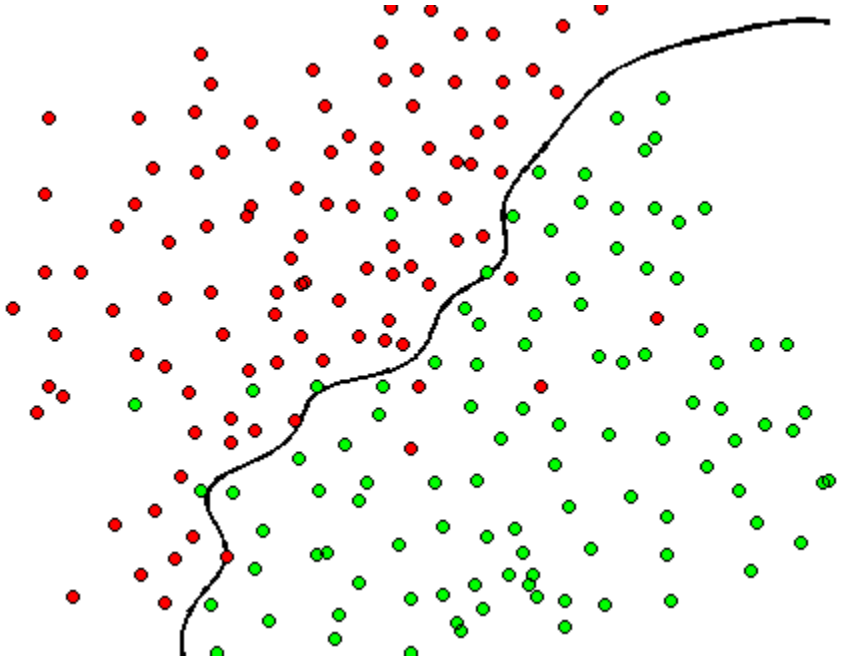
法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 **小象学院**



分类模型(1)

--Robin

目录

- kNN
- 决策树
- 朴素贝叶斯

决策树

- 例子1

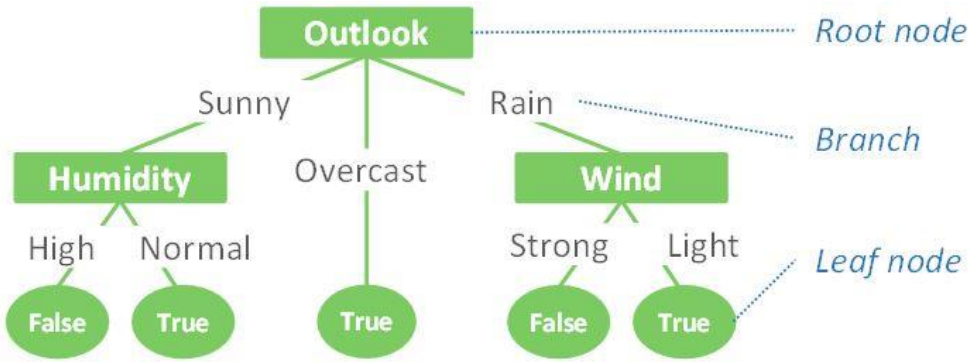

决策树例子
Is it alive?
No
Yes
Does it fly?
No
Yes
Carry > 10 people?
No
Yes
Objects with: Alive == No AND Fly == No AND Carry > 10 == No

