

# 法律声明

---

□ 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师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、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



# 决策树和随机森林实践

---



小象学院  
ChinaHadoop.cn

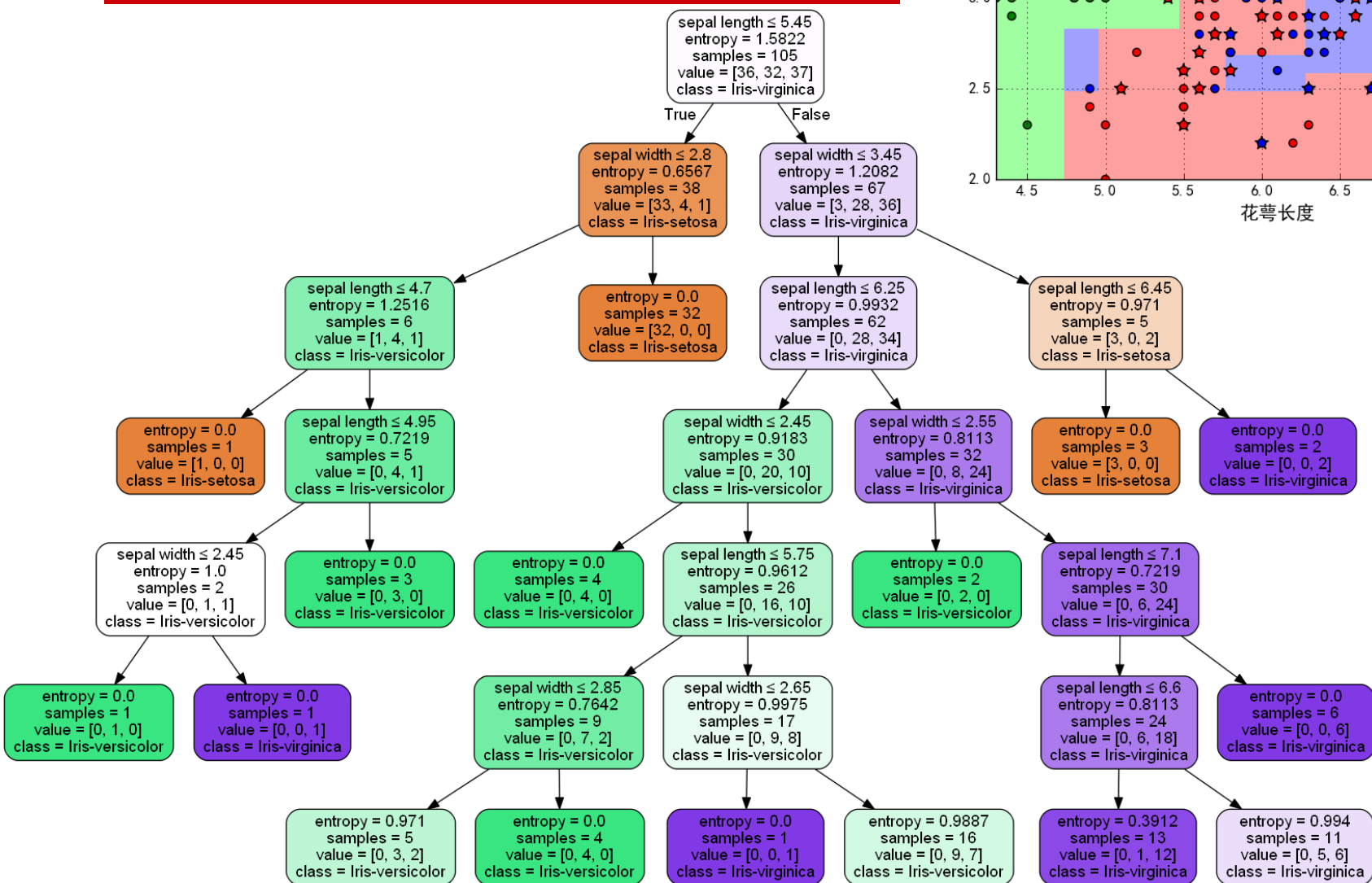
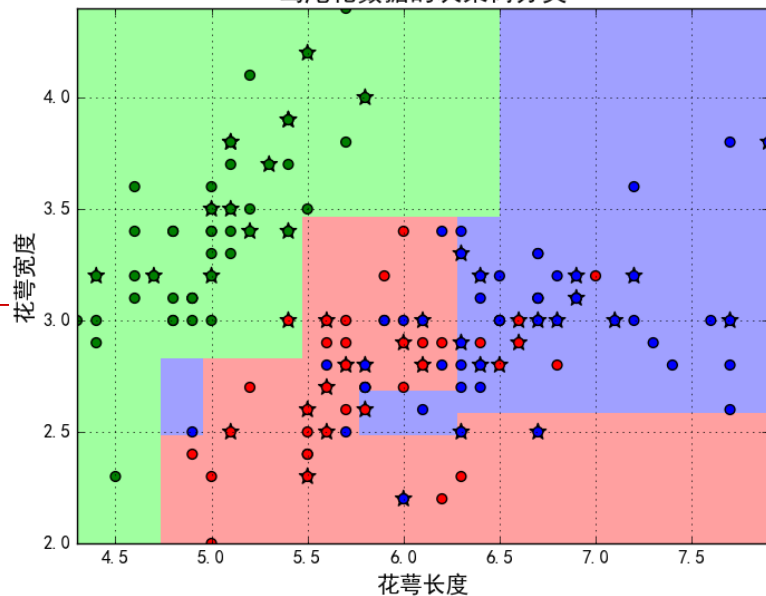
邹博

