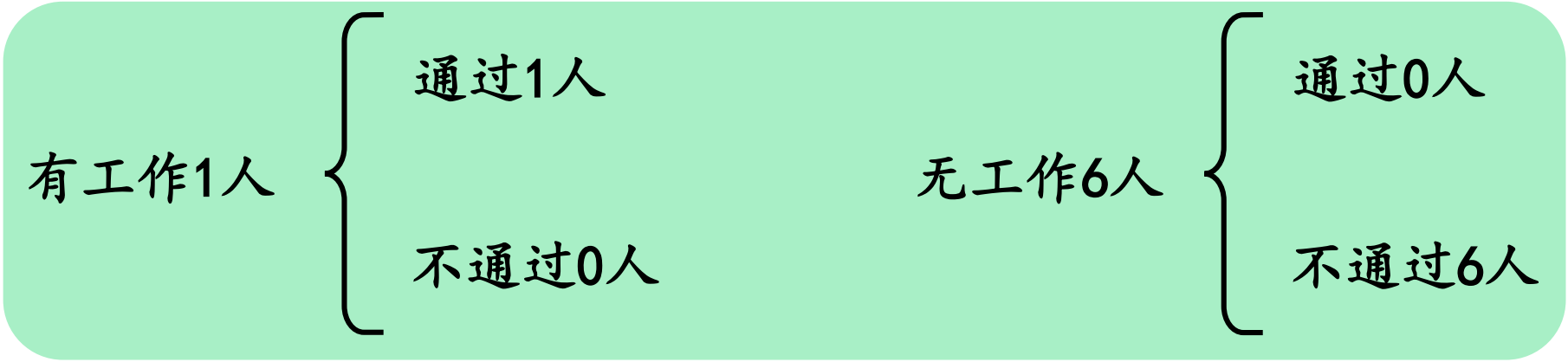
有房产的申请人全部通过贷款，因此



ID	年龄	**有工作**	有房产	信贷情况	是否通过
1	青年	**否**	否	一般	否
2	青年	**否**	否	好	否
3	青年	**是**	否	好	是
5	青年	**否**	否	一般	否
6	中年	**否**	否	一般	否
7	中年	**否**	否	好	否
15	老年	**否**	否	一般	否

有工作1人
- 通过1人
- 不通过0人

无工作6人
- 通过0人
- 不通过6人



$A=$ 工作，$p(A=\text{有工作})=1/7,\ p(A=\text{无工作})=6/7$

$$H(Y|A，\ \text{无房产})=p(A=\text{有工作})H(Y|A=\text{有工作，}\ \text{无房产})+p(A=\text{无工作})H(Y|A=\text{无工作，无房产})$$

$$=\frac{1}{7}[-1\log(1)-0\log(0)]+\frac{6}{7}[-0\log(0)-1\log(1)]$$

$$=0$$

信息增益：

$$G(A，\text{无房产})=H(y|\text{无房产})-H(Y|A=\text{工作，无房产})=0.4101-0=0.4101$$

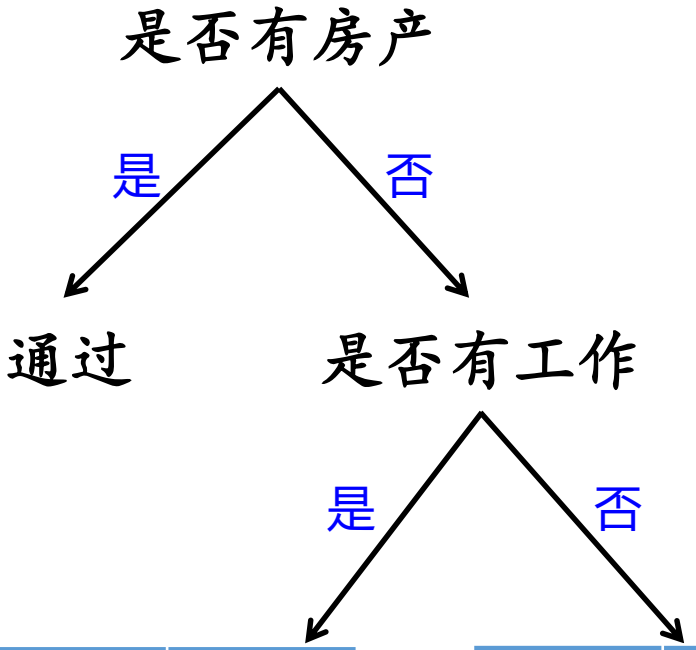
决策树（Decision Tree）

特征名称	条件熵	信息增益
年龄	0.3213	0.0888
工作	0	0.4101
信贷情况	0.2728	0.1373

决策树（Decision Tree）

特征名称	条件熵	信息增益
年龄	0.3213	0.0888
工作	**0**	**0.4101**
信贷情况	0.2728	0.1373

决策树（Decision Tree）

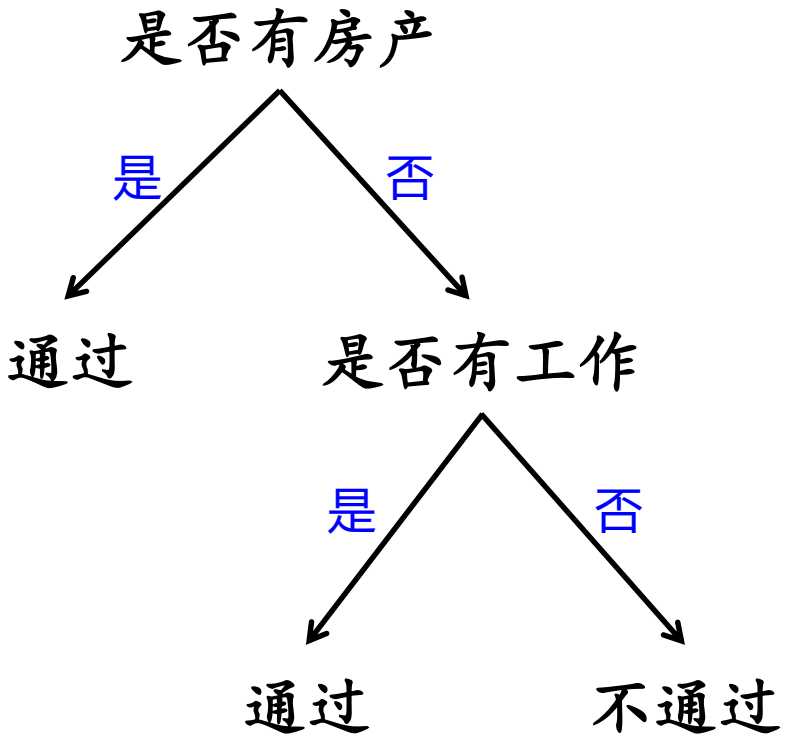


ID	年龄	有工作	有房产	信贷情况	是否通过
3	青年	是	否	好	是

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	青年	否	否	好	否
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
15	老年	否	否	一般	否