决策树

- 例子2

Playing Tennis Example

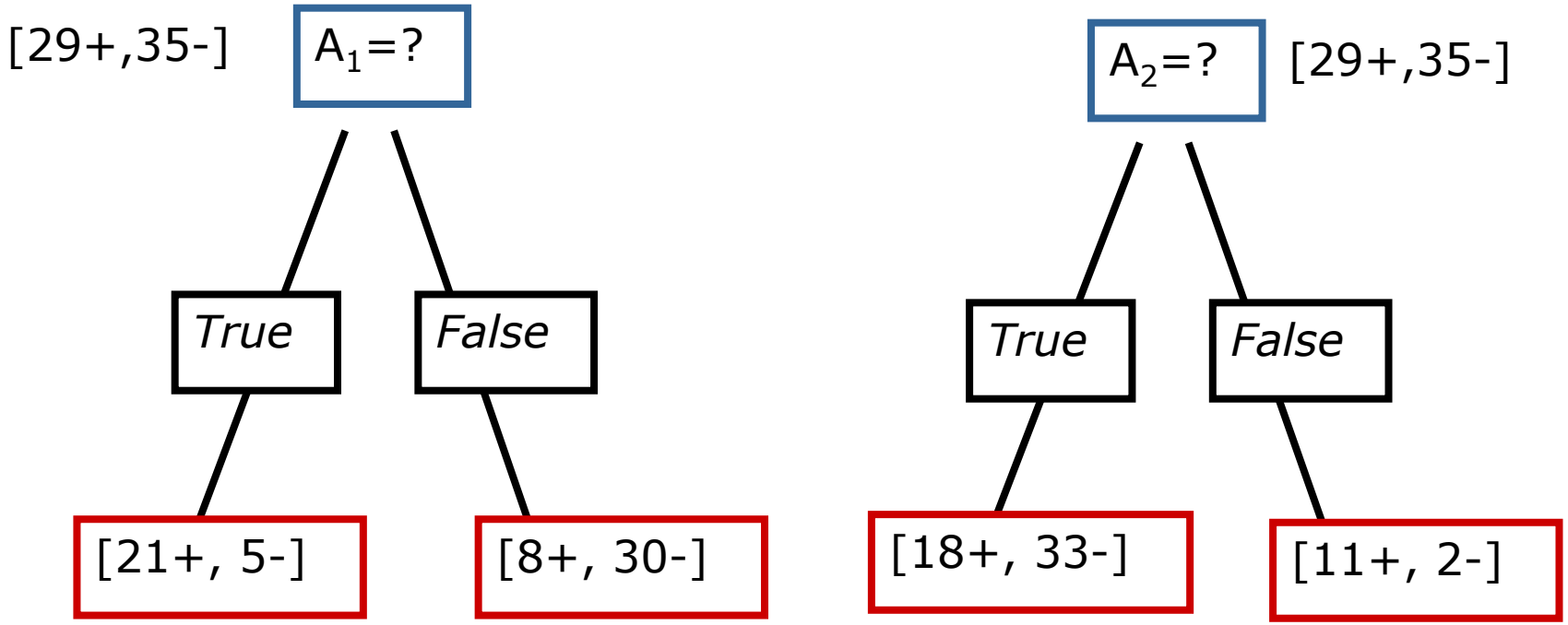
Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



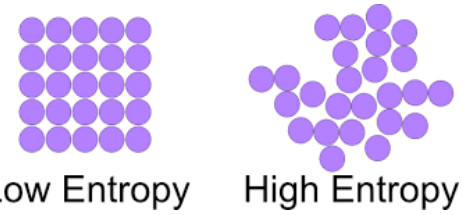
决策树

思考

- 按什么标准选择特征？
- 选择A_1还是A_2作为分割点？



决策树



预备知识

- 熵（Entropy）

在信息论中，设离散随机变量 的概率分布为 ，则概率分布的熵(Entropy)的定义为：

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

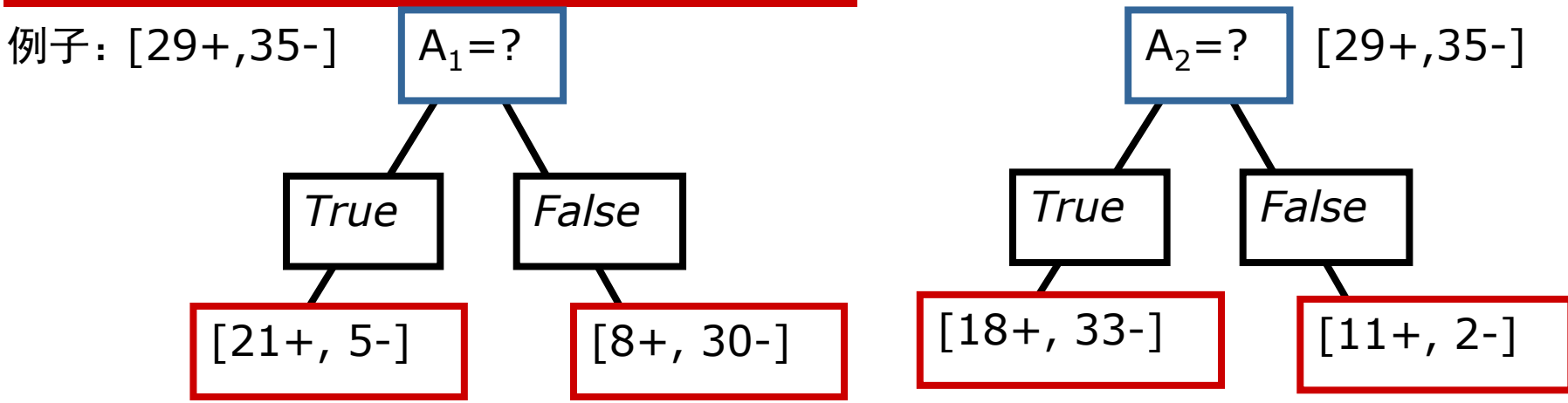
scipy.stats.entropy()

