Python 3 玩儿转机器学习

liuyubobobo

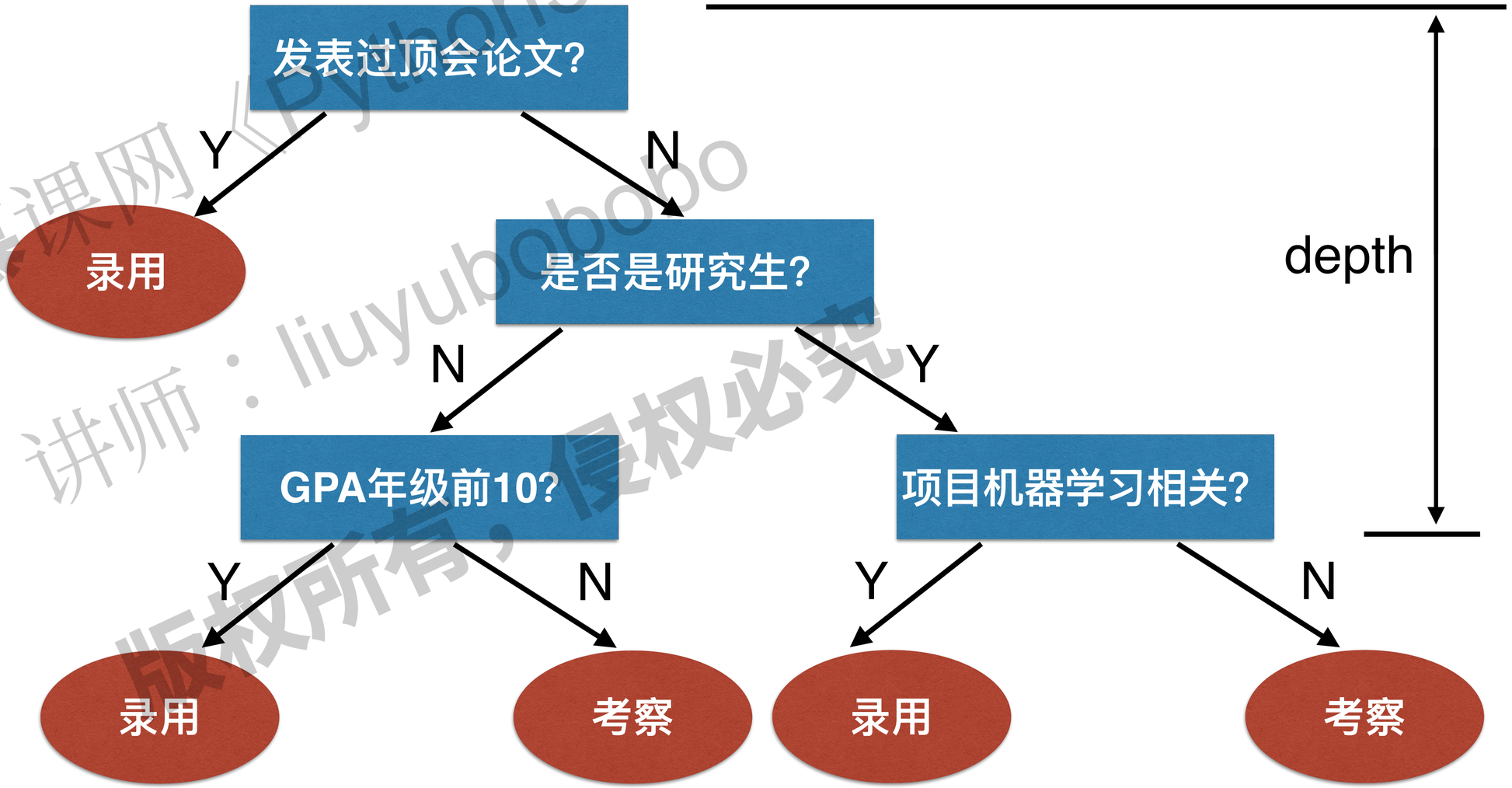
决策树

什么是决策树

什么是决策树

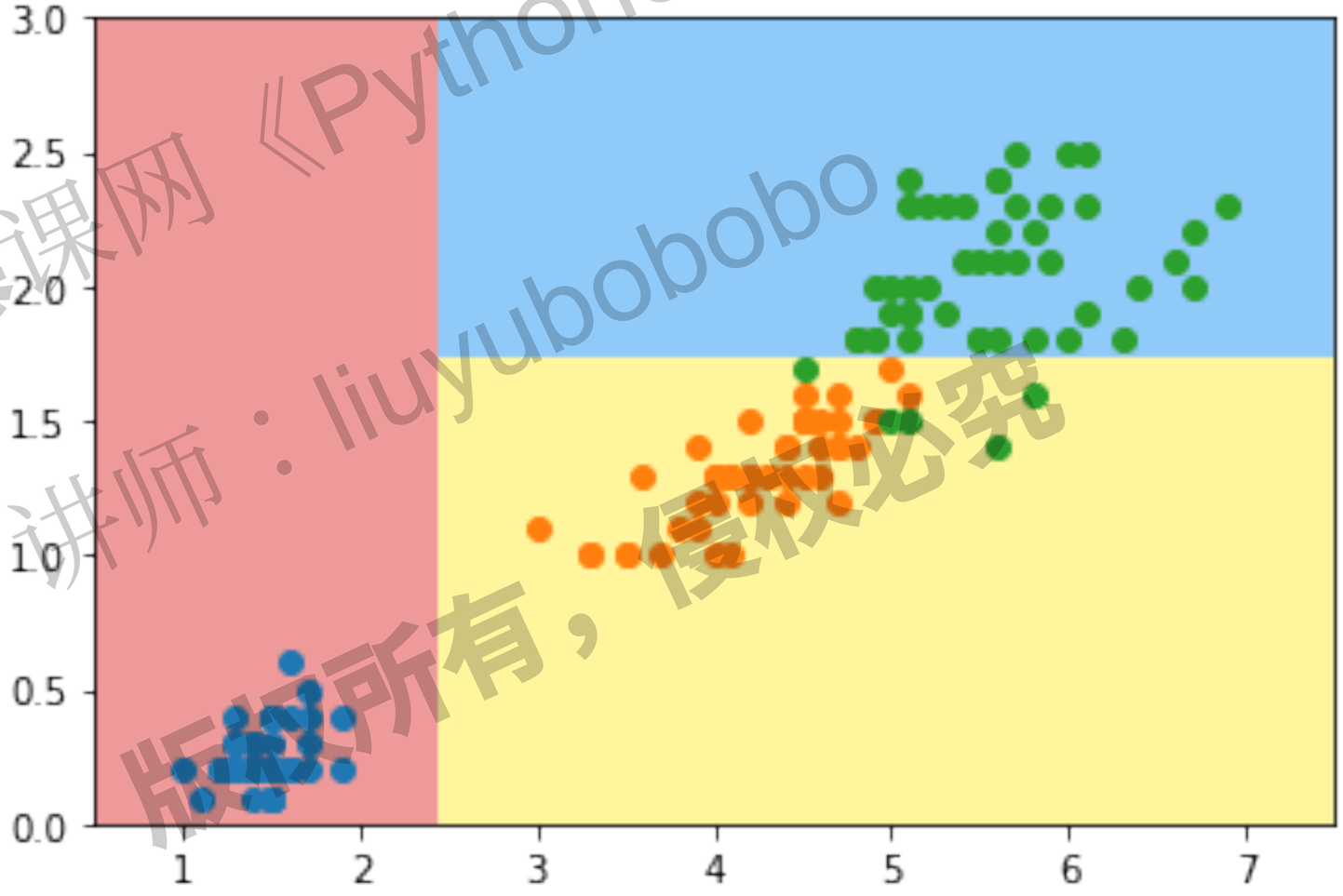
招聘机器学习

算法工程师

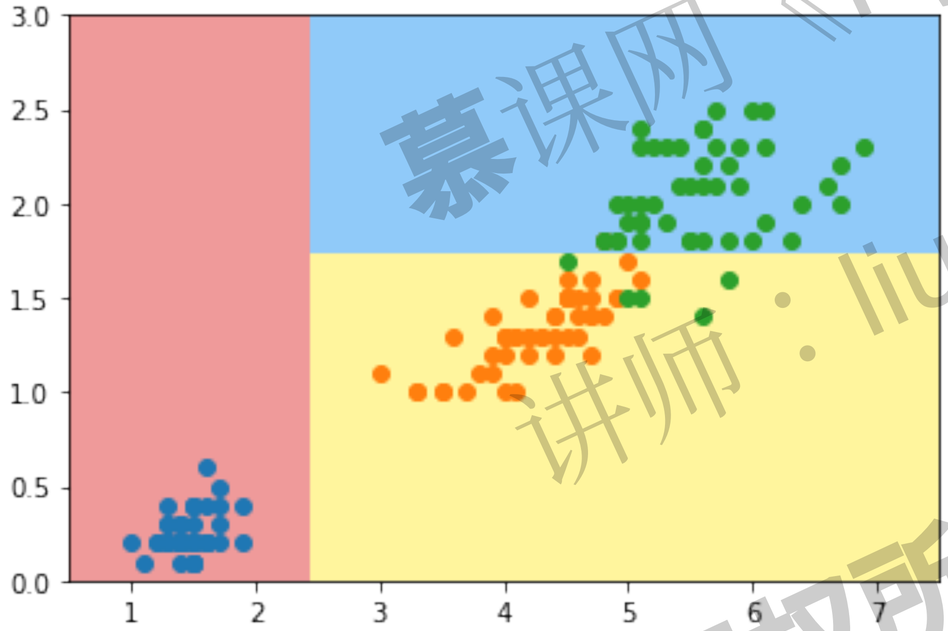


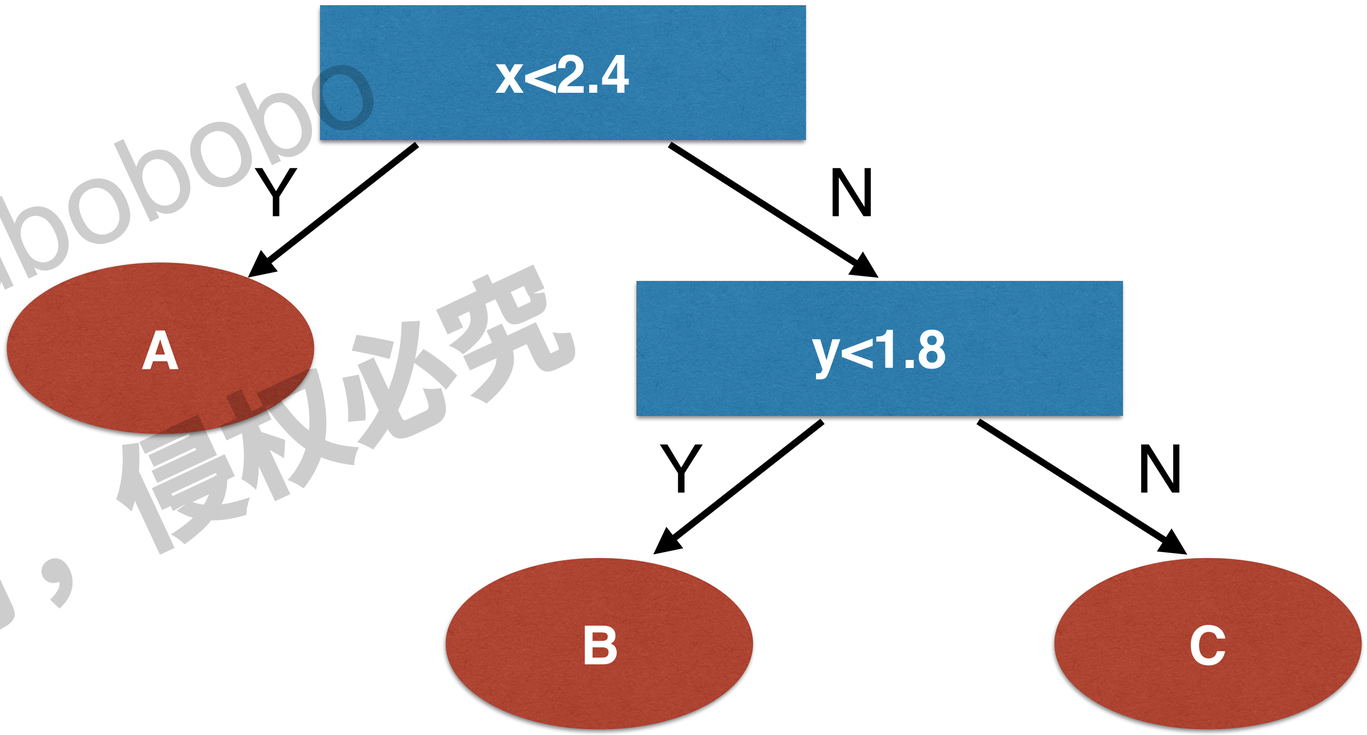
实践：scikit-learn中的决策树

什么是决策树

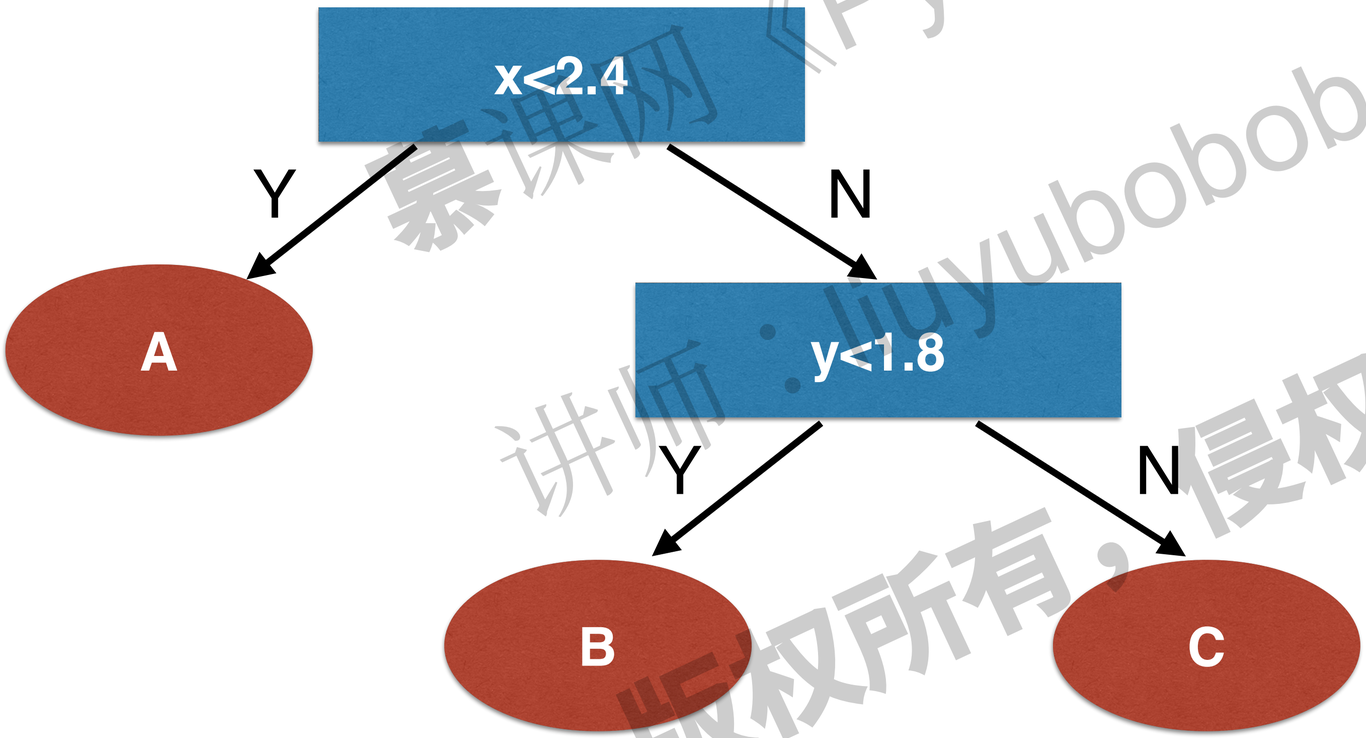


什么是决策树





什么是决策树



非参数学习算法

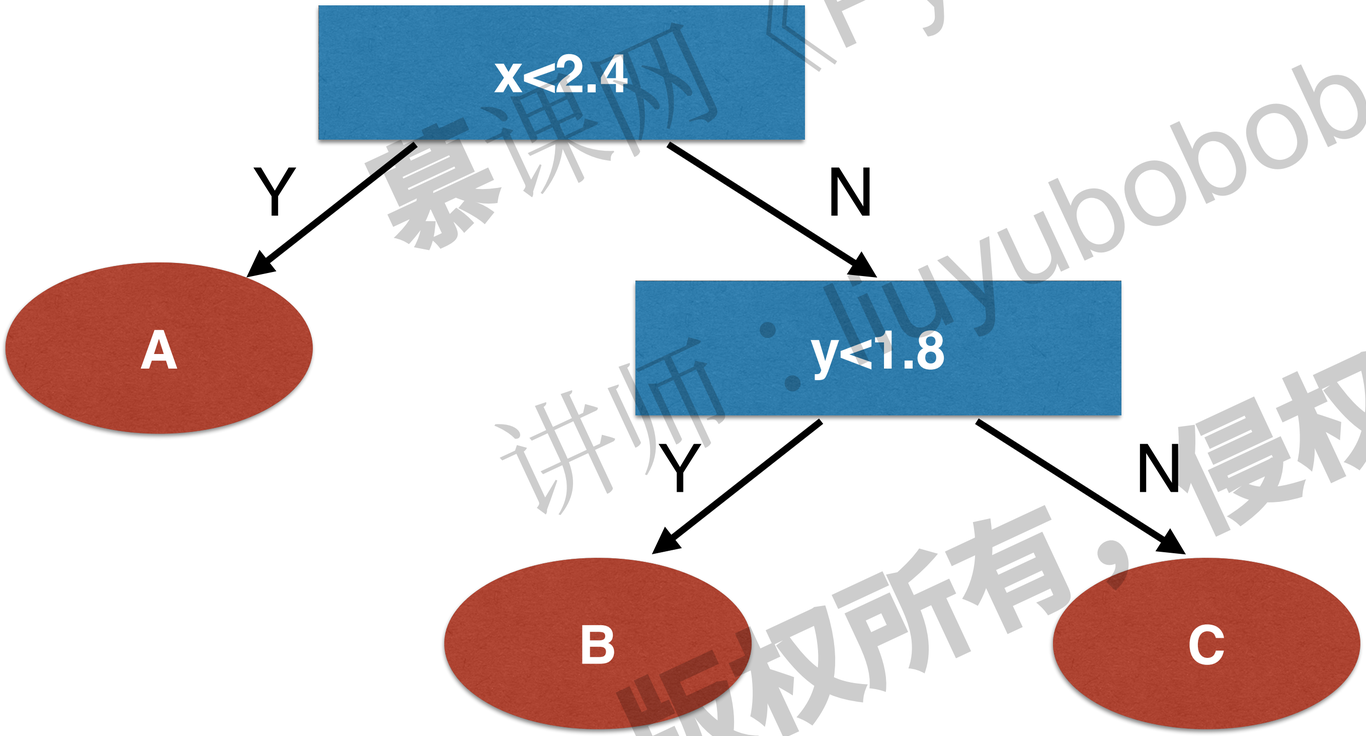
可以解决分类问题

天然可以解决多分类问题

也可以解决回归问题

非常好的可解释性

什么是决策树



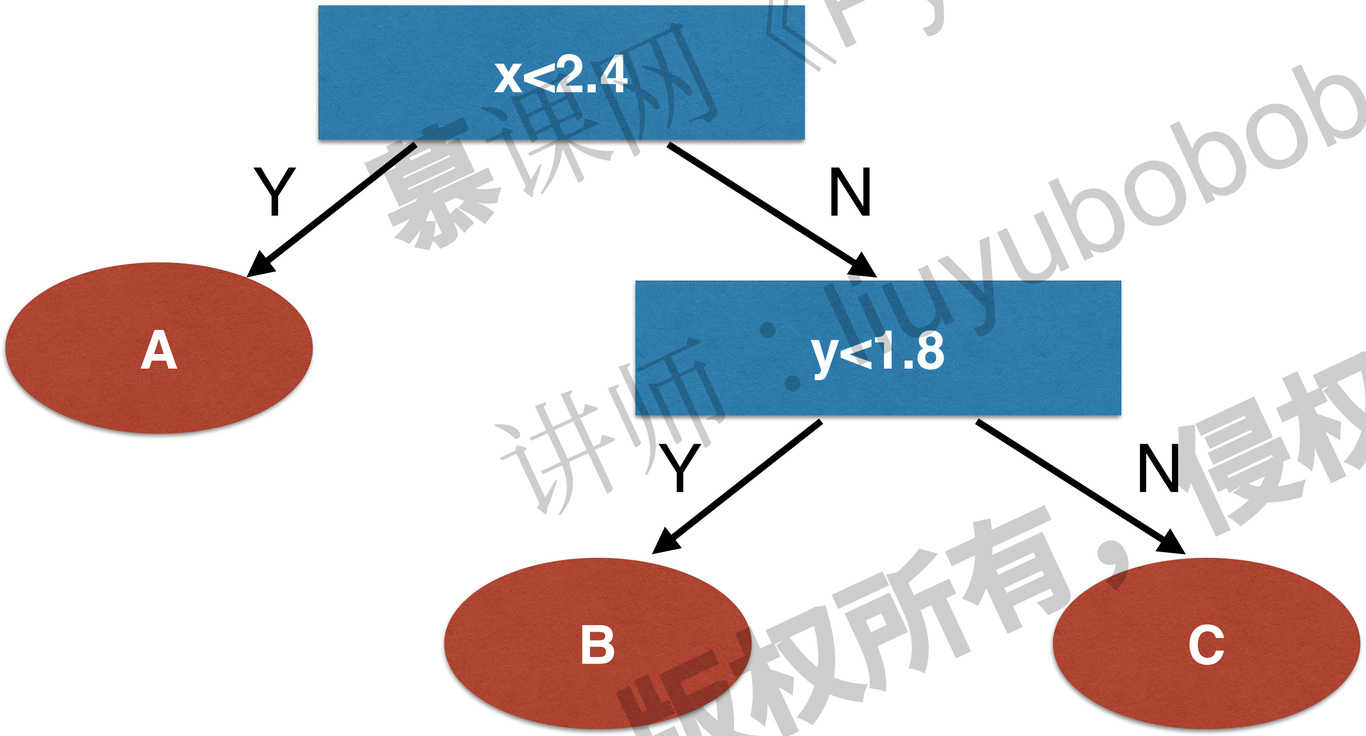