决策树（Decision Tree）

- ID3 算法：使用信息增益作为特征选择标准
- C4.5 算法：使用信息增益比作为特征选择标准



ID	年龄	有工作	有房产	信贷情况	是否通过
1	18.9	否	否	一般	否
2	21.5	否	否	好	否
3	25.0	是	否	好	是
4	29.4	是	是	一般	是
5	31.2	否	否	一般	否
6	33.7	否	否	一般	否
7	34.1	否	否	好	否
8	36.0	是	是	好	是
9	39.2	否	是	非常好	是
10	41.6	否	是	非常好	是
11	45.8	否	是	非常好	是
12	49.5	否	是	好	是
13	52.3	是	否	好	是
14	55.4	是	否	非常好	是
15	61.0	否	否	一般	否

决策树（Decision Tree）

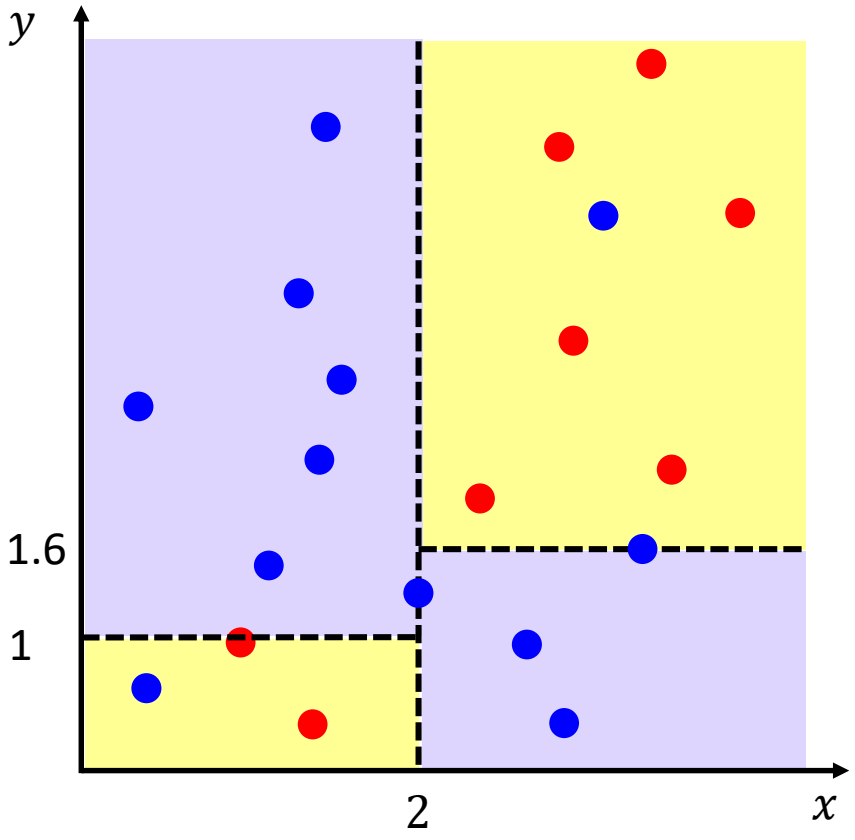
- 在银行贷款的例子中，年龄的切分值有14个。
- 对于连续取值的某个特征A，按照如下方式选择切分点

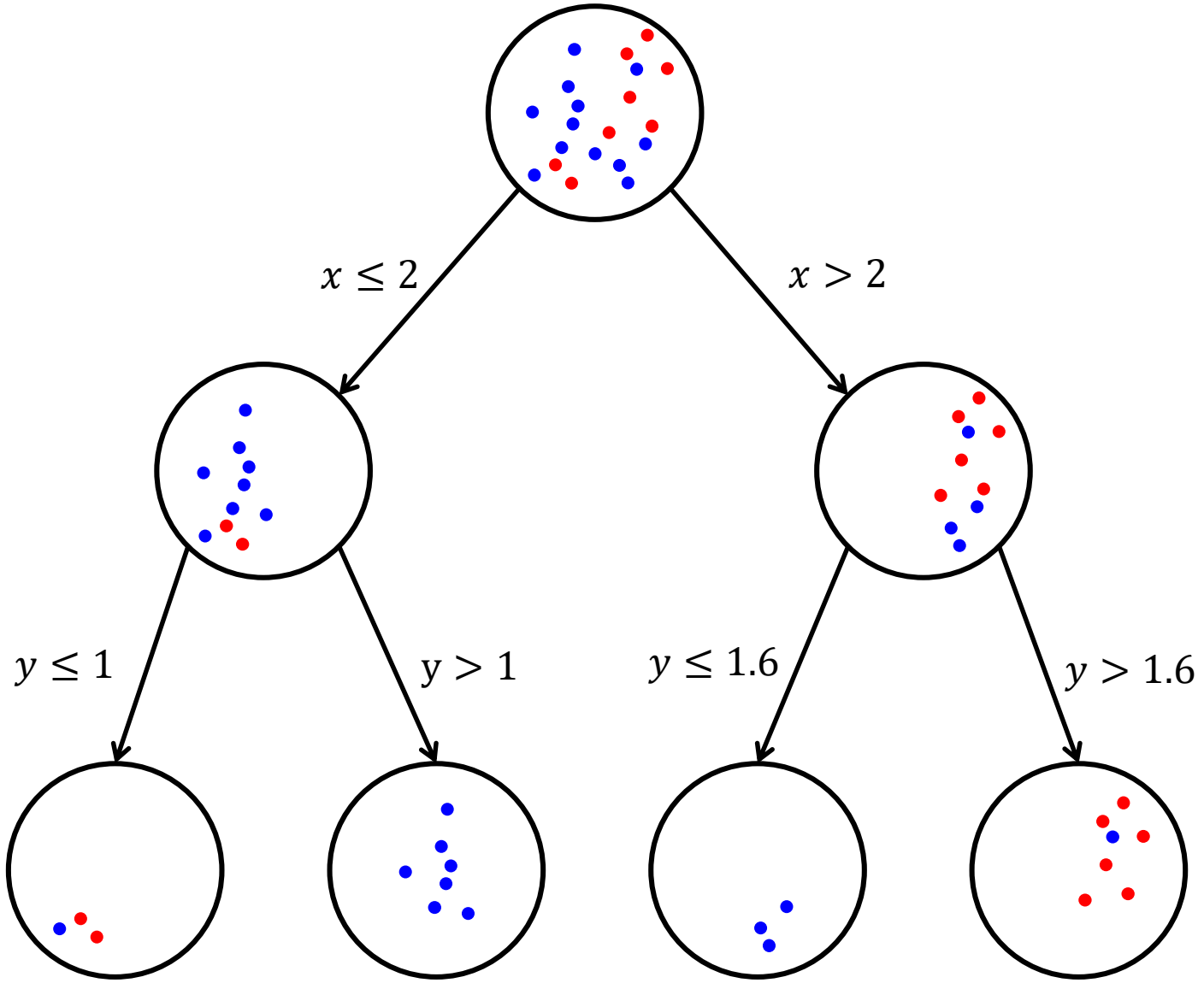
$$\max_{a} H(Y) - H(Y|A=a) \quad 或 \quad \min_{a} H(Y|A=a)$$