- 信息增益（Information Gain）

描述了当使用Q进行编码时，再使用P进行编码的差异。在决策树算法中，信息增益是针对某个特征而言的，就是看一个特征A，系统有它和没它的时候信息量各是多少，两者的差值就是这个特征给系统带来的信息量，即增益

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

决策树



Entropy([29+,35-]) = -29/64 $\log_2$ 29/64 – 35/64 $\log_2$ 35/64 = 0.99

Entropy([21+,5-]) = 0.71

Entropy([8+,30-]) = 0.74

Gain(S, A_1)=Entropy(S)

-26/64*Entropy([21+,5-])

-38/64*Entropy([8+,30-])

=0.27

Entropy([18+,33-]) = 0.94

Entropy([11+,2-]) = 0.62

Gain(S,A_2)=Entropy(S)

-51/64*Entropy([18+,33-])

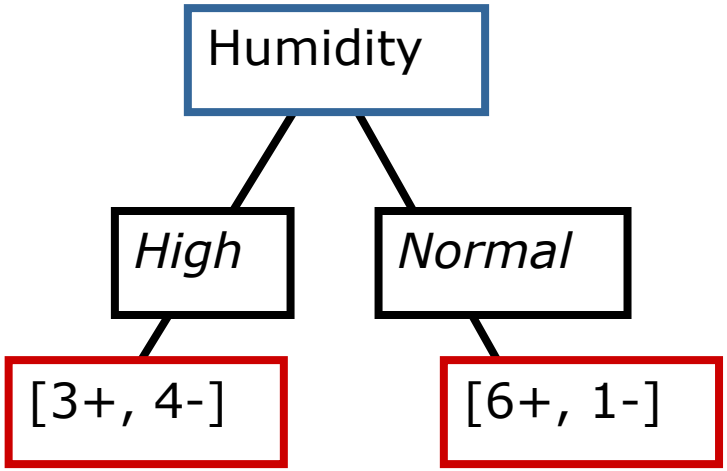
-13/64*Entropy([11+,2-])