问题：

每个节点在哪个维度做划分？

某个维度在哪个值上做划分？

信息熵

信息熵

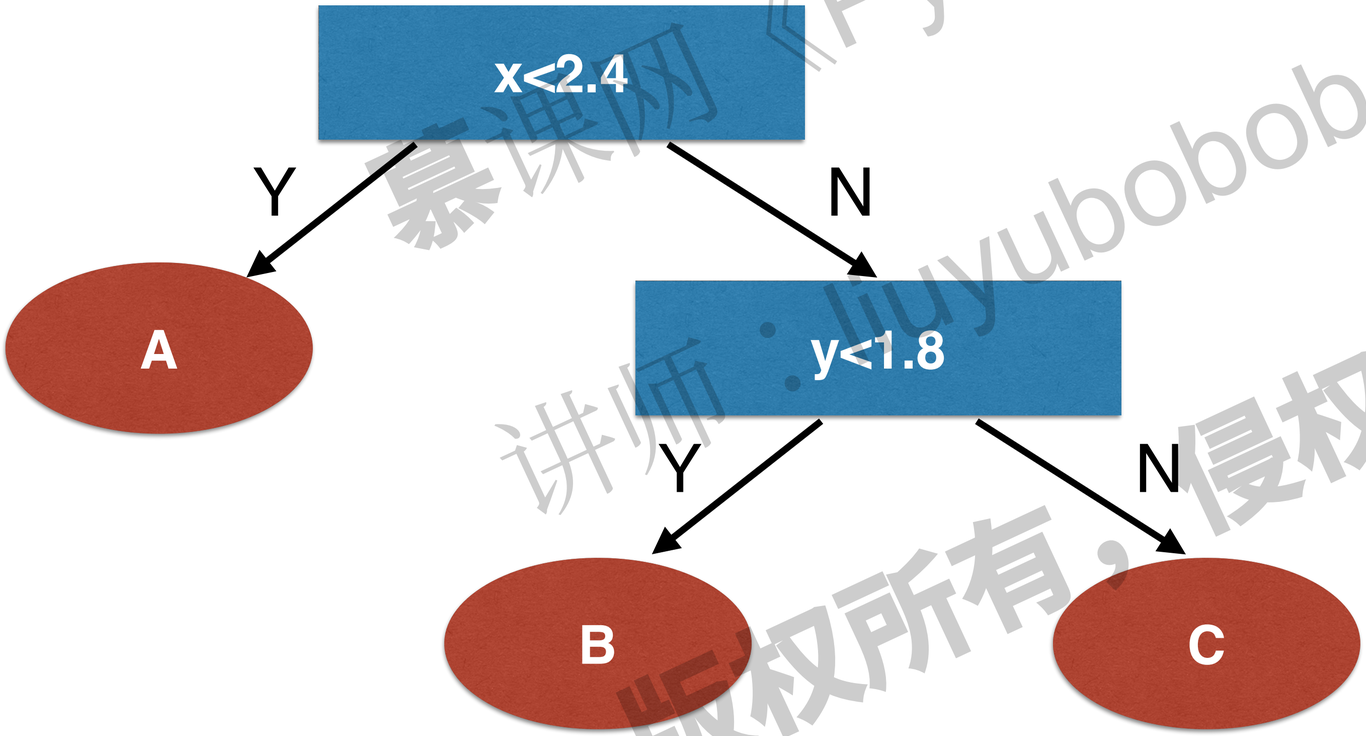


问题：

每个节点在哪个维度做划分？

某个维度在哪个值上做划分？

信息熵



熵在信息论中代表

随机变量不确定度的度量。

熵越大，数据的不确定性越高

熵越小，数据的不确定性越低

信息熵

熵在信息论中代表 随机变量不确定度的度量。

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

信息熵

熵在信息论中代表 随机变量不确定度的度量。

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

$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$

$$H = -\frac{1}{3}\log(\frac{1}{3}) - \frac{1}{3}\log(\frac{1}{3}) - \frac{1}{3}\log(\frac{1}{3})$$

$$= 1.0986$$

$\{\frac{1}{10}, \frac{2}{10}, \frac{7}{10}\}$

$$H = -\frac{1}{10}\log(\frac{1}{10}) - \frac{2}{10}\log(\frac{2}{10}) - \frac{7}{10}\log(\frac{7}{10})$$

$$= 0.8018$$

信息熵

熵在信息论中代表 随机变量不确定度的度量。

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

$\{\frac{1}{10}, \frac{2}{10}, \frac{7}{10}\}$

$$H = -\frac{1}{10}\log(\frac{1}{10}) - \frac{2}{10}\log(\frac{2}{10}) - \frac{7}{10}\log(\frac{7}{10})$$