，其中a的取值没有范围。实际上，只需考虑排序后两个相邻特征值的中位数作为a的取值即可。

ID	年龄	有工作	有房产	信贷情况	是否通过
1	18.9	否	否	一般	否
2	21.5	否	否	好	否
3	25.0	是	否	好	是
4	29.4	是	是	一般	是
5	31.2	否	否	一般	否
6	33.7	否	否	一般	否
7	34.1	否	否	好	否
8	36.0	是	是	好	是
9	39.2	否	是	非常好	是
10	41.6	否	是	非常好	是
11	45.8	否	是	非常好	是
12	49.5	否	是	好	是
13	52.3	是	否	好	是
14	55.4	是	否	非常好	是
15	61.0	否	否	一般	否

决策树（Decision Tree）



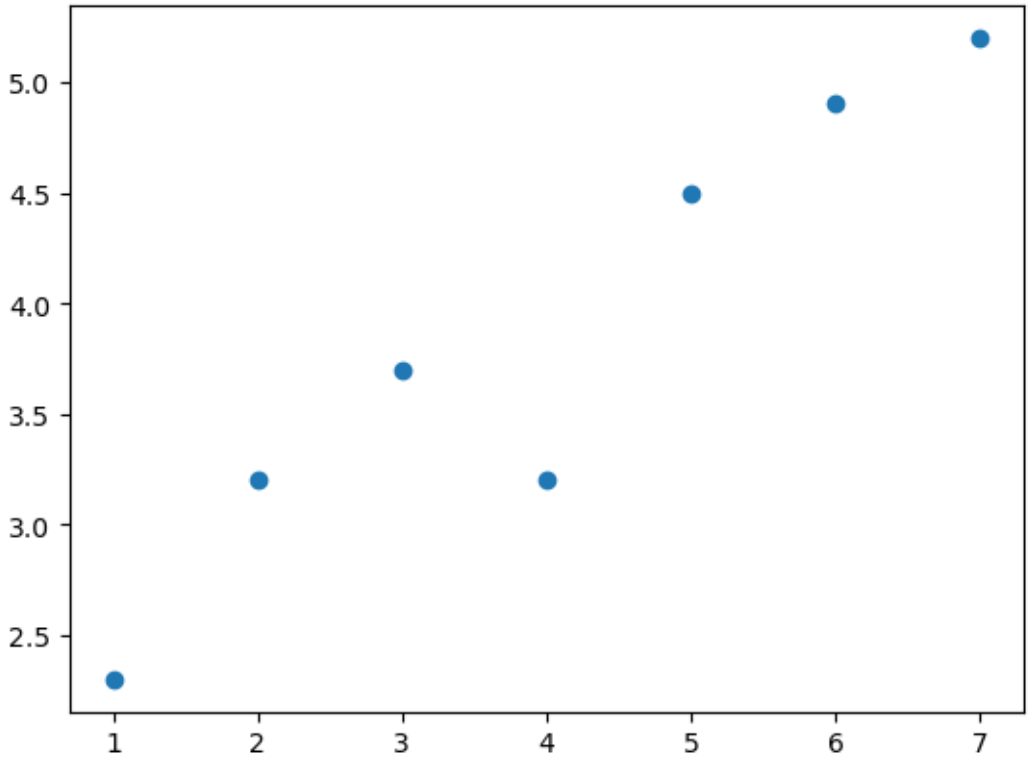


决策树（Decision Tree）

回归树

- 标签为连续值.
- 选择一个特征，分裂后方差和最小

(1,2.3), (2,3.2), (3,3.7), (4,3.2), (5,4.5), (6,4.9), (7,5.2)

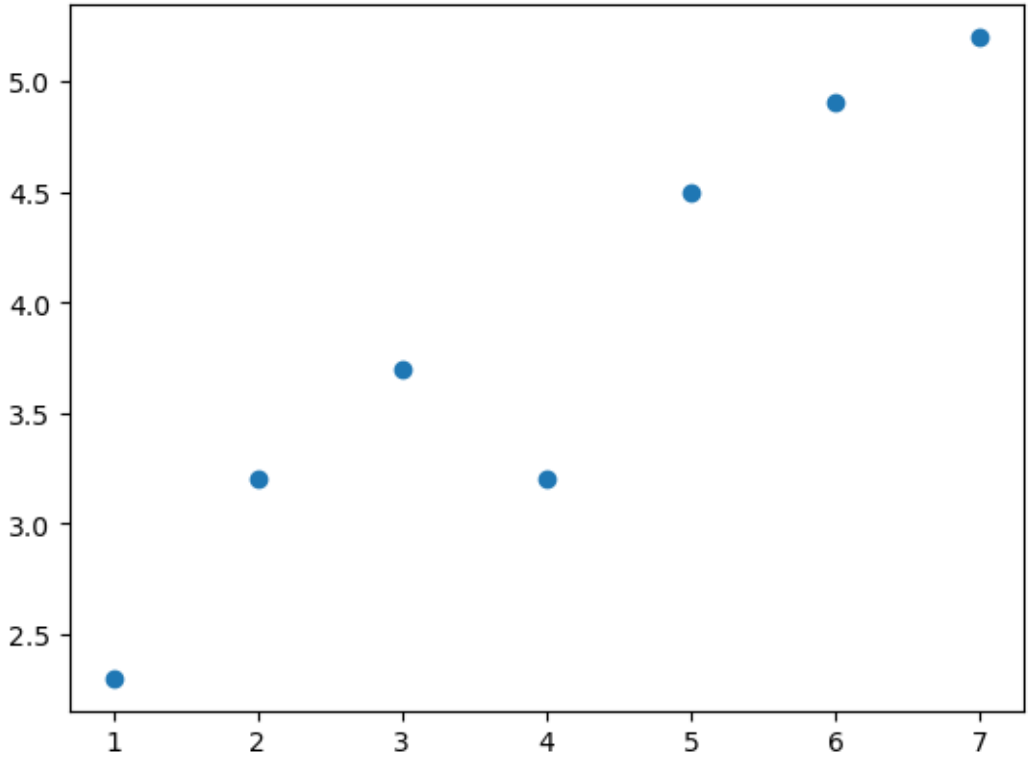


决策树（Decision Tree）

回归树

- 标签为连续值
- 选择一个特征，分裂后方差和最小

(1,2.3), (2,3.2), (3,3.7), (4,3.2), (5,4.5), (6,4.9), (7,5.2)