=0.12

决策树

例子：

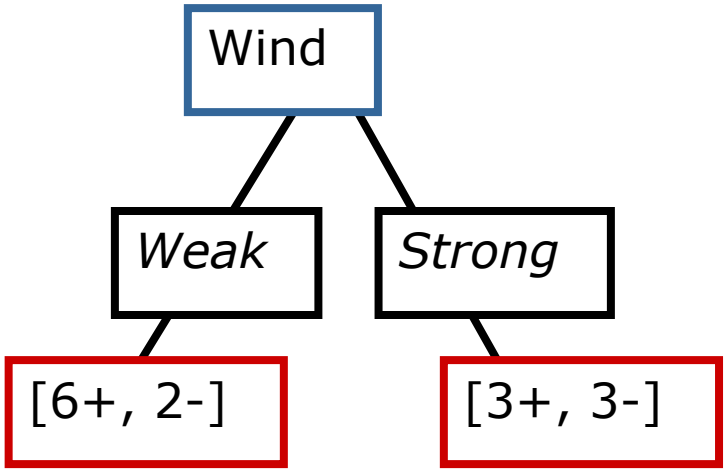
S=[9+,5-]
E=0.940



E=0.985　　E=0.592

Gain(S,Humidity)
=0.940-(7/14)*0.985
– (7/14)*0.592
=0.151

S=[9+,5-]
E=0.940



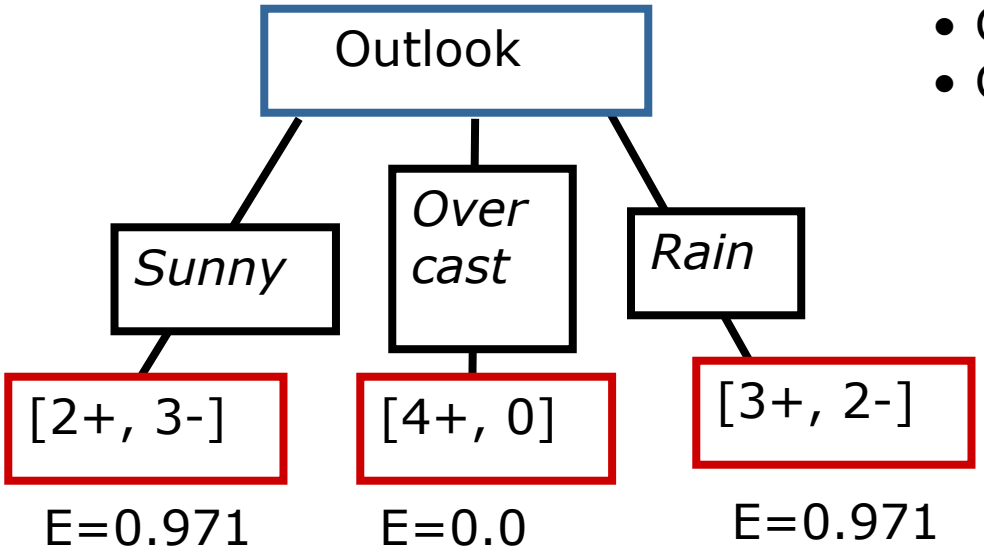
E=0.811