$$= 0.8018$$

$\{1,0,0\}$

$$H = -1 \cdot \log(1) = 0$$

信息熵

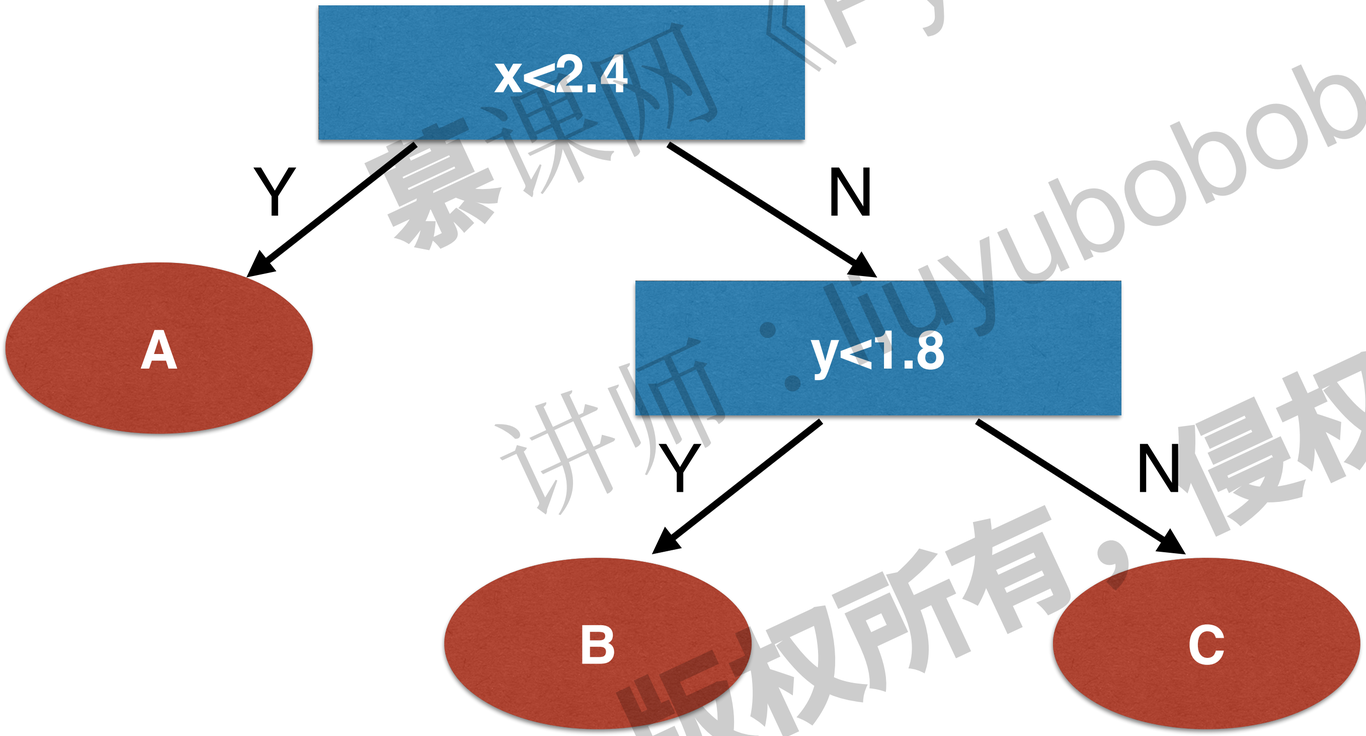
熵在信息论中代表 随机变量不确定度的度量。

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

$$H = -x\log(x) - (1-x)\log(1-x)$$

实践：信息熵的函数图像

信息熵



问题：

每个节点在哪个维度做划分？

某个维度在哪个值上做划分？

划分后使得信息熵降低

实践：模拟使用信息熵进行划分的过程

基尼系数

基尼系数

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

基尼系数

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$

$G = 1 - (\frac{1}{3})^2 - (\frac{1}{3})^2 - (\frac{1}{3})^2$

$= 0.6666$

$\{\frac{1}{10}, \frac{2}{10}, \frac{7}{10}\}$

$G = 1 - (\frac{1}{10})^2 - (\frac{2}{10})^2 - (\frac{7}{10})^2$

$= 0.46$

$\{1,0,0\}$

$G = 1 - 1^2 = 0$

基尼系数

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

$$G = 1 - x^2 - (1-x)^2$$

$$= 1 - x^2 - 1 + 2x - x^2$$

$$= -2x^2 + 2x$$

实践：模拟使用基尼系数进行划分的过程

信息熵 vs 基尼系数

信息熵的计算比基尼系数稍慢。

scikit-learn中默认为基尼系数。

大多数时候二者没有特别的效果优劣

CART

CART

Classification And Regression Tree

根据某一个维度d和某一个阈值v进行二分

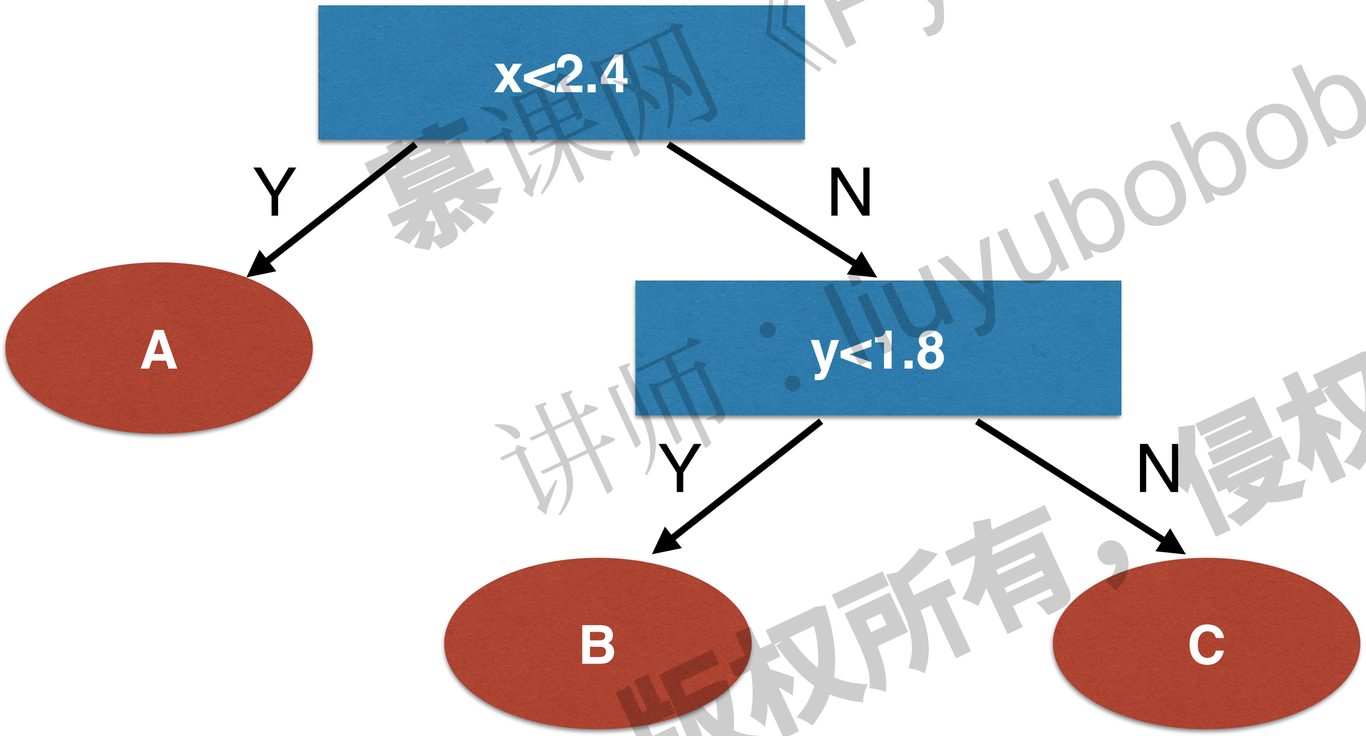
CART

scikit-learn的决策树实现：CART

ID3, C4.5, C5.0

http://scikit-learn.org/stable/modules/tree.html

复杂度



预测：O(logm)

训练：O(n*m*logm)