$$\bar{y}_{\text{left}} = \frac{1}{4}(2.3 + 3.2 + 3.7 + 3.2) \approx 3.10$$

$$\bar{y}_{\text{right}} = \frac{1}{3}(4.5 + 4.9 + 5.2) \approx 4.86$$

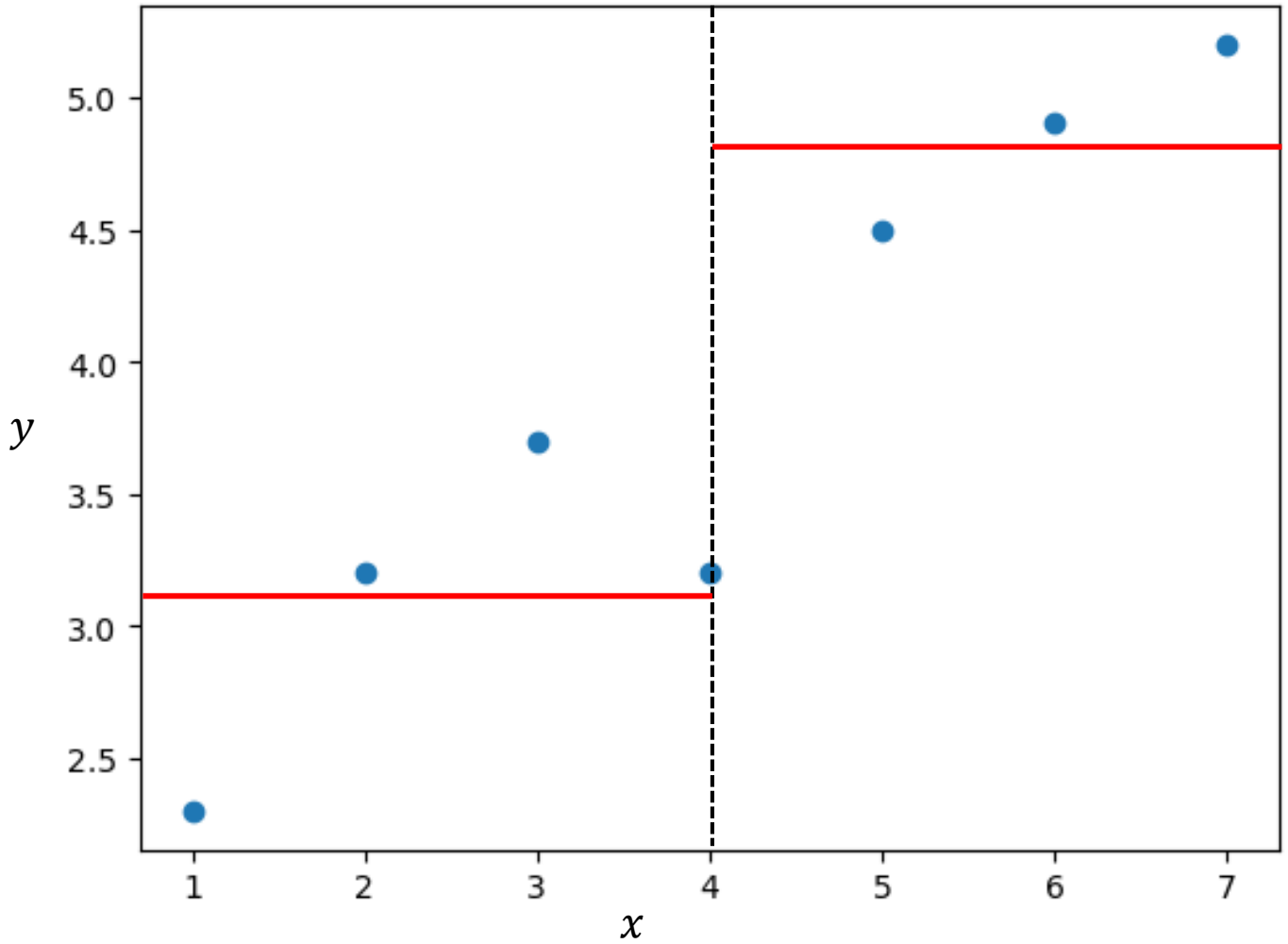
$$Var_{\text{left}} = \frac{1}{4}\left((2.3 - 3.10)^2 + (3.2 - 3.10)^2 + (3.7 - 3.10)^2 + (3.2 - 3.10)^2\right) \approx 0.26$$

$$Var_{\text{right}} = \frac{1}{3}\left((4.5 - 4.86)^2 + (4.9 - 4.86)^2 + (5.2 - 3.86)^2\right) \approx 0.08$$

$$Var = Var_{\text{left}} + Var_{\text{right}} = 0.26 + 0.08 = 0.34$$

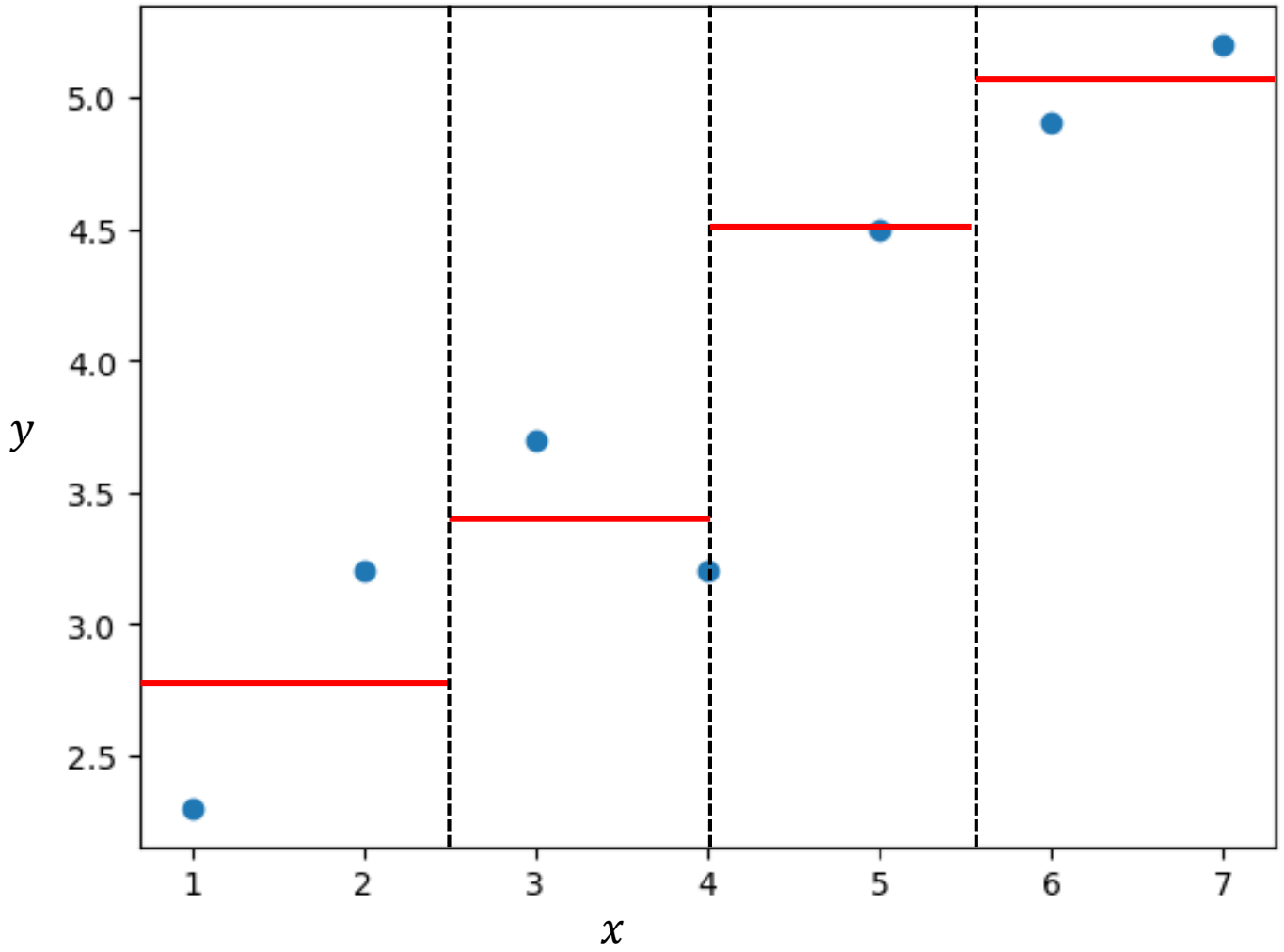
决策树（Decision Tree）

回归树　(1,2.3), (2,3.2), (3,3.7), (4,3.2), (5,4.5), (6,4.9), (7,5.2)



