19 随机森林

19.1 本集内容简介

- 随机森林
- 决策树的训练算法
- 包外误差
- 计算变量的重要性

19.2 集成学习简介

集成学习是一种机器学习里面的哲学思想。

比如医生会诊，叫了9个医生，让每个医生出一个诊断结果，6个医生说有病，3个医生说没有病。这样就刻意预测这个人得病。

集成学习简介
集成学习（ensemble learning）是机器学习中的一种思想，它通过多个模型的组合形成一个精度更高的模型，参与组合的模型称为弱学习器（weak learner）
在预测时使用这些弱学习器模型联合进行预测
训练时需要用训练样本集依次训练出这些弱学习器

Bagging　随机森林
Boosting　AdaBoost算法

- bagging 直接用来投票
- 随机森林
- Boosting 加权和
- AdaBoost算法

另外在训练当中，构建样本弱学习器（ weaker learner ）。一般也会用决策树，做弱学习器。

19.3 Bootstrap抽样

采样与抽样，

抽样：从样本中抽出一些样本，

- 有放回抽样
- 无放回抽样，抽一次，少一个样本，每个样本，最多只能抽一次。
- Boostrap是有放回的抽样
- 当n趋向于正无穷
- 包外样本OOB

Bootstrap抽样
所谓抽样是指从一个样本数据集中随机选取一些样本，形成新的数据集
有放回抽样，一个样本被抽中之后会放回去，在下次抽样时还有机会被抽中
无放回抽样，一个样本被抽中之后就从抽样集中去除，下次不会再参与抽样，因此一个样本最多只会被抽中一次