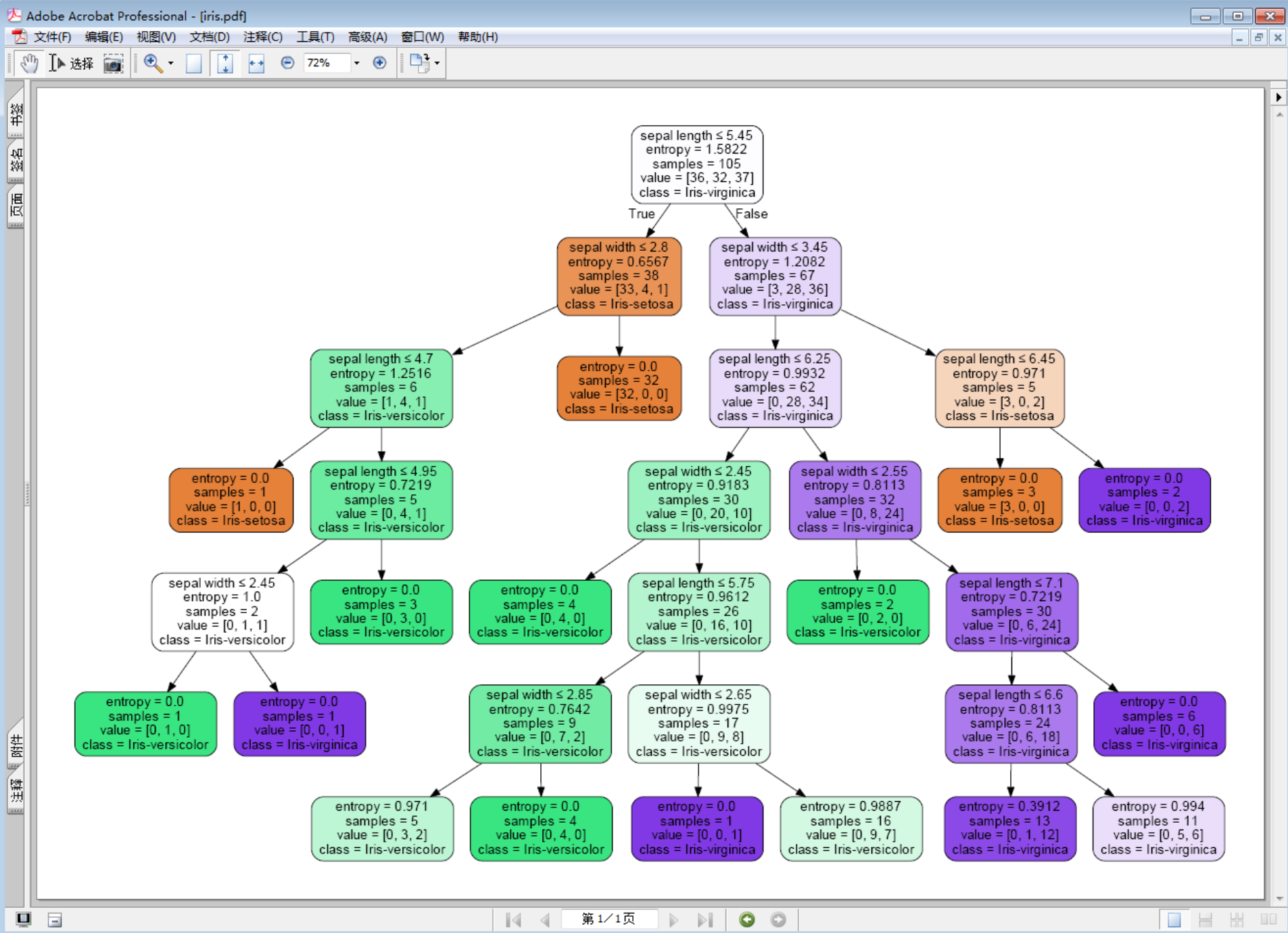
# 主要内容

---

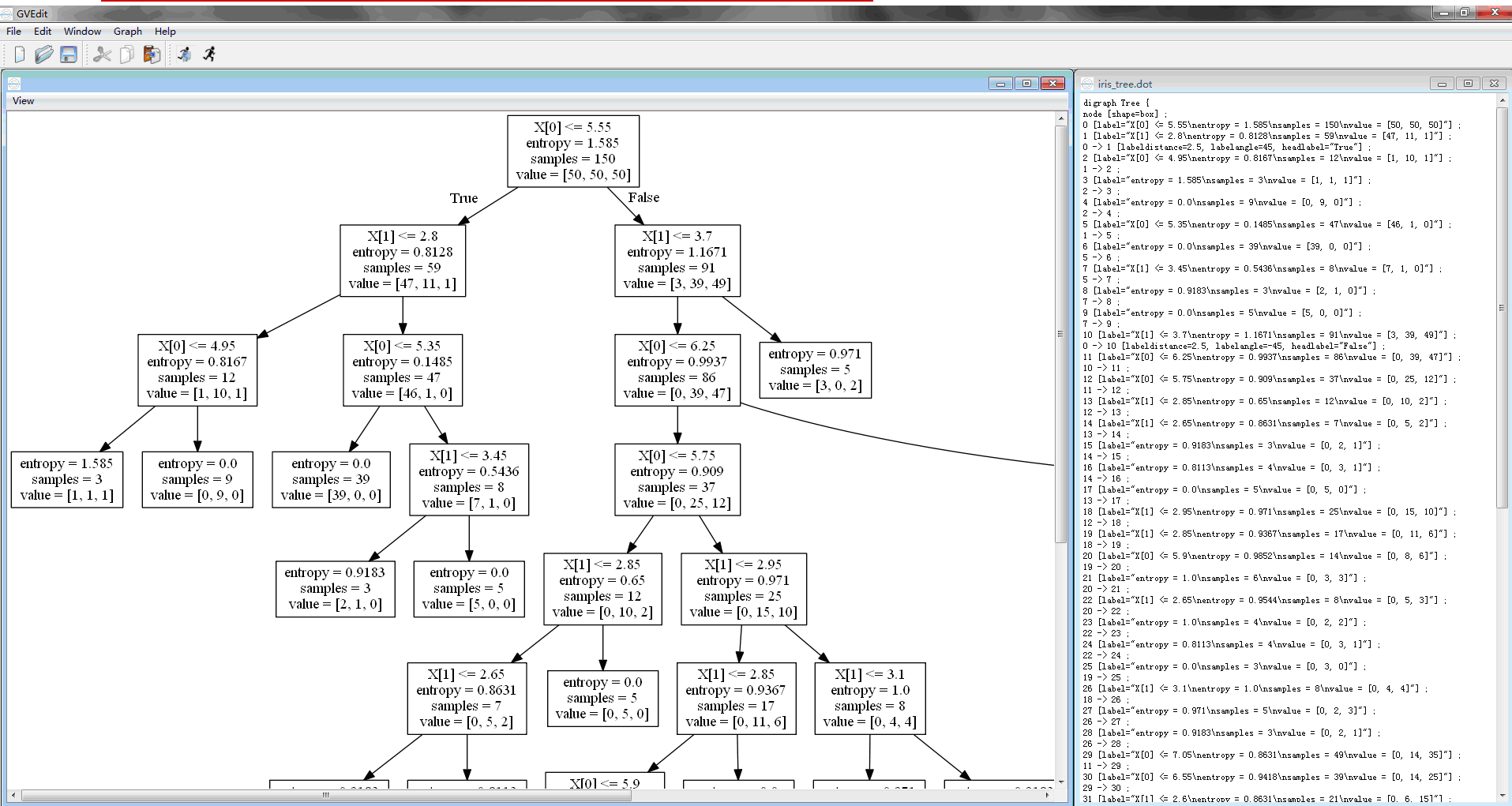
- 决策树分类鸢尾花数据
- 随机森林分类鸢尾花数据
- 决策树的可视化
- 决策树回归
- 多输出决策树
  