剪枝: 降低复杂度，解决过拟合

实践：决策树中的超参数

CART

min_samples_split

min_samples_leaf

min_weight_fraction_leaf

max_depth

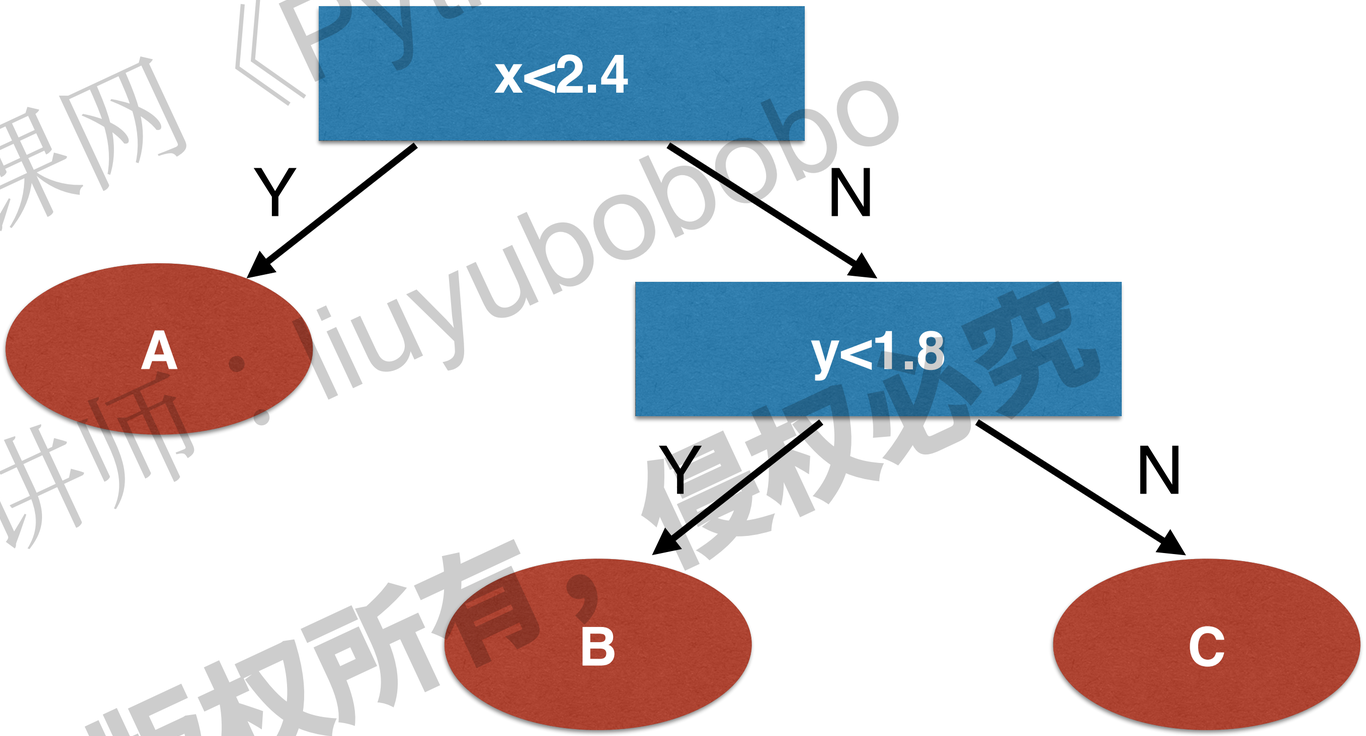
max_leaf_nodes

min_features

http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

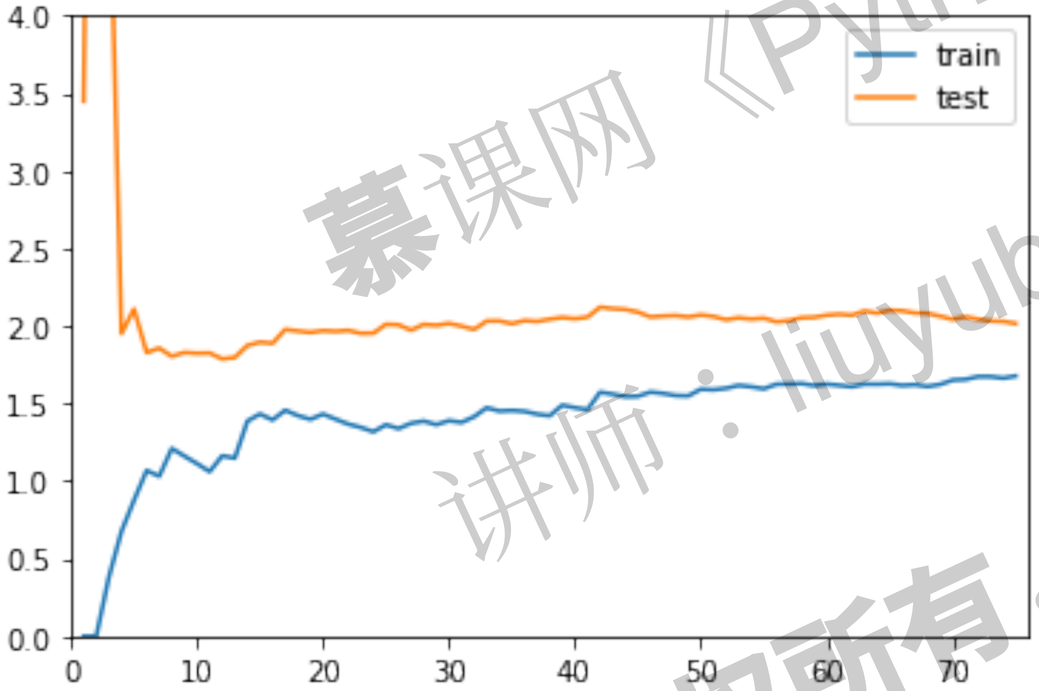
决策树解决回归问题

决策树解决回归问题