Gain(S,Wind)
=0.940-(8/14)*0.811
– (6/14)*1.0
=0.048

决策树

例子：

S=[9+,5-]
E=0.940



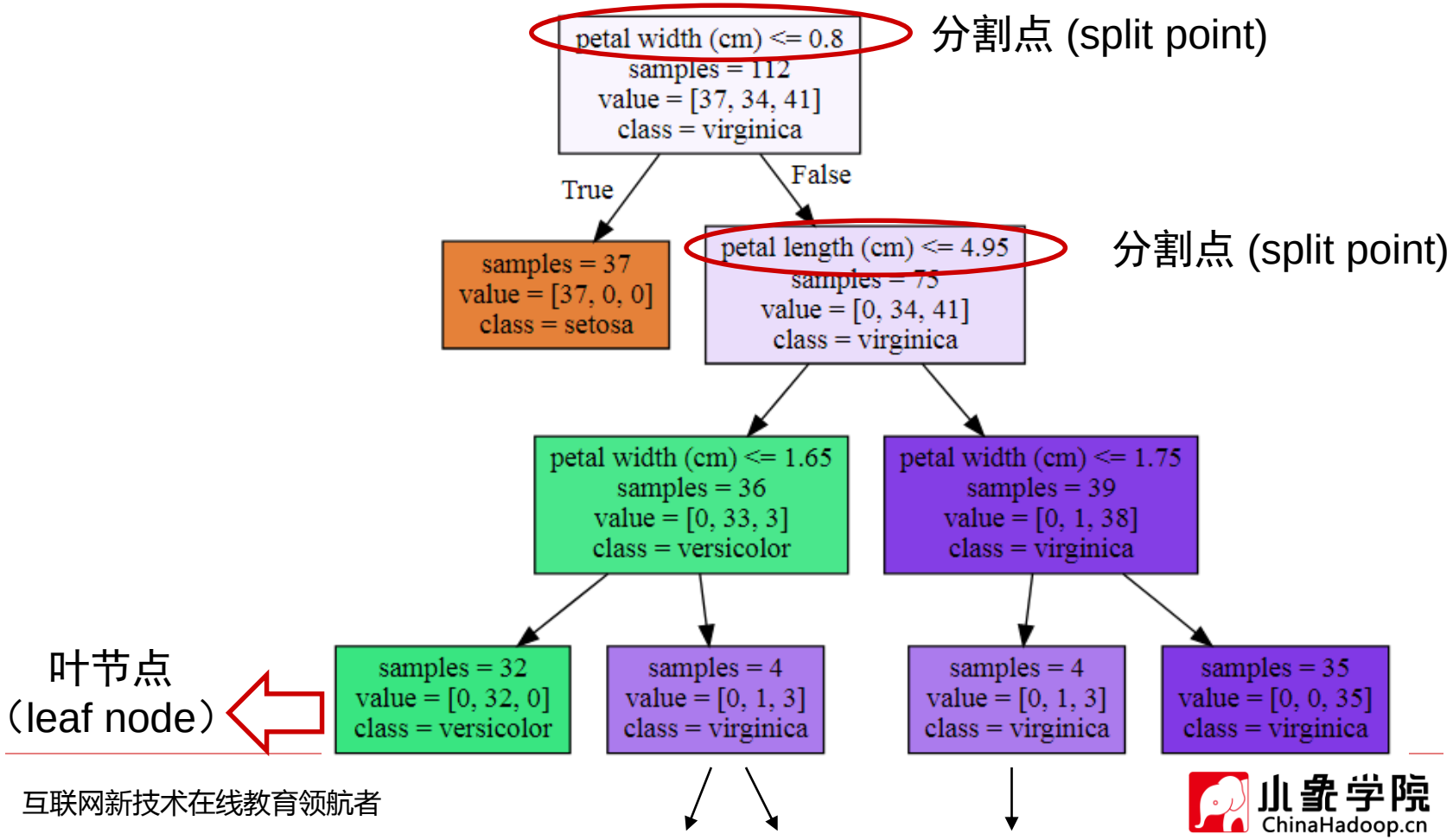
- Gain(S,Outlook) =0.247
- Gain(S,Humidity) =0.151
- Gain(S,Wind) =0.048
- Gain(S,Temperature) =0.029