- 思考：树的深度与过拟合

# 鸢尾花数据决策树

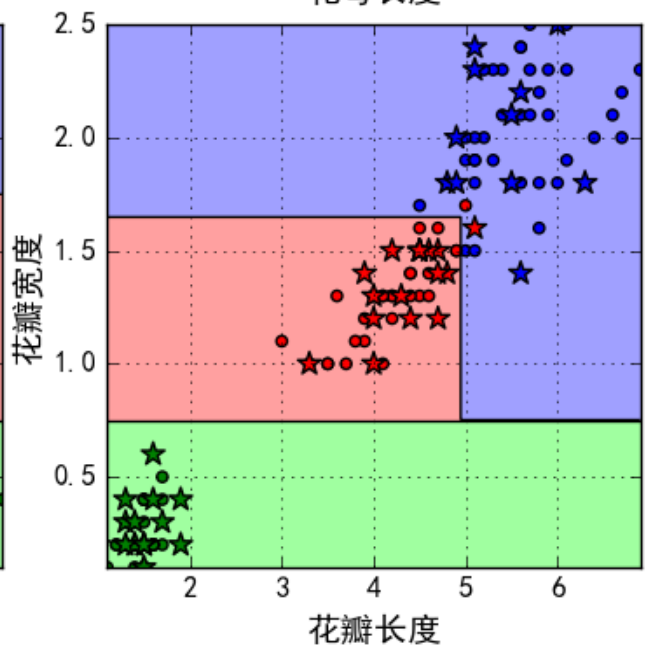
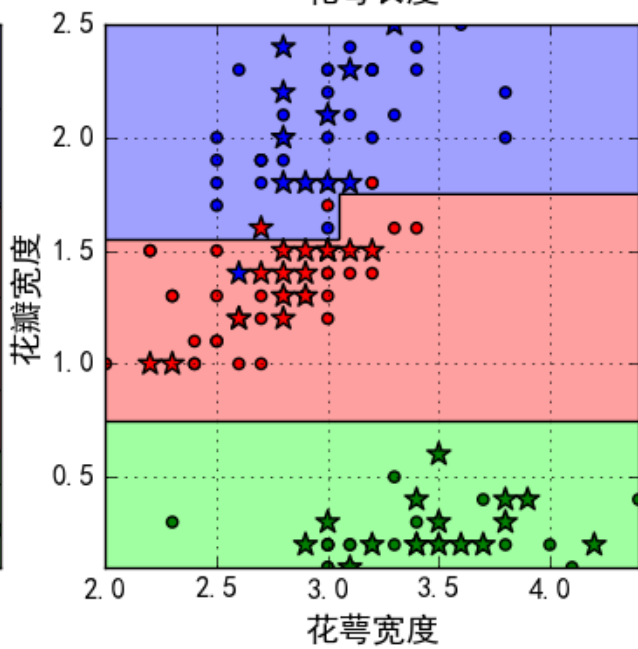
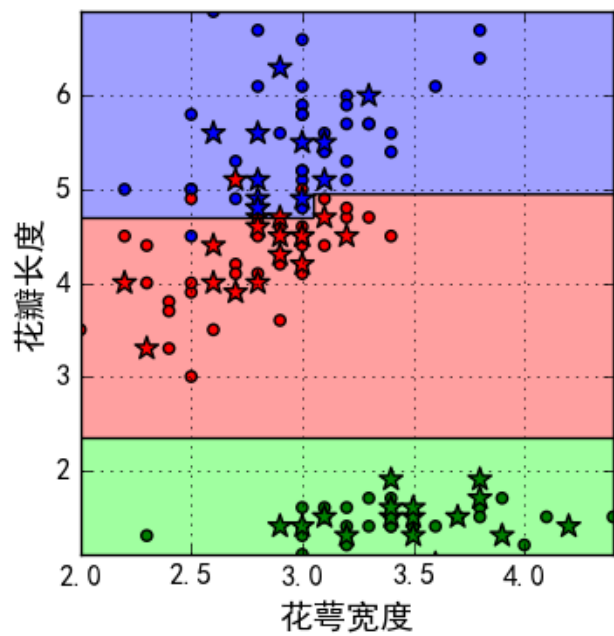
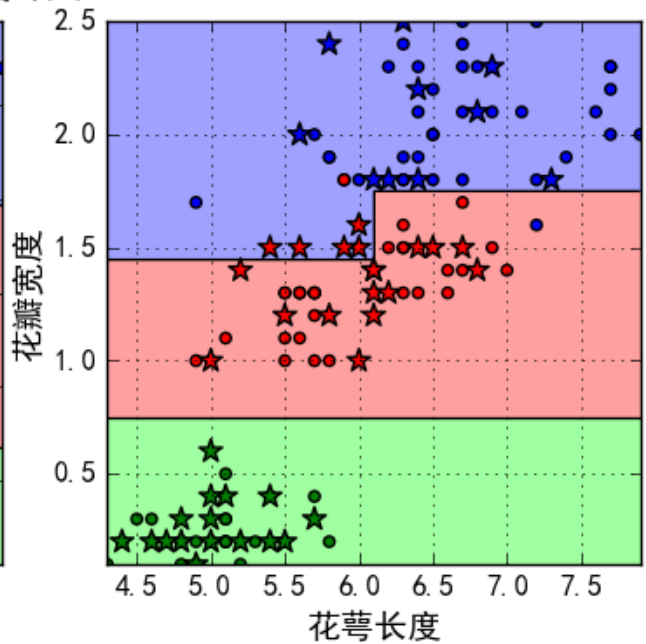
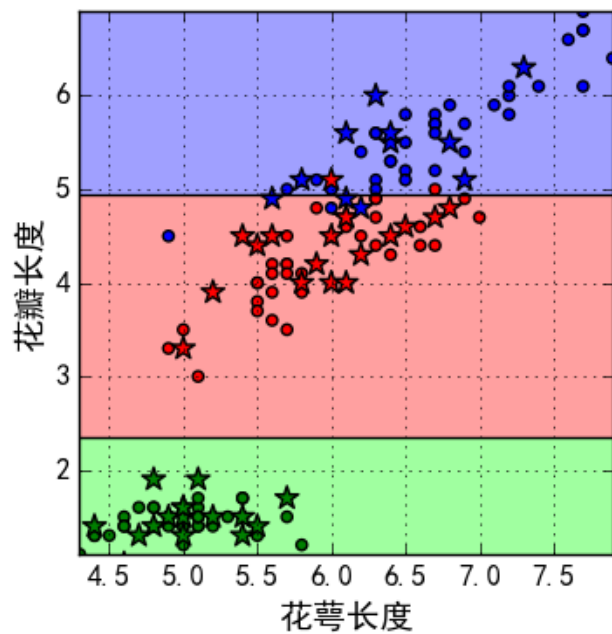
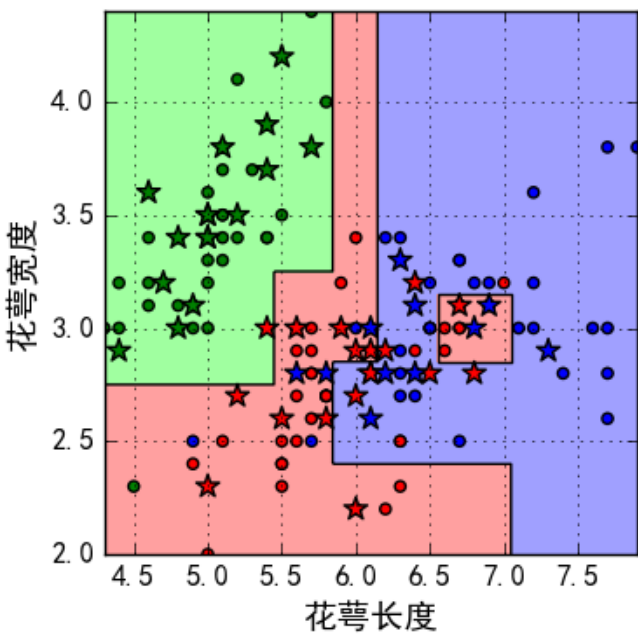