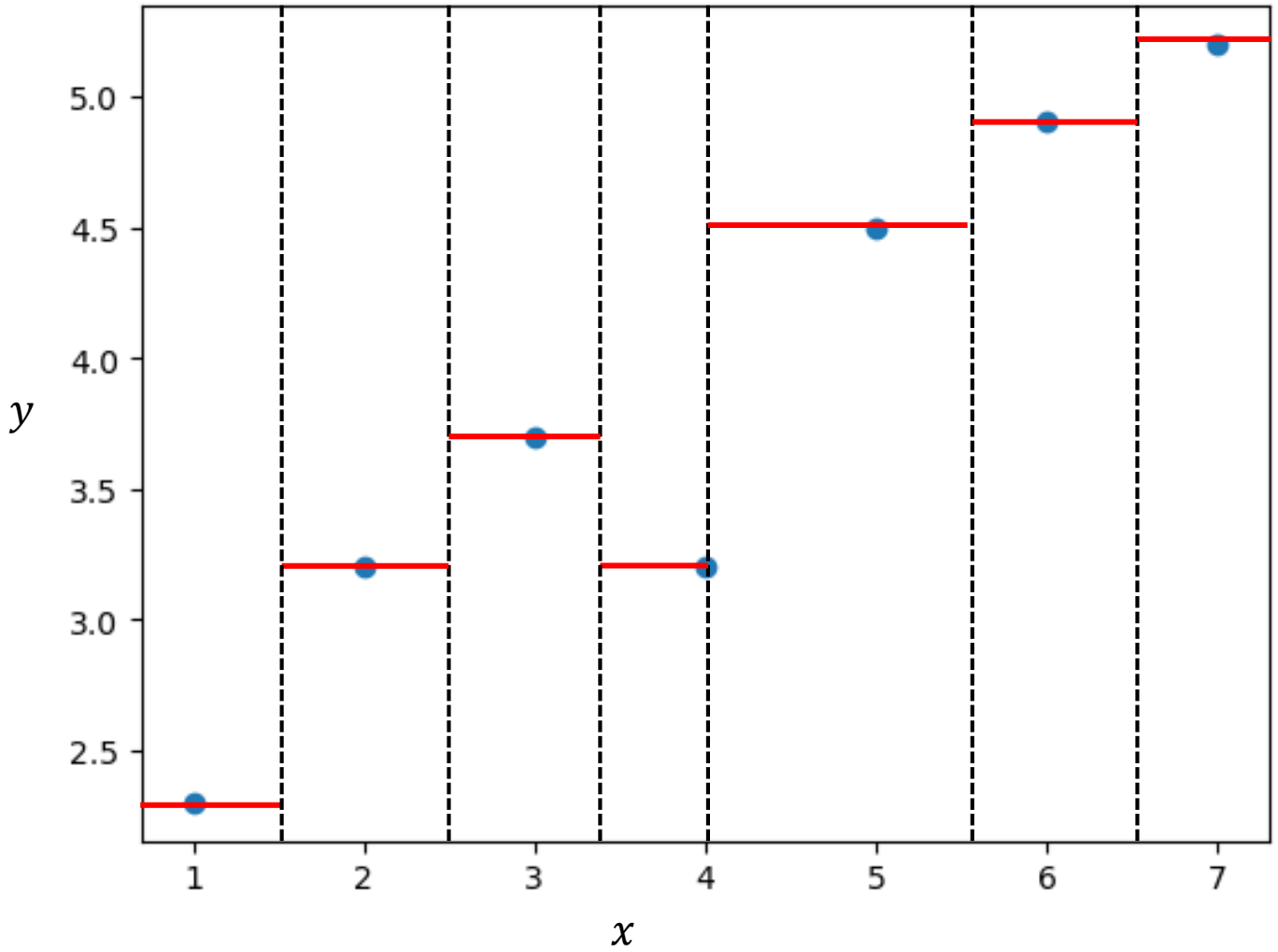
决策树（Decision Tree）

回归树　(1,2.3), (2,3.2), (3,3.7), (4,3.2), (5,4.5), (6,4.9), (7,5.2)



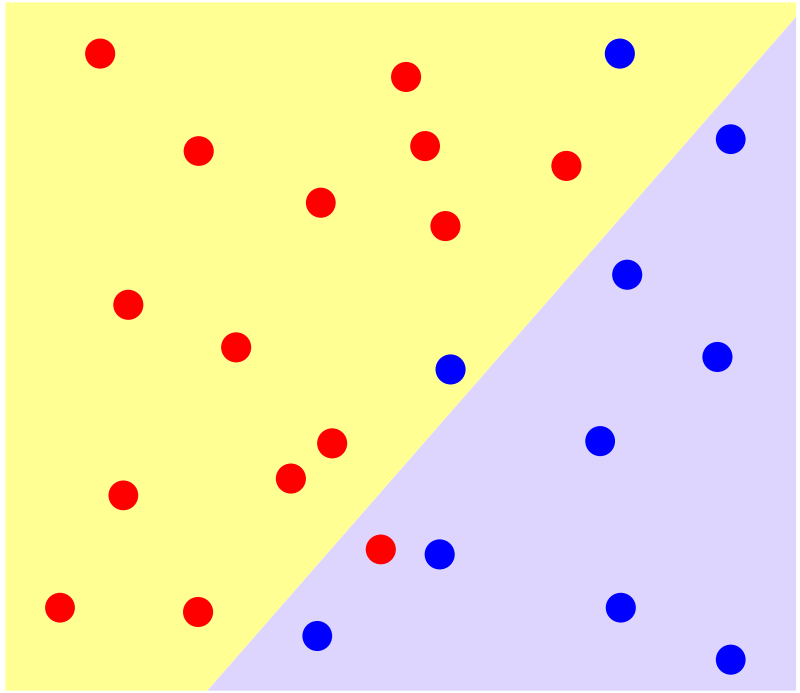
决策树（Decision Tree）

回归树　(1,2.3), (2,3.2), (3,3.7), (4,3.2), (5,4.5), (6,4.9), (7,5.2)