特征为连续值？
- 离散化连续值
  - 二分法

Gain(S,Outlook)
=0.940-(5/14)*0.971
-(4/14)*0.0 – (5/14)*0.0971
=0.247

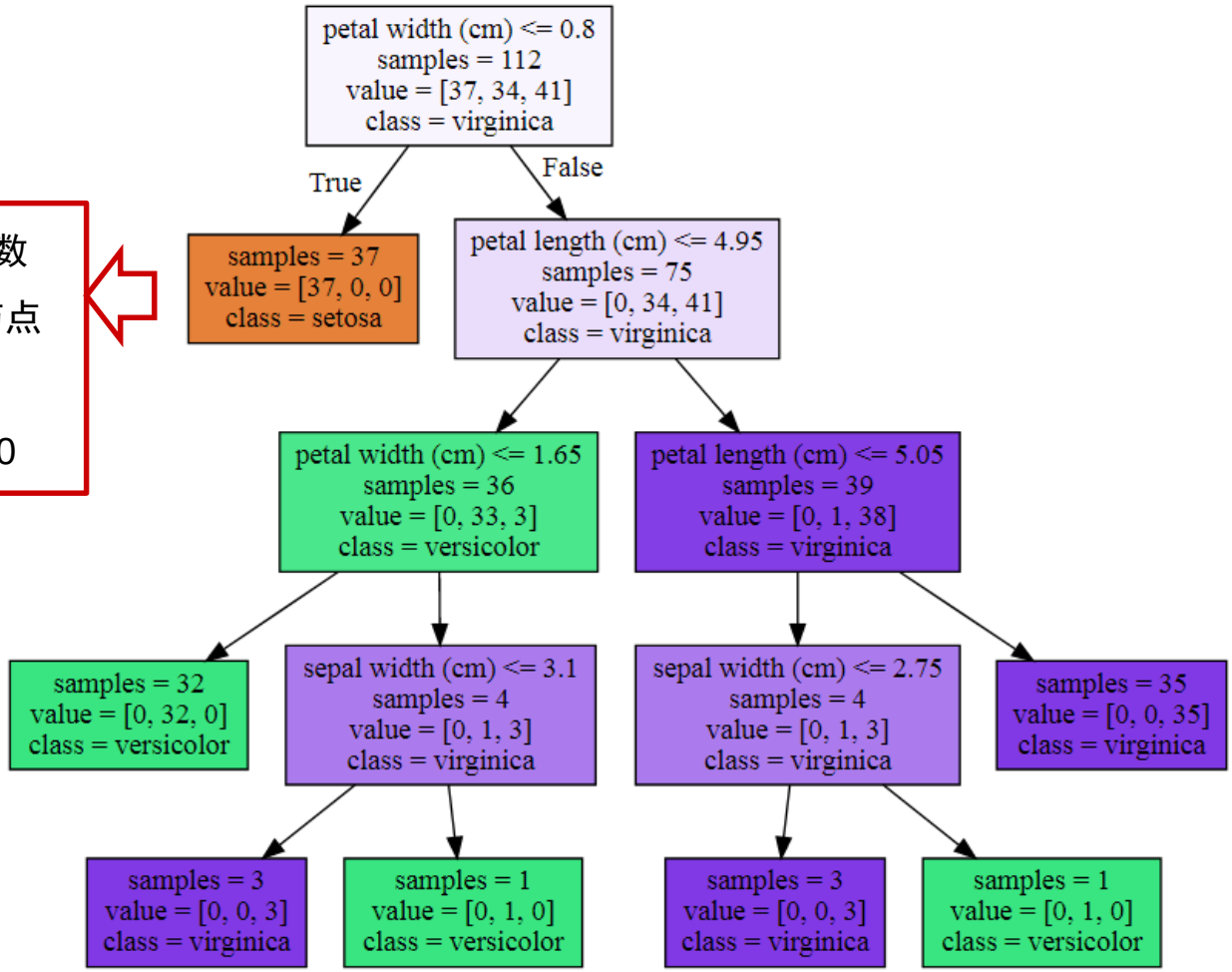
决策树

- 在iris数据集上使用决策树



决策树

- 每个叶节点包含不同分类的样本个数
- 如：values=[37, 0, 0]表示在该叶节点中，属于setosa的样本个数为37，versicolor, viginica的样本个数均为0



决策树

- 构建树的过程：

  ID3　　ID4.5　　CART

  1. 从根节点开始，计算所有特征值的信息增益（信息增益比、基尼指数），选择计算结果最大的特征作为根节点
  2. 根据算出的特征建立子节点，执行第1步，直到所有特征的信息增益（信息增益比）很小或没有特征可以选择为止

- 直接按照以上步骤构建树容易产生过拟合
- 防止过拟合：减少模型的复杂度。简化决策树->剪枝（pruning）
  - 预剪枝（pre-pruning），构造树的同时进行剪枝
  - 后剪枝（post-pruning），决策树构建完后再进行剪枝
  - 关于“剪枝” 的详细资料可参考：

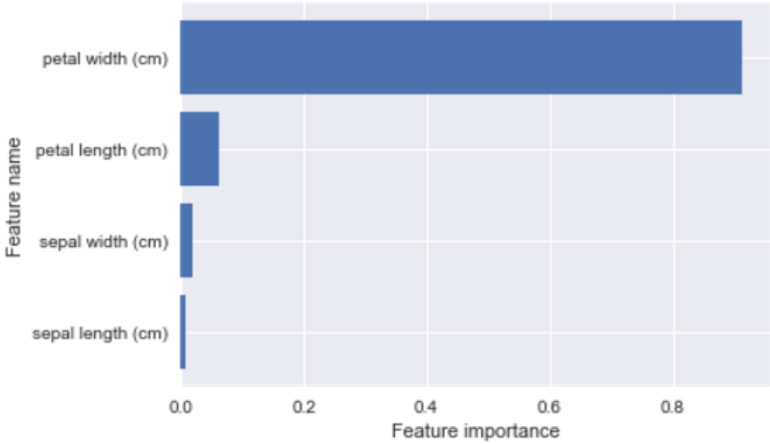
    http://www.saedsayad.com/decision_tree_overfitting.htm

sklearn.tree.DecisionTreeClassifier

决策树

- sklean中决策树重要的参数
  - max_depth：树的最大深度（分割点的个数），最常用的用于减少模型复杂度防止过拟合的参数
  - min_samples_leaf：每个叶子拥有的最少的样本个数
  - max_leaf_nodes：树中叶子的最大个数
- 实际应用中，通常只需要调整max_depth就已足够防止决策树模型的过拟合

- feature importance：
  - 得分范围：0-1
  - 得分为0：特征在预测时没有作用
  - 得分为1：单独使用该特征即可完成预测
  - 每个特征的得分总和为1



决策树

- 决策树的优缺点

优点	缺点
• 容易可视化，容易解释 • 无需对特征进行归一化处理 • 可适用于混合特征类型的数据集（连续性特征，类别型特征等）	• 即使剪枝后也很难避免过拟合 • 通常需要进行ensemble才能达到较好的效果（如：随机森林）

联系我们

小象学院：互联网新技术在线教育领航者

– 微信公众号：**小象学院**