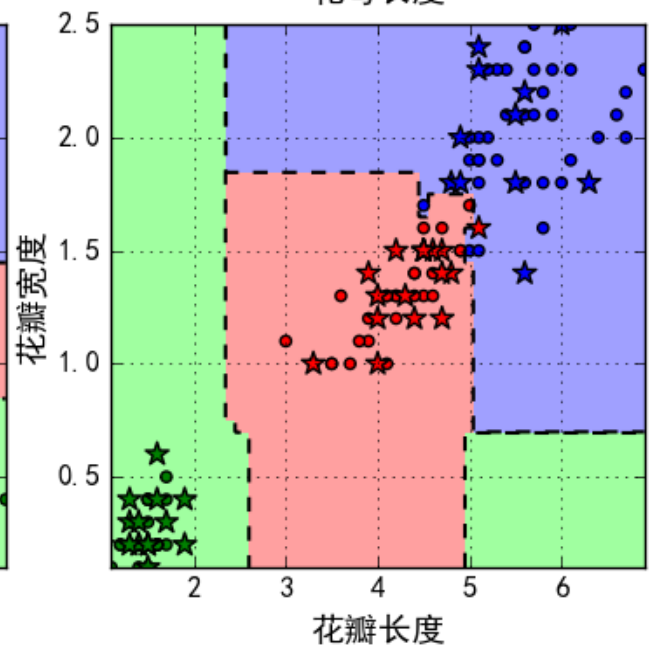
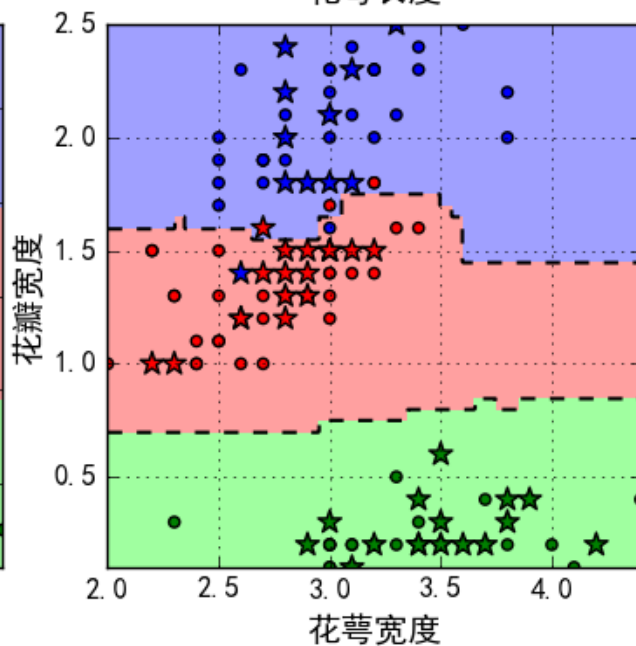
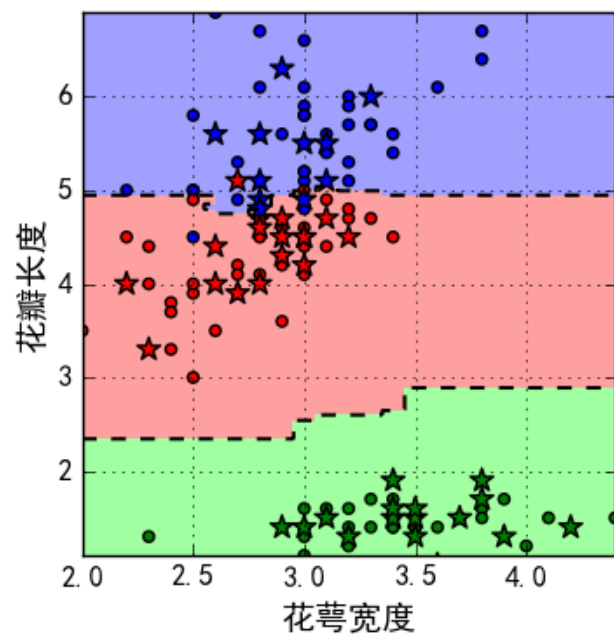
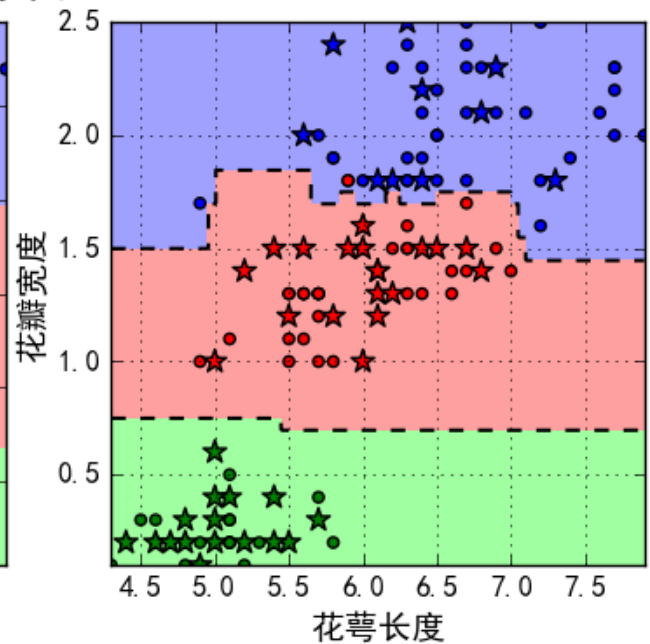
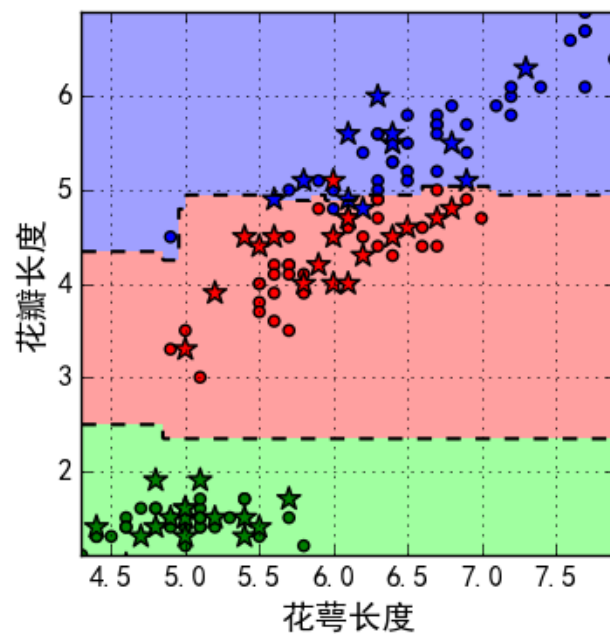
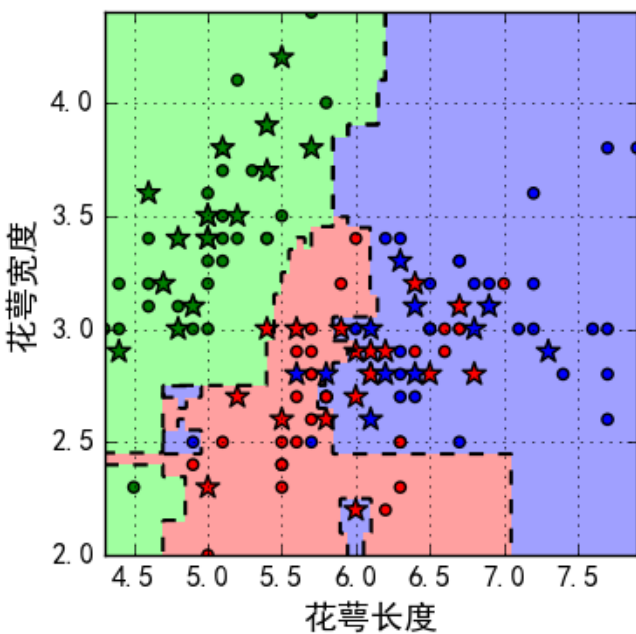
# Graphviz