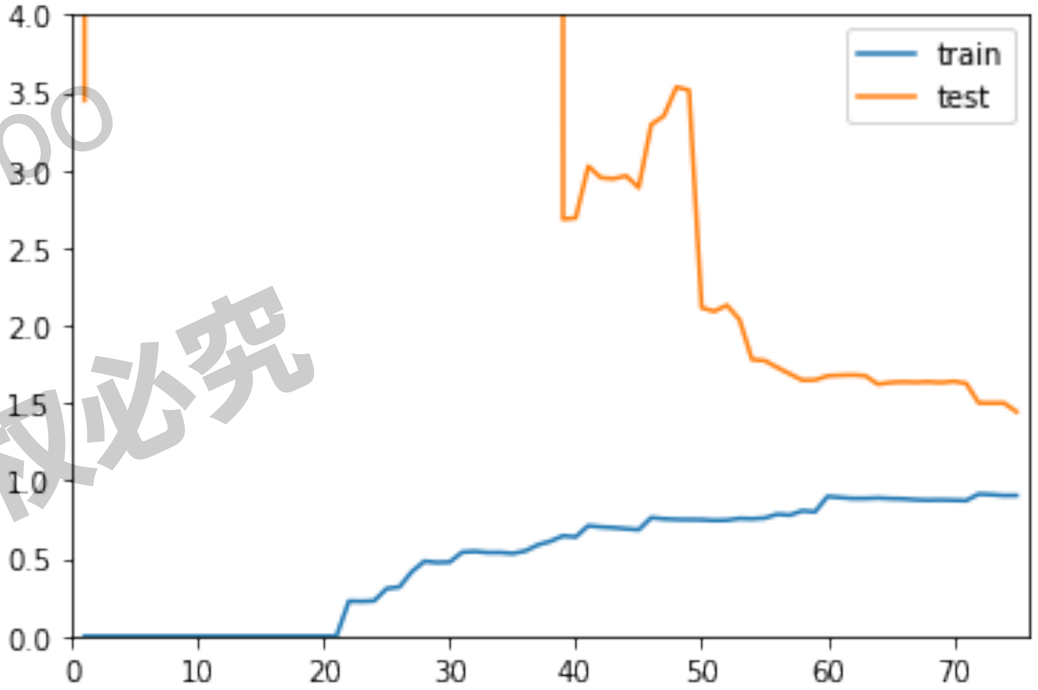


实践：scikit-learn中的决策树解决回归问题

学习曲线

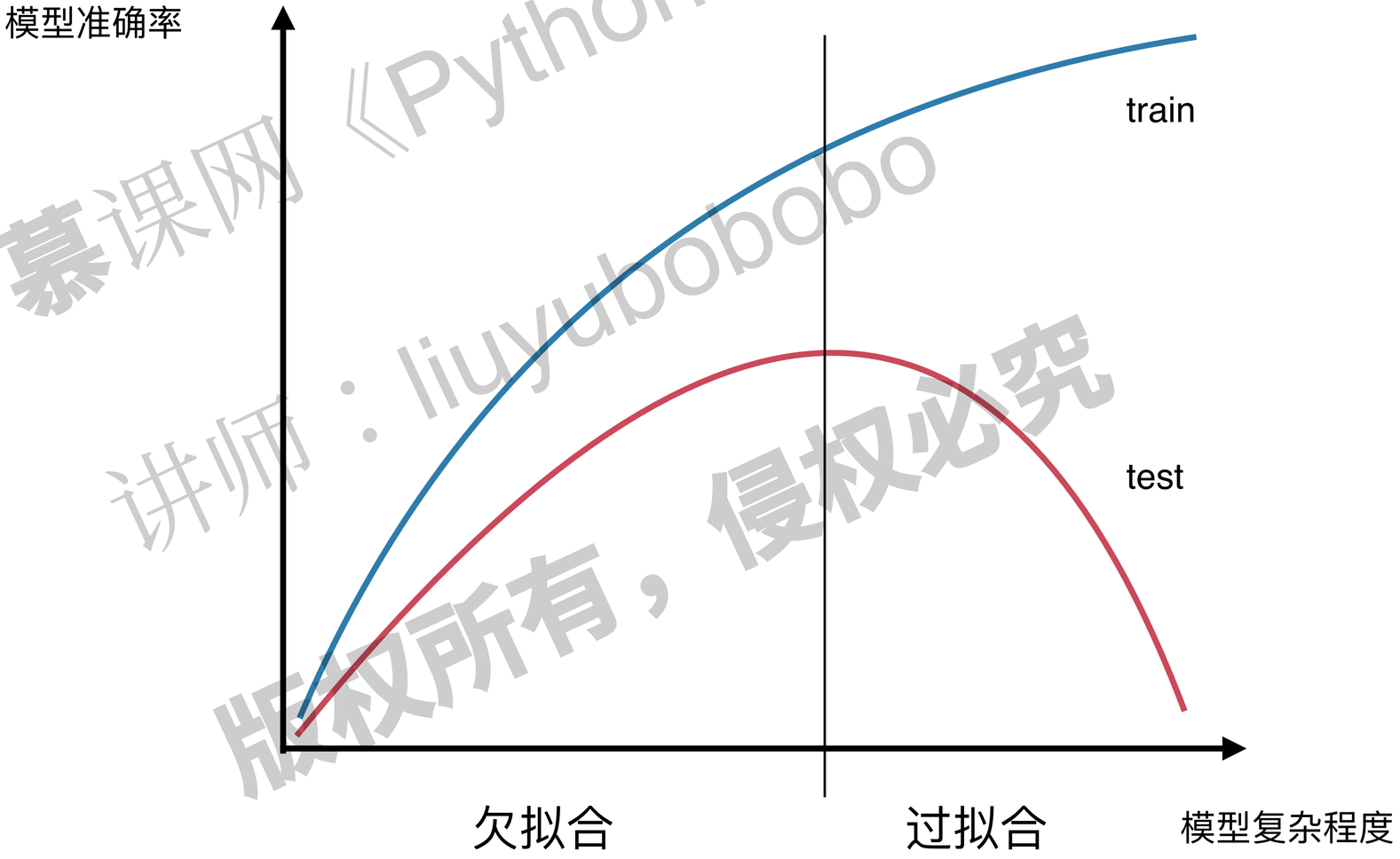


欠拟合



过拟合

模型复杂度曲线

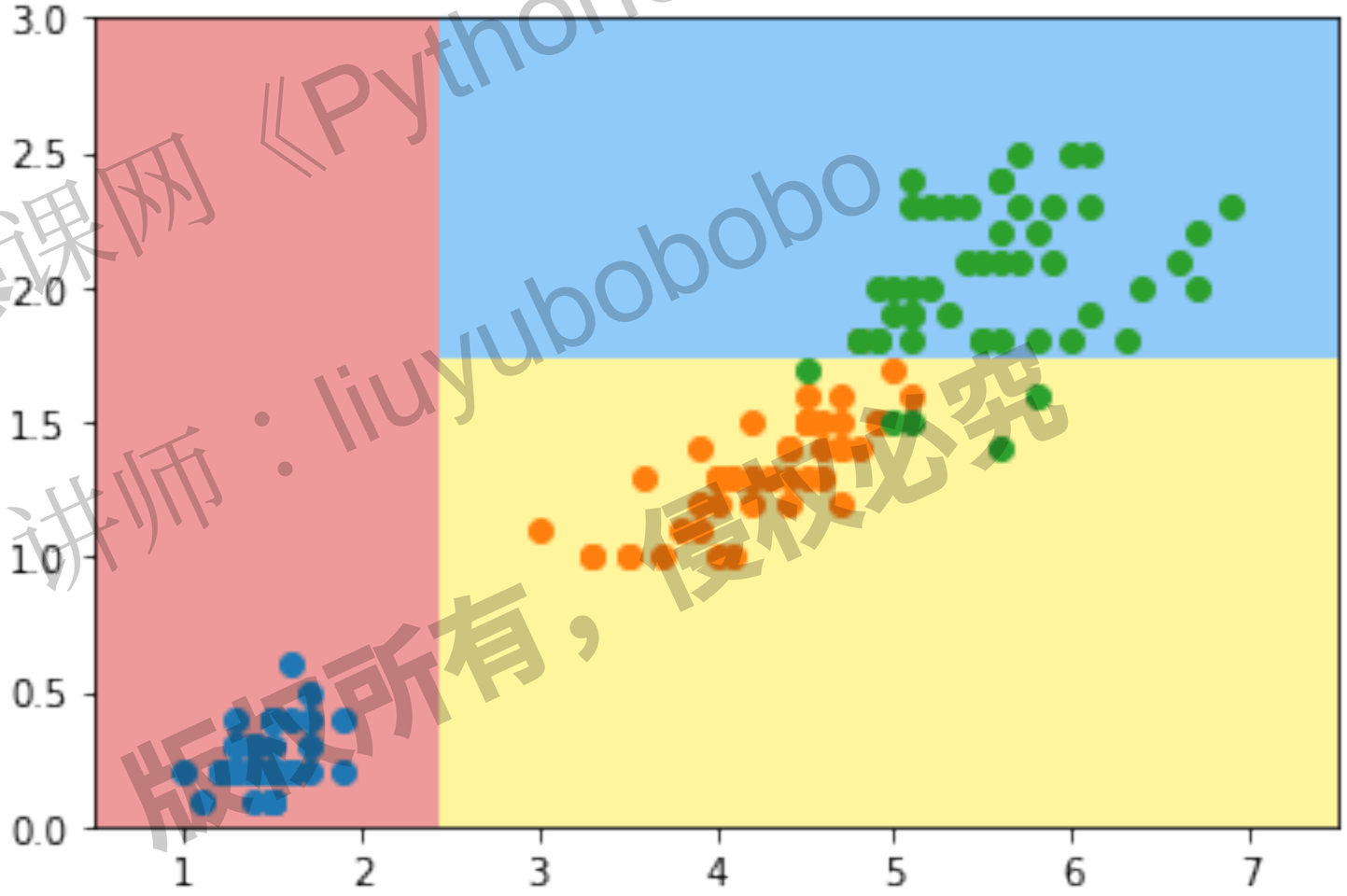


课程补充代码

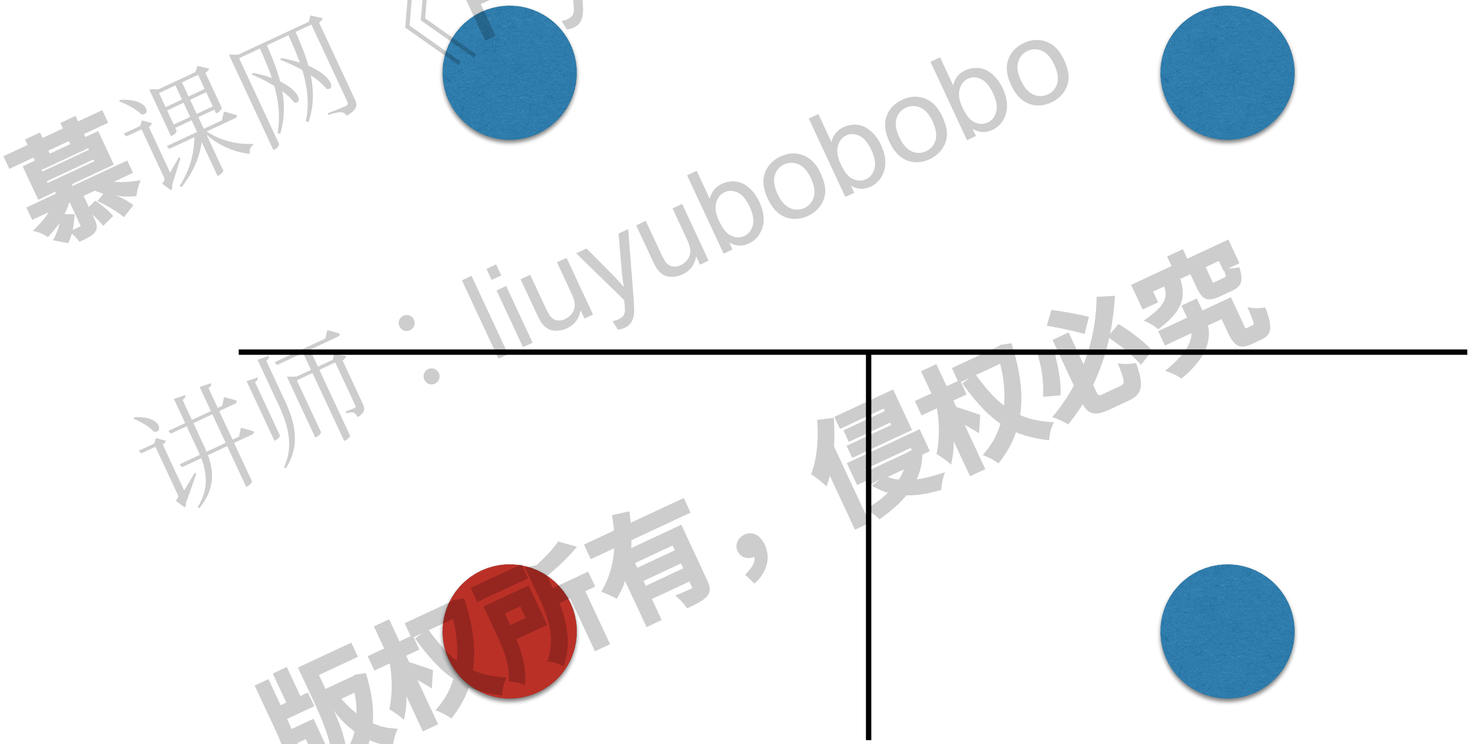
https://github.com/liuyubobobo/Play-with-Machine-Learning-Algorithms