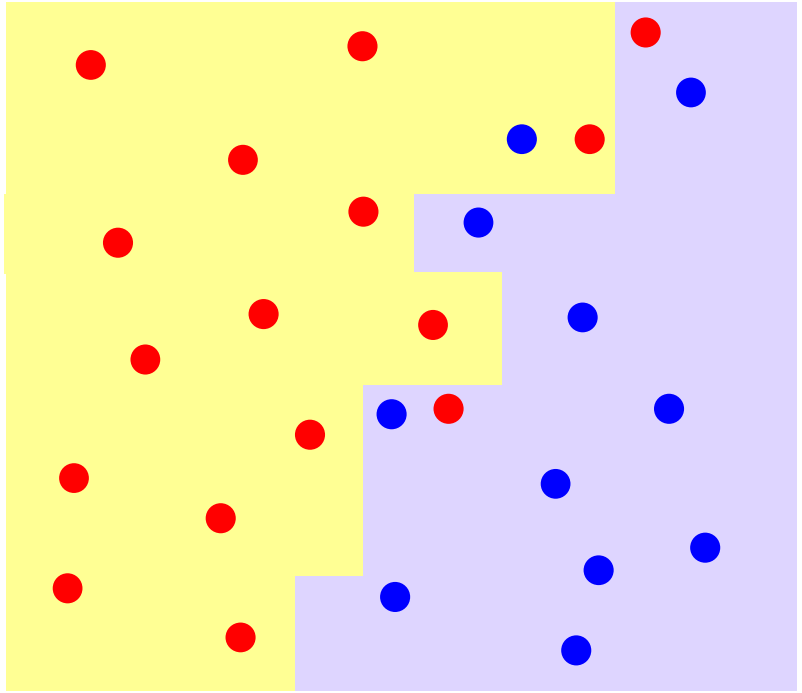


决策树（Decision Tree）

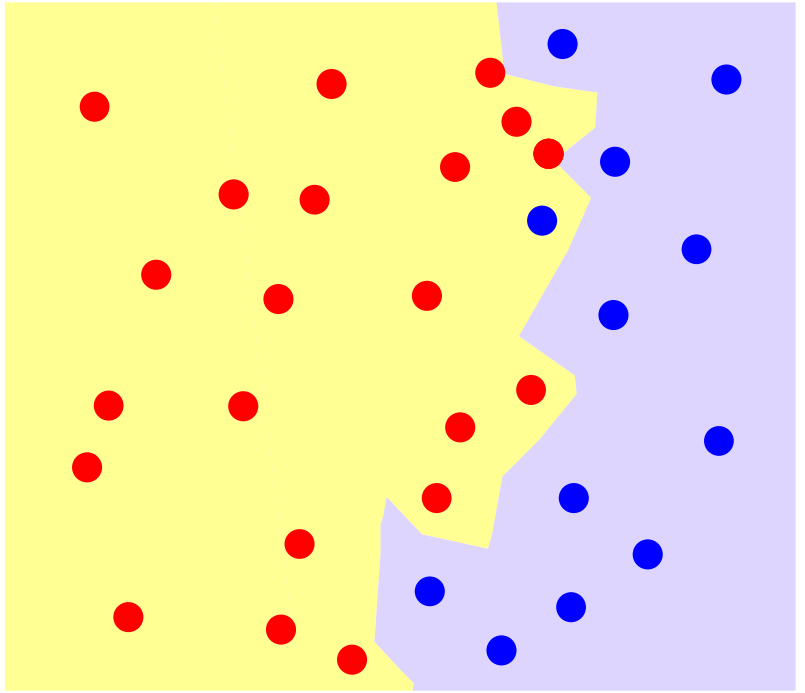
- 样本有两个特征，下图是三个分类器的分类区域和分类边界
- 逻辑回归、k近邻算法、决策树算法分别对应哪幅图?



(a)



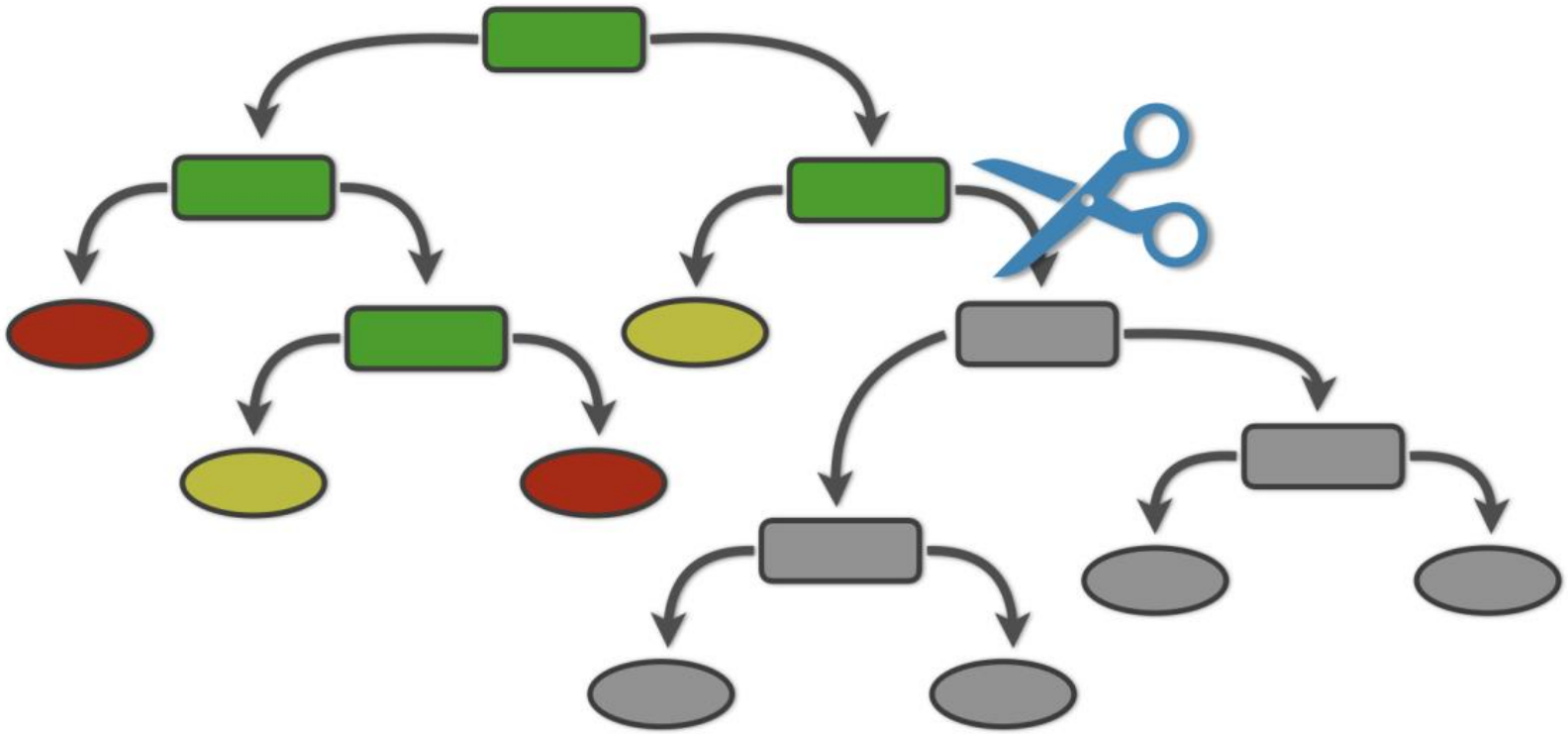
(b)