决策树对鸢尾花数据两特征组合的分类结果



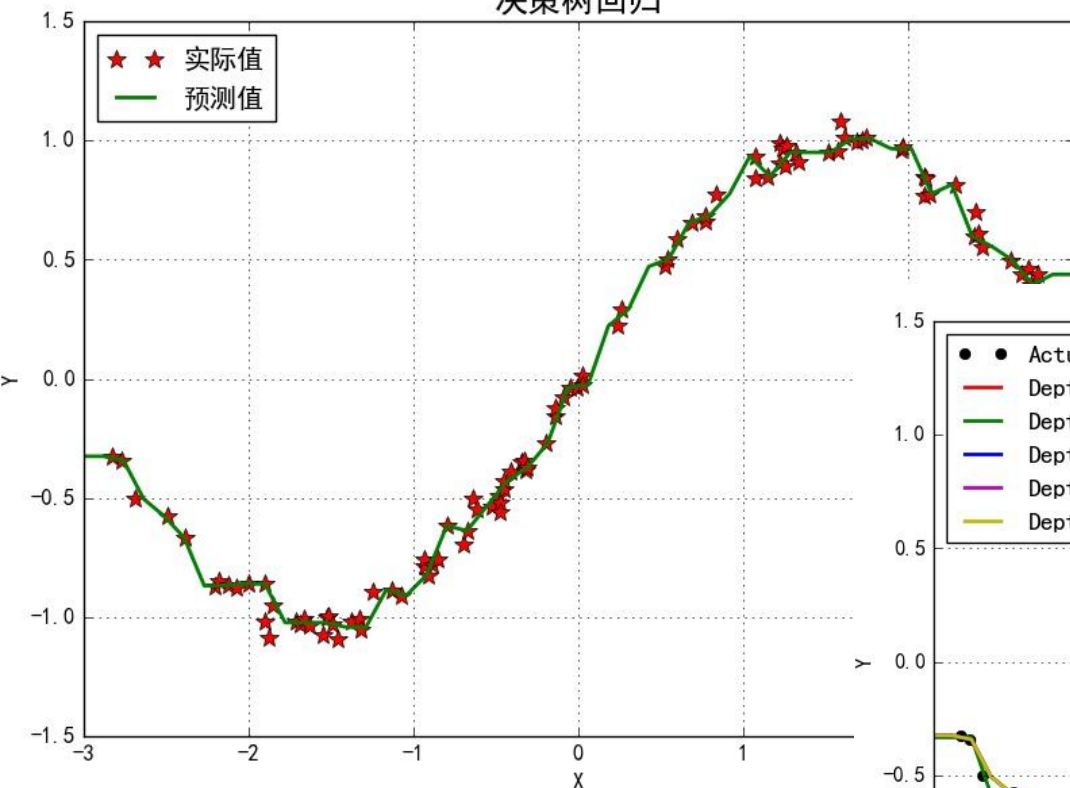
随机森林对鸢尾花数据两特征组合的分类结果



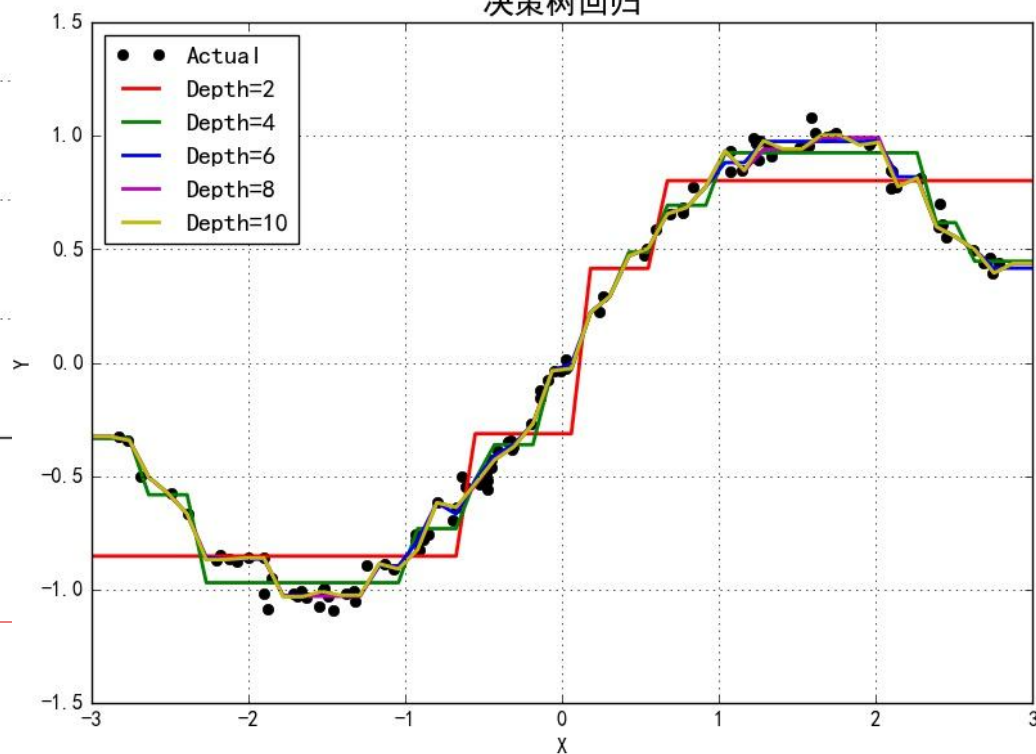