决策树的局限性

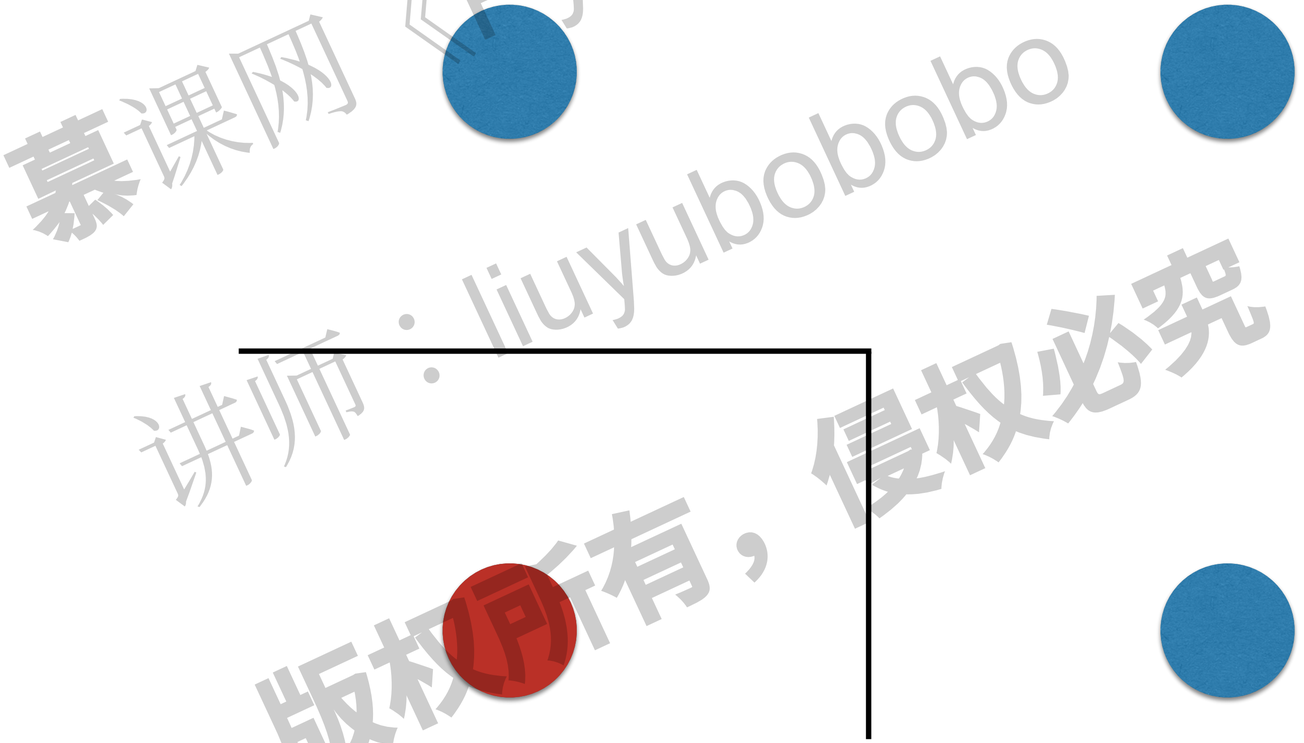
决策树的局限性



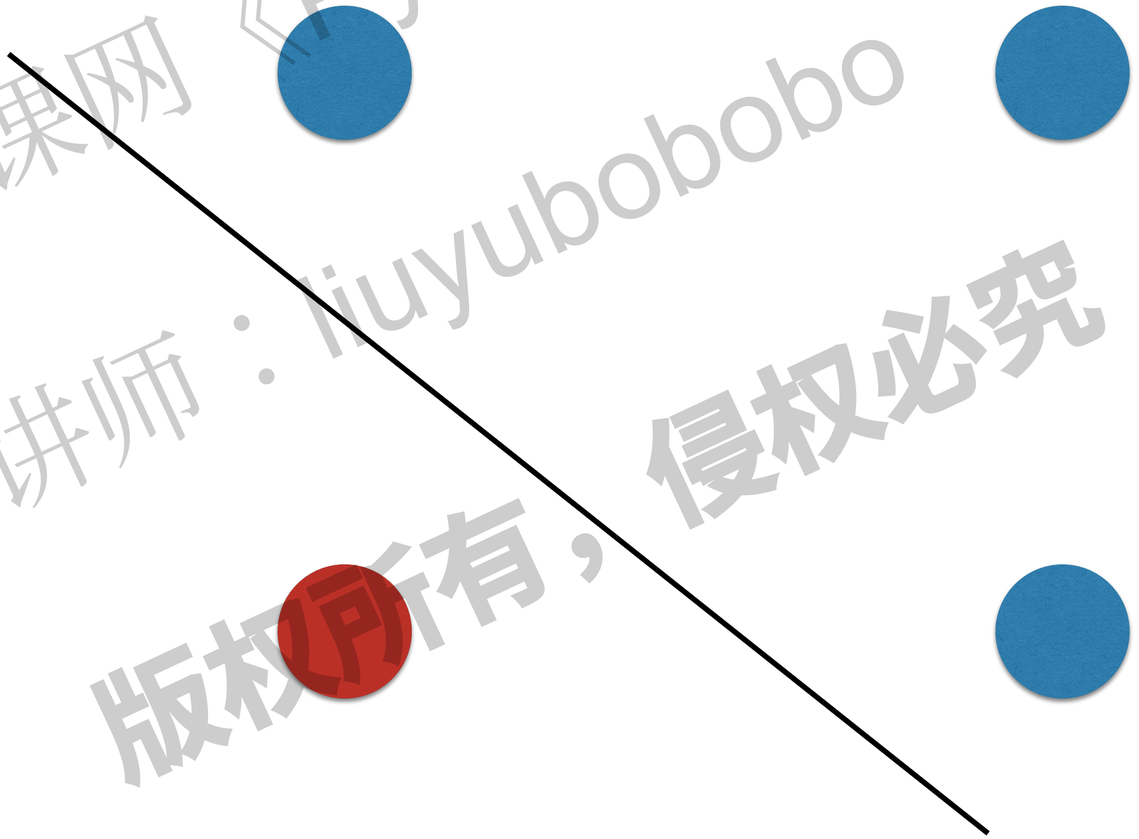
决策树的局限性



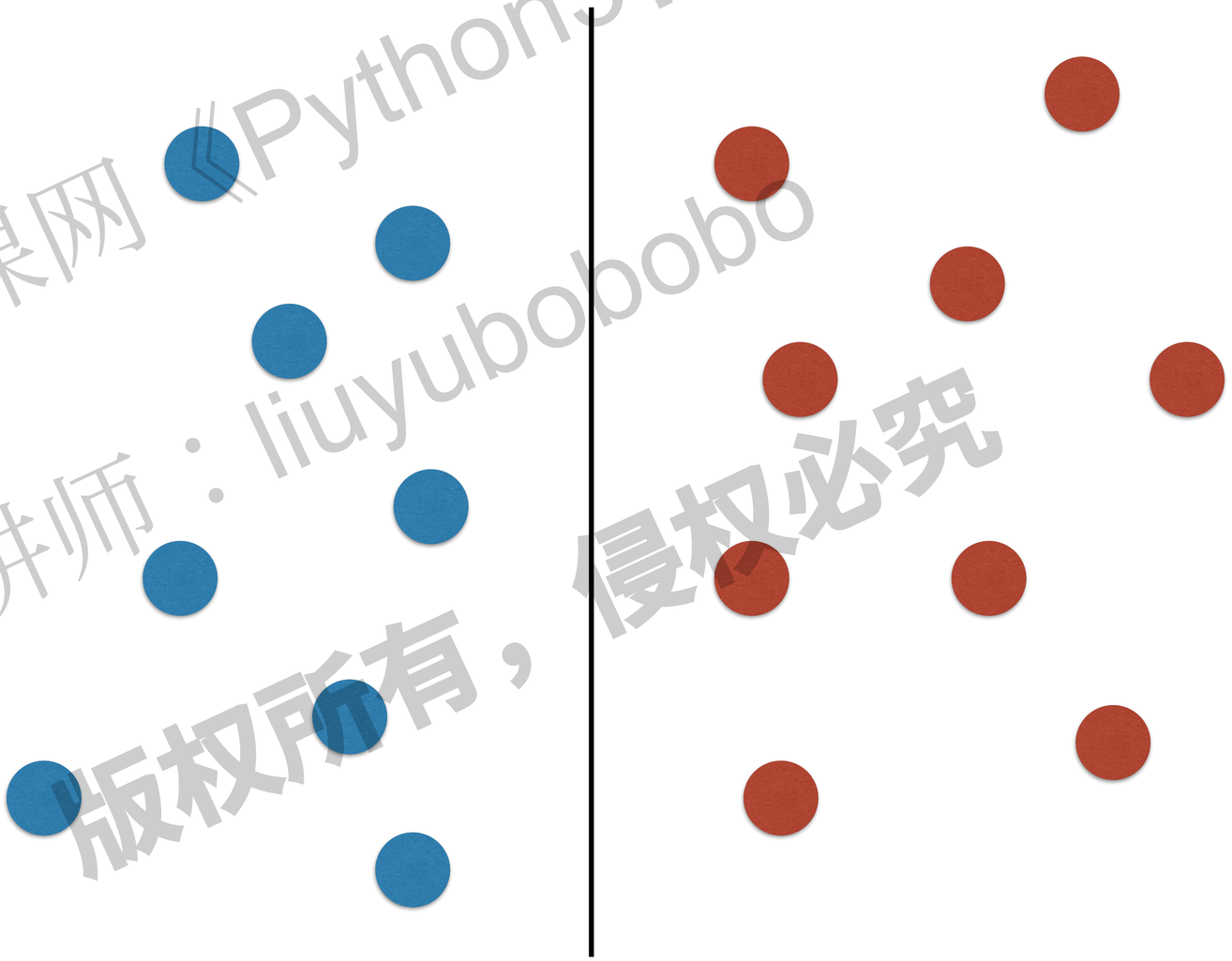
决策树的局限性

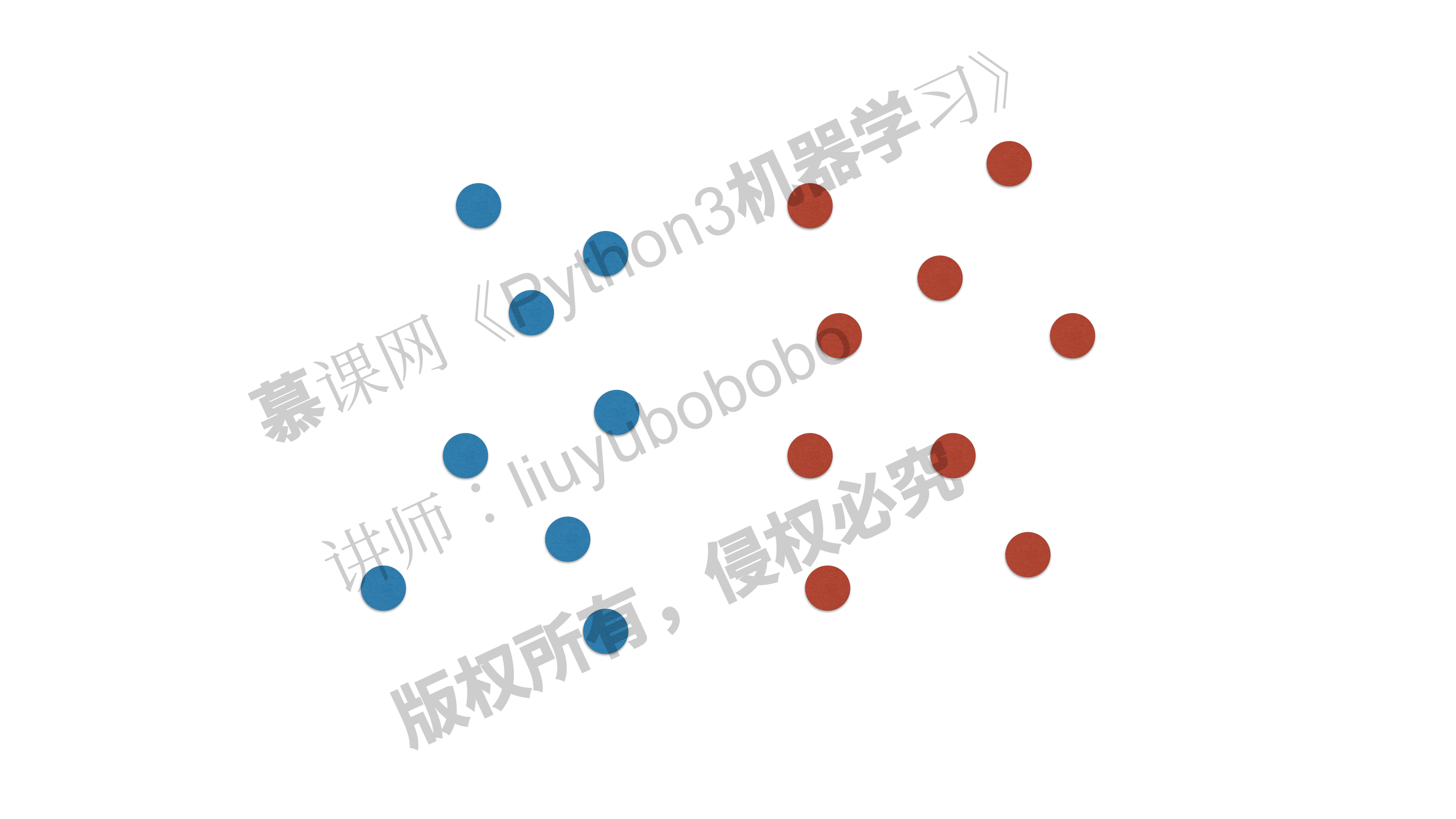