(c)

决策树（Decision Tree）

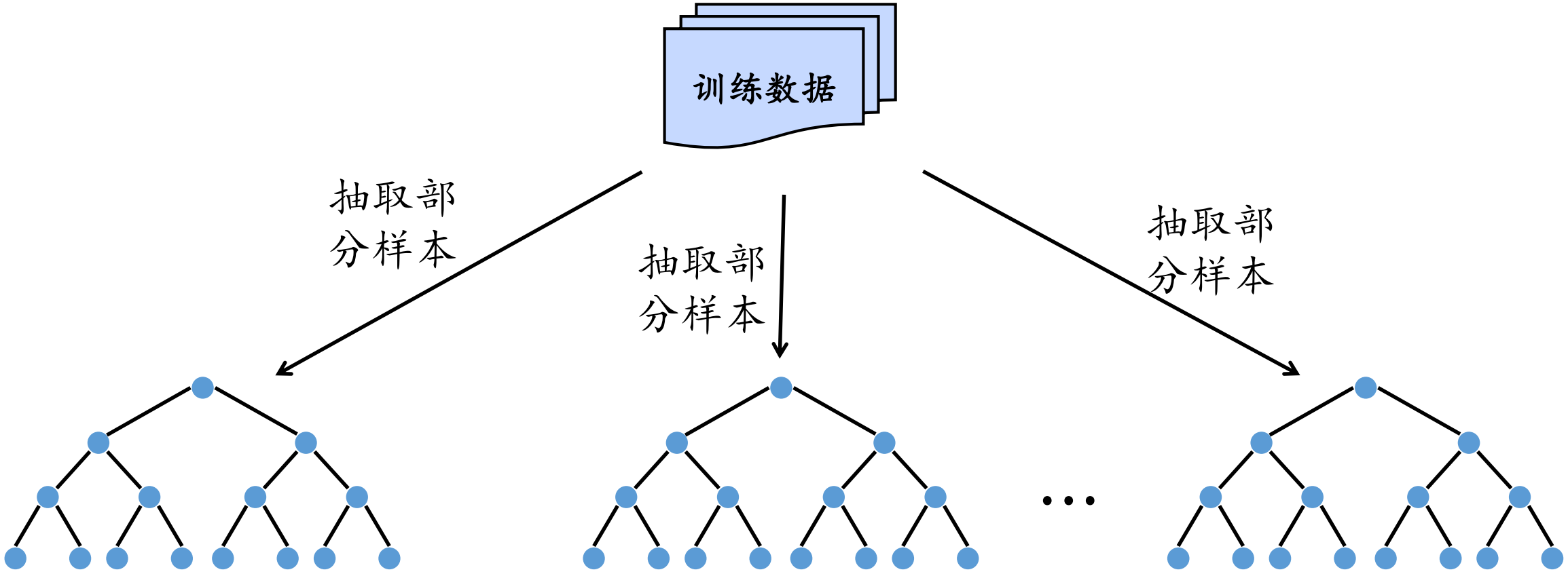
- 优点?
- 缺点: 过拟合
  - ✓ 解决方案：剪枝



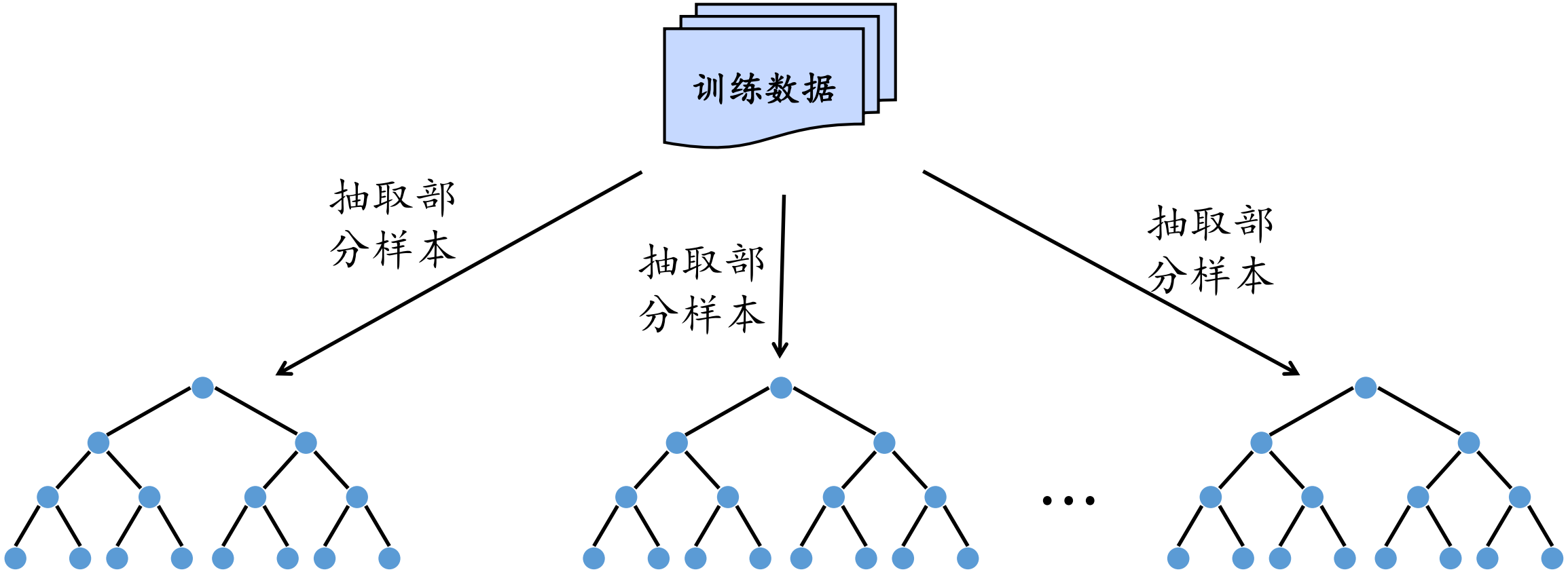
Random forest（随机森林）

- 无法使用单个决策树，因为无法解决过拟合问题。
- 构造多个决策树，通过投票方式作出最后预测。
- 每个决策树使用的训练数据一样，训练得出的决策树也一样。
- 多个决策树应该不一样，如何做到？（尽量让所有决策树不相关）