# 决策树用于拟合

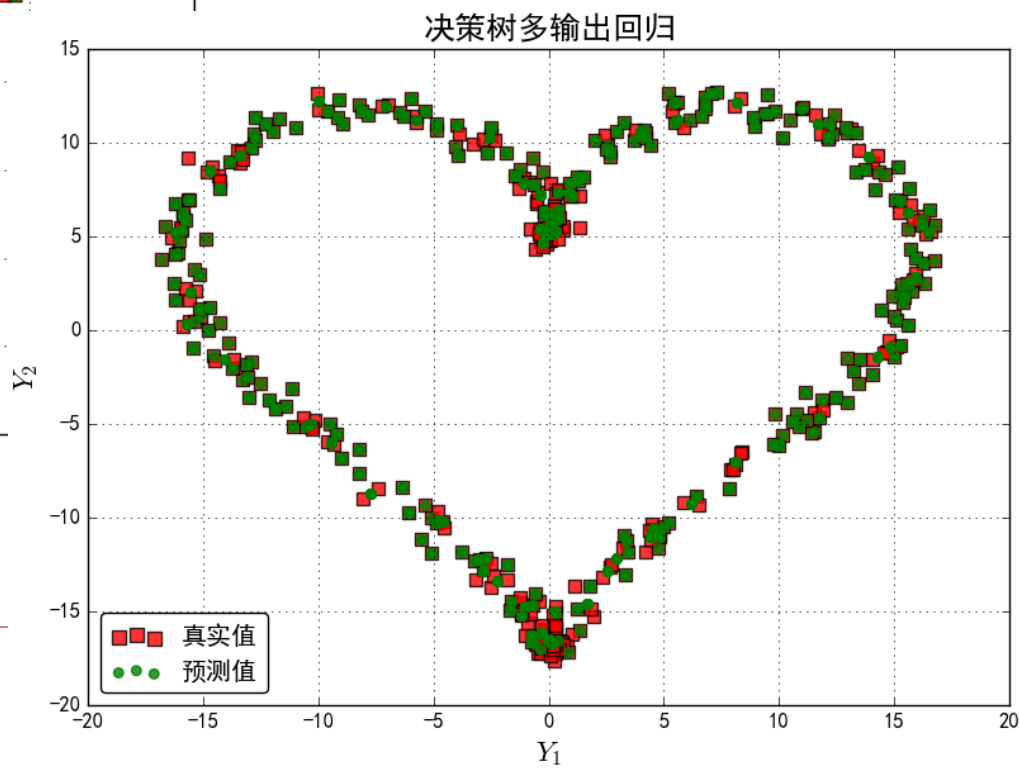
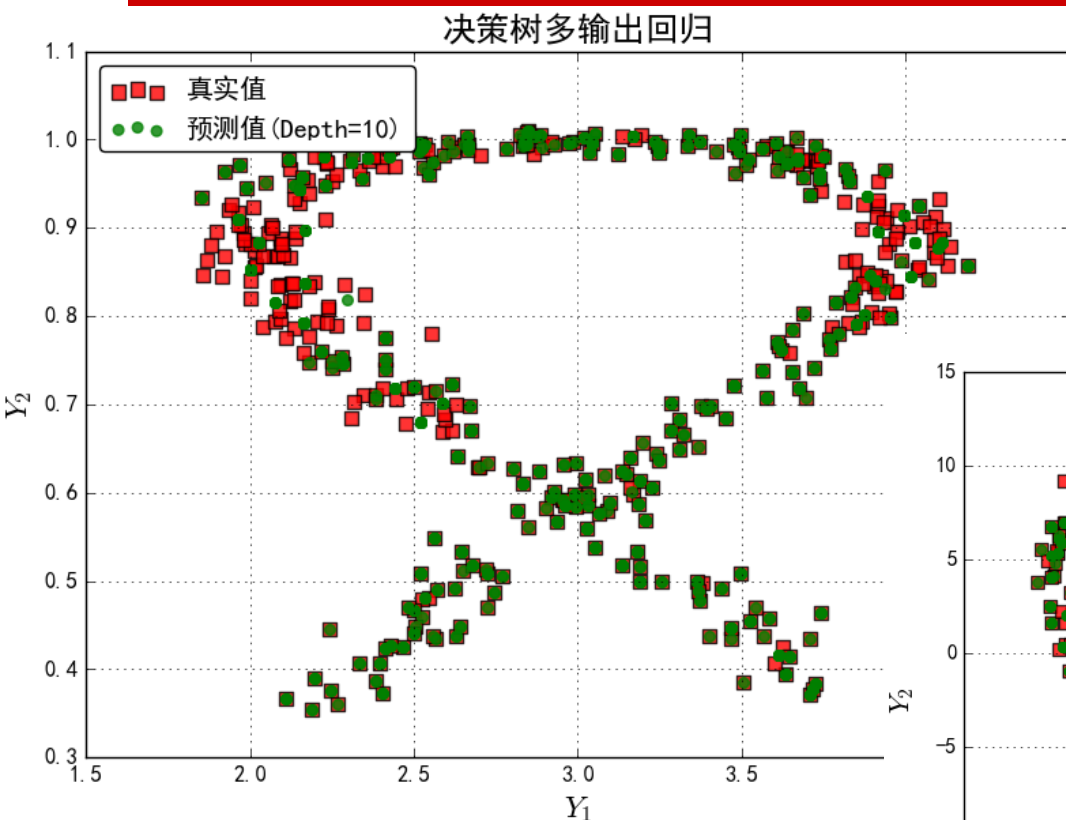
决策树回归