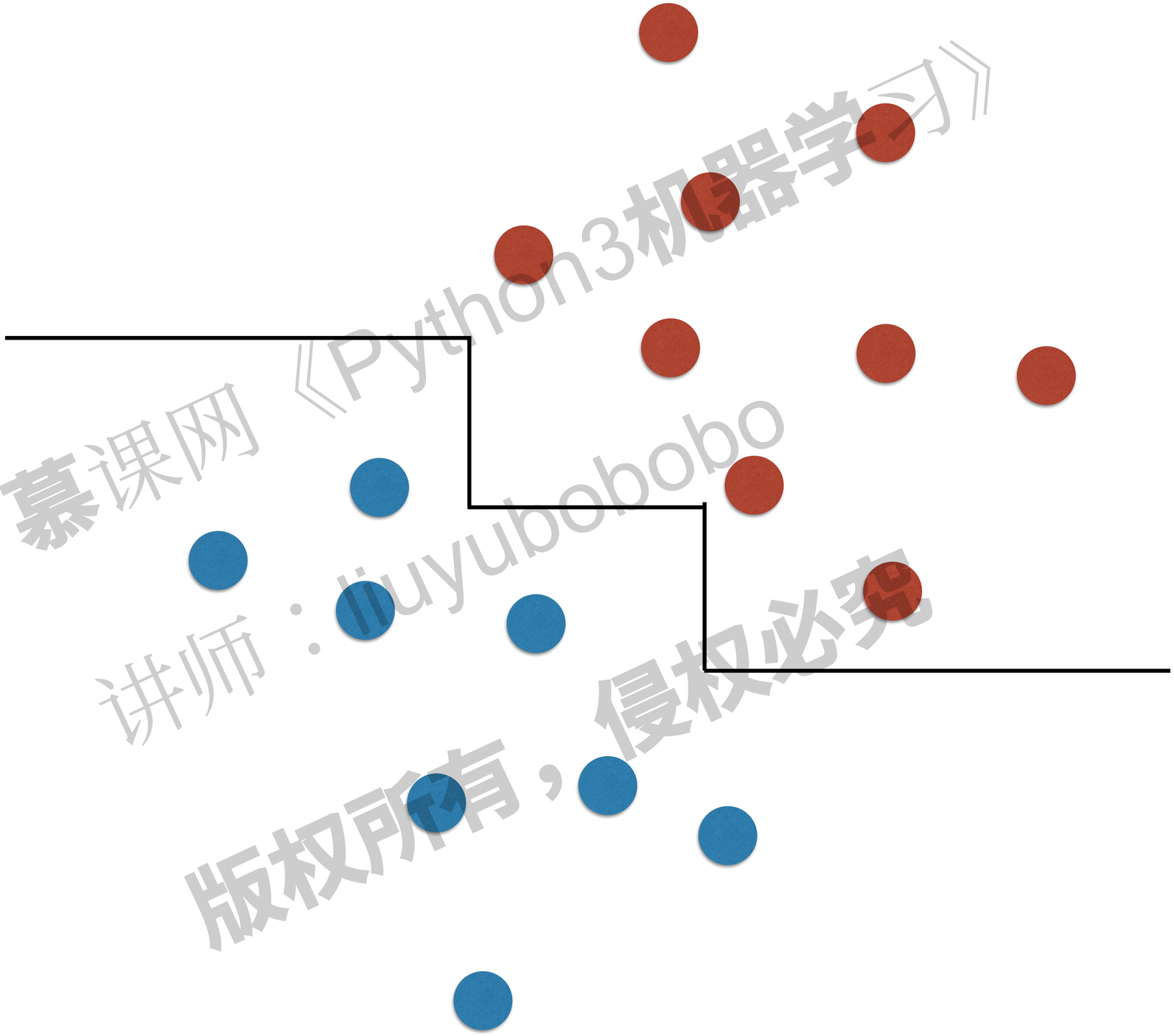
决策树的局限性



决策树的局限性




慕课网《Python3机器学习》
讲师：liuyubobobo
版权所有，侵权必究


慕课网《Python3机器学习》
讲师：liuyubobobo
版权所有，侵权必究

决策树的局限性

对个别数据敏感

实践：决策树对个别数据敏感

其他

欢迎大家关注我的个人公众号：是不是很酷

Python 3 玩儿转机器学习

liuyubobobo