Random forest（随机森林）



Random forest（随机森林）

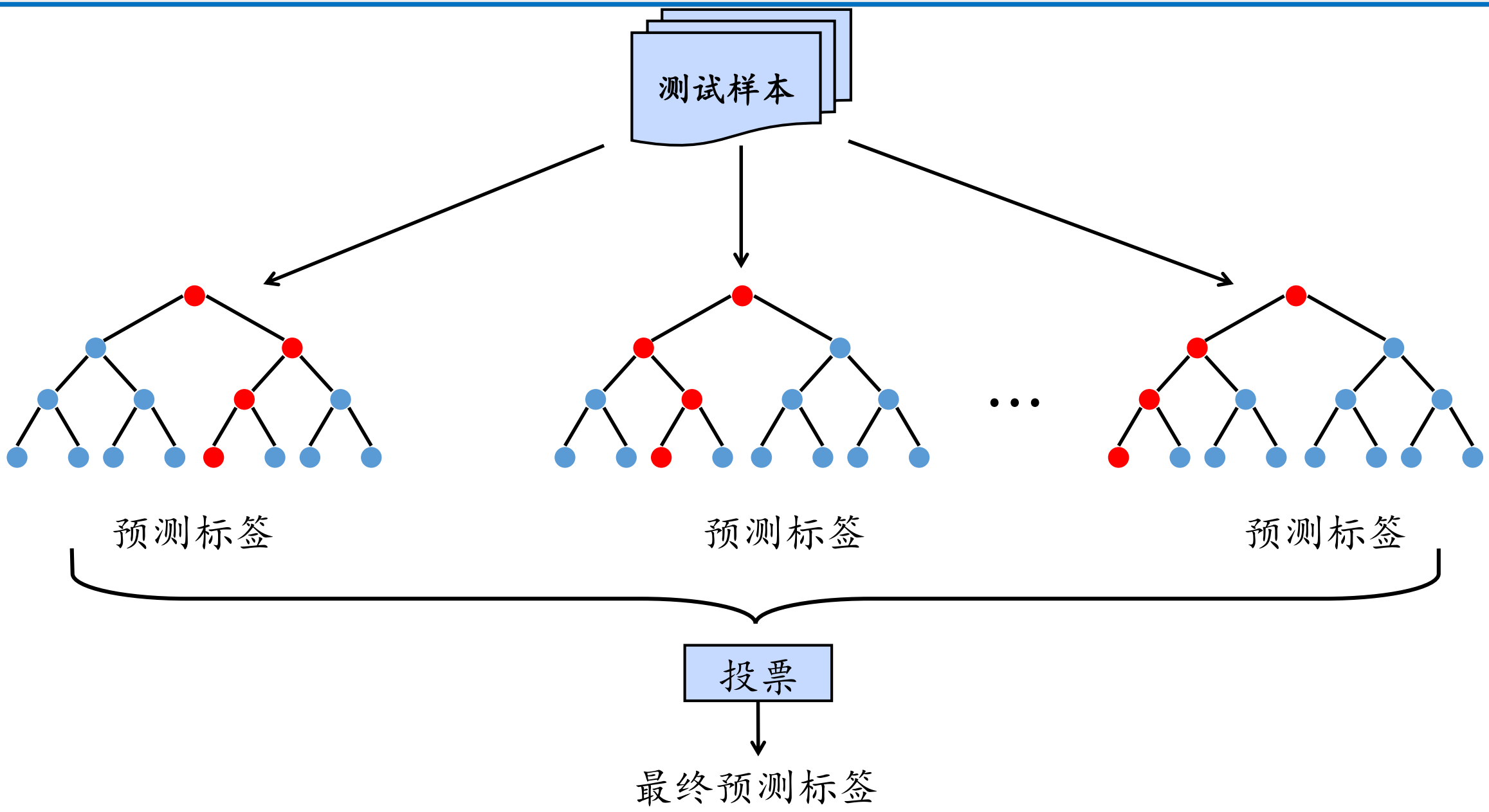


每次随机选择一部分特征，用于分裂

ID	年龄	有工作	有房产	信贷情况	是否通过
1	青年	否	否	一般	否
2	**青年**	**否**	**否**	**好**	**否**
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	**青年**	**否**	**否**	**一般**	**否**
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	**中年**	**是**	**是**	**好**	**是**
9	**中年**	**否**	**是**	**非常好**	**是**
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	**老年**	**是**	**否**	**非常好**	**是**
15	老年	否	否	一般	否

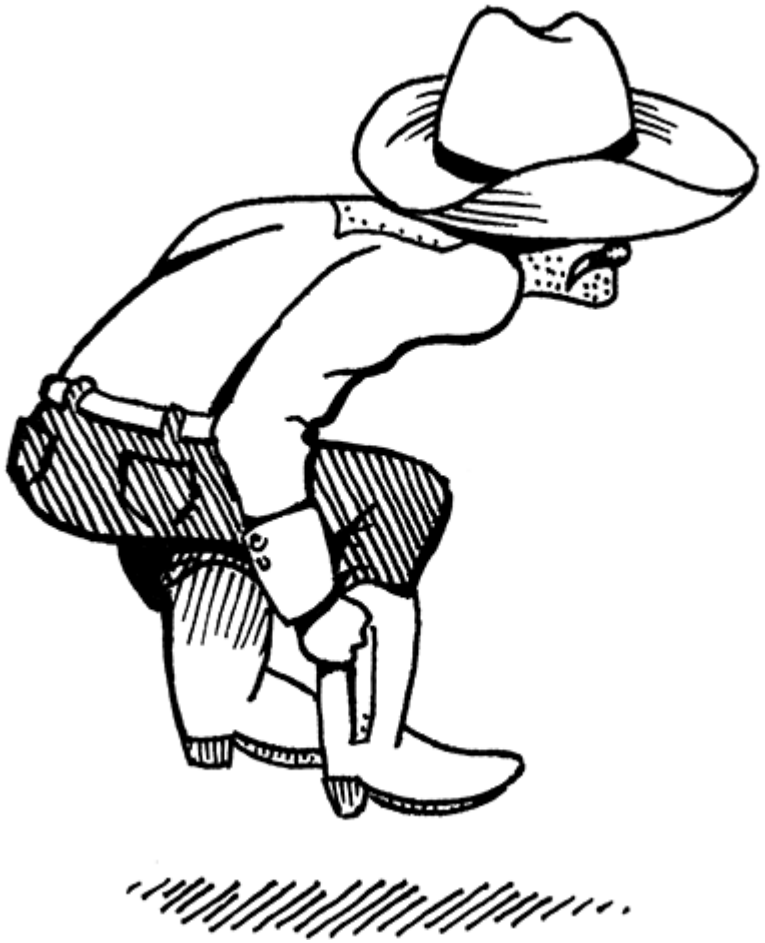
- 候选特征{年龄、房产}
- 抽取样本{2,5,8,9,14}

Random forest（随机森林）



Bootstrapping

Pull yourself up by your bootstraps.



Bootstrapping, Bagging, Ensemble, Boosting

- Bootstrapping: sampling with replacement. 1/e samples will never be picked
- 有放回的随机抽取方法，1/e样本不会被抽取到

- Bagging(bootstrapping aggregation): samples by bootstrapping. Train many models by these samples. Make prediction by voting.
- 通过bootstrapping方法产生训练样本，再训练多个同类型的分类器，通过通票产生最终结果

- Ensemble model: train many classifiers and make prediction by voting.
- 训练多个分类器，通过投票产生最终结果。

- Boosting(提升算法): combine many weak classifiers to get a strong classifier.
- 结合多个弱分类器，得到一个强分类器。