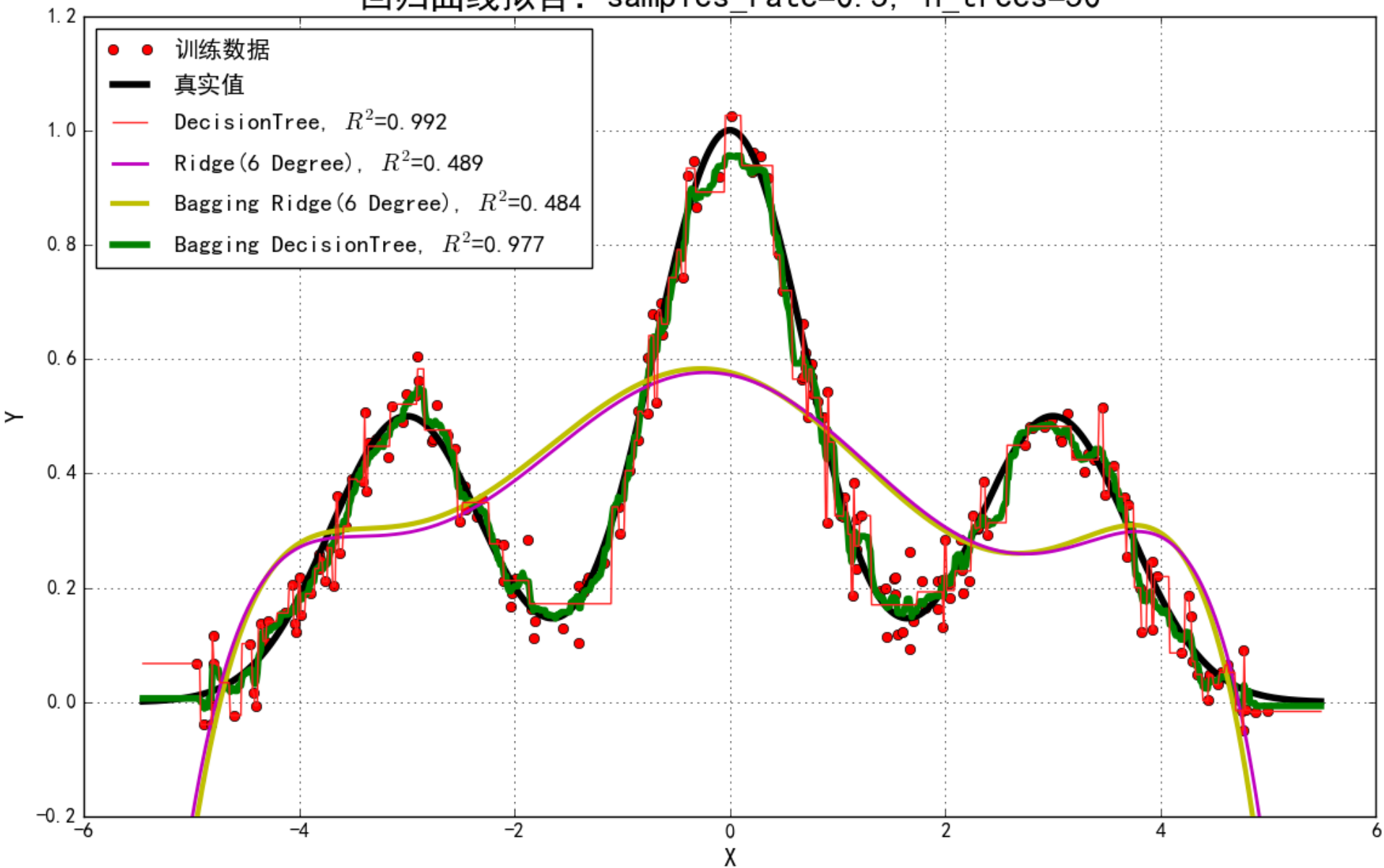
决策树回归



# 多输出的决策树回归



回归曲线拟合: samples\_rate=0.5, n\_trees=50



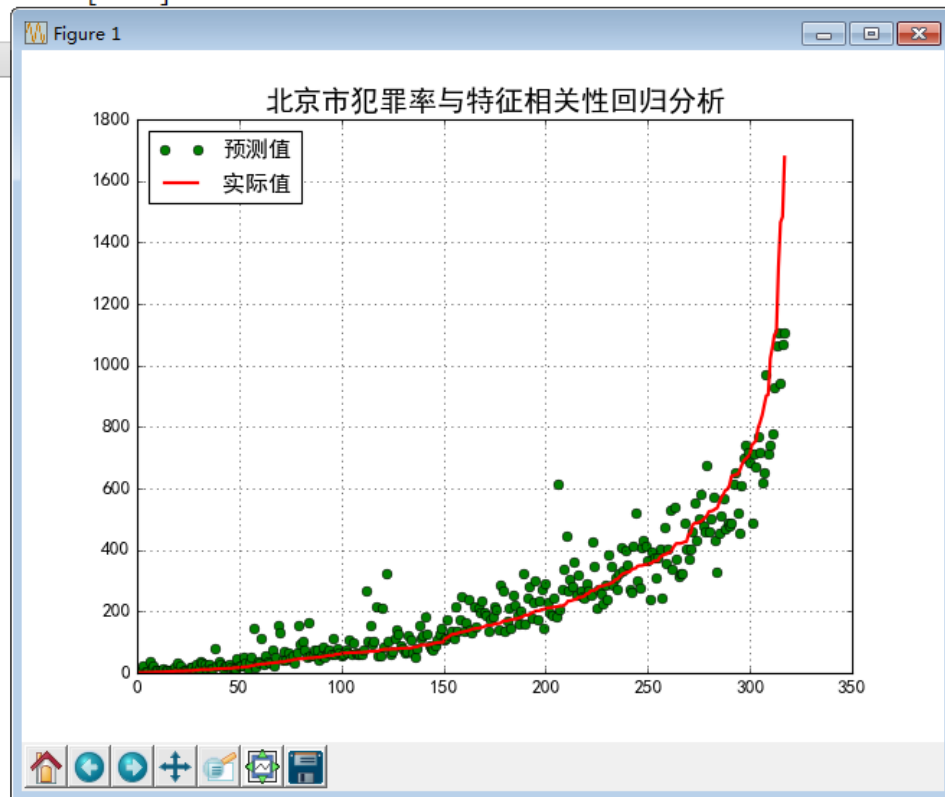
# 再谈北京市区域犯罪率分析

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	地区	盗窃案件数	批发和零售业数量	交通运输仓储邮政业数量	房地产业数量	住宿和餐饮业数量	卫生和社会工作数量	居民服务修理服务业数量	大型单位数量	中型单位数量	小微单位数量	金融业单位数量	液化石油气	能源合计	从业人员	销售费用	营业收入	营业税及附加	总产值	利润总额	人员支出
1	安定门街道办事处	87	222	1	88	22	2	88	12	12	88	8	17612	2222	2222	7822222	2222222	2222222	8222222	2222222	
2	安定镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
3	安贞街道办事处	122	88	12	12	88	12	88	12	12	88	12	1222	2222	1222	8822222	8822222	8822222	8822222	8822222	
4	奥运村街道办事处	88	222	12	12	88	12	88	12	12	88	12	1222	2222	1222	8822222	8822222	8822222	8822222	8822222	
5	八宝山街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
6	八宝山镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
7	八达岭镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
8	八角街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
9	八里庄街道办事处(朝阳)	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
10	八里庄街道办事处(海淀)	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
11	白纸坊街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
12	百泉街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
13	百善镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
14	宝山镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
15	北房镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
16	北京经济技术开发区	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
17	北七家镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
18	北石槽镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
19	北太平庄街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
20	北务镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
21	北下关街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
22	北小营镇	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
23	北新桥街道办事处	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	

# Code

```
model = RandomForestRegressor(n_estimators=100, criterion='mse', max_depth=10, min_samples_split=5,
                             max_features=0.6, oob_score=True)
model.fit(x, y)
print 'OOB Score = ', model.oob_score_
y_hat = model.predict(x)
rmse = np.sqrt(np.mean((y_hat - y)**2))
print 'RMSE = ', rmse, 'Predict Score = ', rmse / np.mean(y)
feature_importances = np.array(zip(columns, model.feature_importances_))
feature_importances[:, 1] = feature_importances[:, 1].astype(np.float)
feature_importances.sort(axis=0)
feature_importances = feature_importances[::-1]
for fi in feature_importances:
```

房地产业数量 0.0240475164437  
总产值 0.0196562211564  
居民服务修理服务业数量 0.0194228549579  
小微单位数量 0.0187625408241  
大型单位数量 0.0184318382301  
地铁线路 0.0182858907058  
地铁站 0.016660182329  
卫生和社会工作数量 0.0144664713616  
劳务费 0.0137140251708  
利润总额 0.0128121358596  
公用支出 0.0113859310541  
公交线路 0.0113687880457  
公交站 0.0112938309575  
住宿和餐饮业数量 0.0107283288012  
从业人员 0.00870104496645  
人员支出 0.00541540683171  
交通运输仓储邮政业数量 0.0053253206699  
中型单位数量 0.00385674089549



# 作业

---

- 使用决策树做任意数据集的分类。
  - 离散变量
- 使用随机森林做数据回归。
  - 连续变量

# 我们在这里

□ <http://wenda.ChinaHadoop.cn>

■ 视频/课程/社区

□ 微博

■ @ChinaHadoop

■ @邹博\_机器学习

□ 微信公众号

■ 小象学院

■ 大数据分析挖掘

互联网新技术在线教育领航者

小象问答 搜索标题、用户 全站内容搜索 提问 首页 动态 发现 话题 通知

全部 招聘求职 机器学习 大数据平台技术 DCon 大数据行业应用 NoSQL数据库 数据科学 江湖救急

发现 最新 推荐 热门 等待回复

graphviz has no attribute 'write' 贡献  
邹博 回复了问题 • 2 人关注 • 1 个回复 • 3 次浏览 • 2017-04-09 15:48

sklearn中如何理解Pipeline机制 贡献  
数据分析与数据挖掘 邹博 回复了问题 • 2 人关注 • 1 个回复 • 28 次浏览 • 2017-04-09 15:39

关于9.Logistic回归的ppt中第9页的对数线性函数 贡献  
机器学习 邹博 回复了问题 • 3 人关注 • 3 个回复 • 39 次浏览 • 2017-04-09 15:35

关于“贝叶斯估计中，最大后验概率估计就是结构化风险最小化的例子：当模型是条件概率分布，损失函数为对数损失函数，模型的复杂度由模型的先验概率表示，结构化风险最小化就等价于最大后验概率估计” 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 26 次浏览 • 2017-04-09 15:27

关于连续值的预测 贡献  
咨询 邹博 回复了问题 • 2 人关注 • 1 个回复 • 31 次浏览 • 2017-04-09 15:24

拉格朗日对偶函数为什么一定是凸函数 贡献  
数据科学 邹博 回复了问题 • 2 人关注 • 2 个回复 • 26 次浏览 • 2017-04-09 15:20

梯度下降公式中的斯堪J 是 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 29 次浏览 • 2017-04-09 15:17

深度学习适合做预测吗？ 贡献  
深度学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 27 次浏览 • 2017-04-09 15:15

关于6.4PCA\_FeatureSelection.py中plt.legend的参数疑问 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 28 次浏览 • 2017-04-09 15:04

@邹博 有哪些可以下载数据源的网站？ 贡献  
数据分析与数据挖掘 邹博 回复了问题 • 4 人关注 • 1 个回复 • 31 次浏览 • 2017-04-09 14:53

LDA主题模型 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 29 次浏览 • 2017-04-09 14:45

代码10.6bagging\_ridged老师提到了采样率设为0.2能够使峰值部分的数据被体现出来。这是为什么呢？ 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 22 次浏览 • 2017-04-09 14:26

GraphViz's executables not found 贡献  
机器学习 邹博 回复了问题 • 3 人关注 • 2 个回复 • 23 次浏览 • 2017-04-09 13:47

决策树中关于feature\_importances代码的问题 贡献  
机器学习 邹博 回复了问题 • 2 人关注 • 1 个回复 • 6 次浏览 • 2017-04-09 13:11

专题  
招聘求职  
大数据行业应用  
数据科学  
系统与编程  
云计算技术

热门话题 更多 >  
机器学习 907 个问题, 230 人关注  
spark 387 个问题, 172 人关注  
hadoop 1059 个问题, 155 人关注  
python数据分析 171 个问题, 28 人关注  
数据分析与数据挖掘 54 个问题, 111 人关注

热门用户 更多 >  
小心巴 14 个问题, 0 次赞同  
又又V 45 个问题, 22 次赞同  
铁甲无声 10 个问题, 0 次赞同  
带刀锦衣卫 13 个问题, 0 次赞同

---

感谢大家！

恳请大家批评指正！