Bootstrap是有放回抽样
从n个样本中有放回的抽取出n个样本

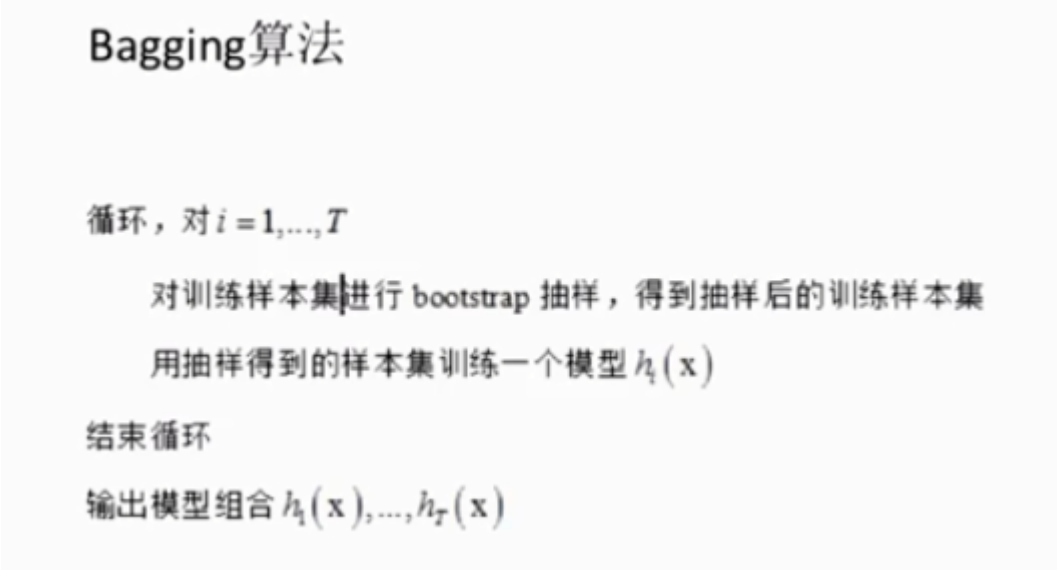
$$(1-1/n)^n$$

$$\lim_{n\to+\infty}(1-1/n)^n = 1/e$$

0.368
在整个抽样中所有样本大约有36.8%没有被抽中。这部分样本称为包外（Out Of Bag，简称OOB）数据

19.4 Bagging算法

有了有放回抽样以后，在机器学习里面，我们就有了bagging算法。算法框架如下



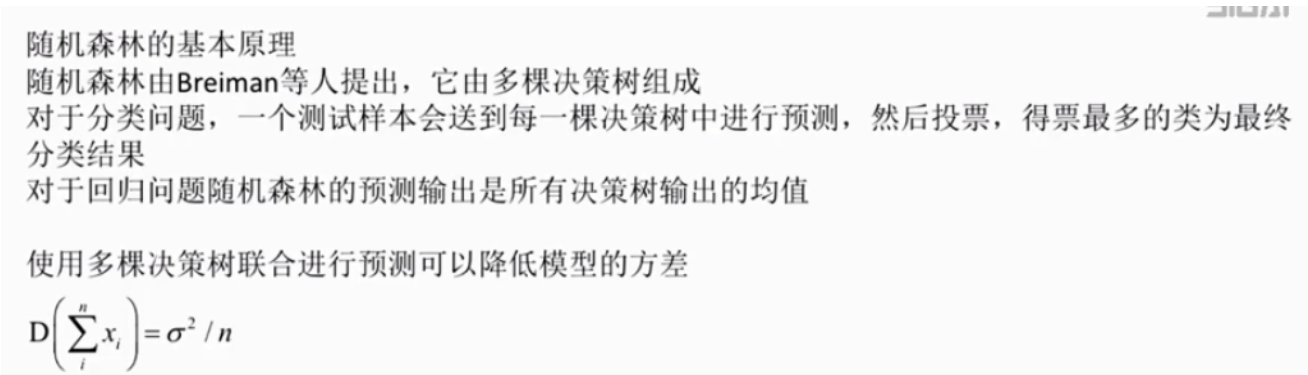
最后输出弱学习器的组合。

19.5 随机森林的基本原理

前面讲的bagging是一个抽象的框架，可以用决策树，贝叶斯分类器等算法。

前面没有指明弱学习器是什么类型的。

假设我们采用决策树的话，这种算法就称为随机森林。除了可以用分类问题，还可以用回归问题。对于分类问题，随机森林有效的减小方差。对于回归问题，我们最终的回归树，弱学习器越多，他的方差越小。



在概率论和数理统计里面，有图中公式的结论。

19.6 训练算法

训练算法
训练时依次训练每一棵决策树，每棵树的训练样本都是从原始训练集中进行随机抽样得到
在训练决策树的每个节点时所用的特征也是随机抽样得到的，即从特征向量中随机抽出部分特征参与训练
随机森林对训练样本和特征向量的分量都进行了随机采样

样本的随机抽样可以用均匀分布的随机数构造
对特征分量的采样是无放回抽样，可以用随机洗牌算法实现

决策树的数量
使用的特征的数量

在训练决策树的时候，我们要寻找最佳分裂，决策树做了些处理：

对行进行有放回的抽样，对列不放回抽样 。可以用随机洗牌算法。STL算法。

样本的随机抽样可以用均匀分布的随机数构造
对特征分量的采样是无放回抽样，可以用随机洗牌算法实现

决策树的数量多少合适？深度多少合适？这个没有固定的规则，需要人工设定

随机洗牌算法？

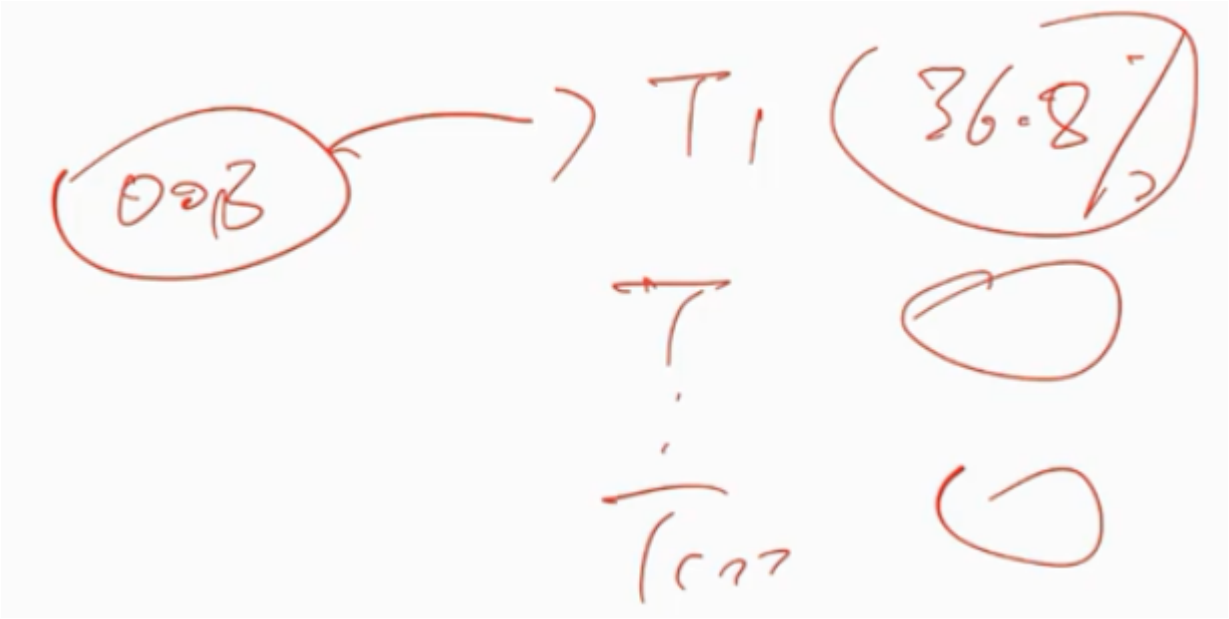
19.7 包外误差

包外误差

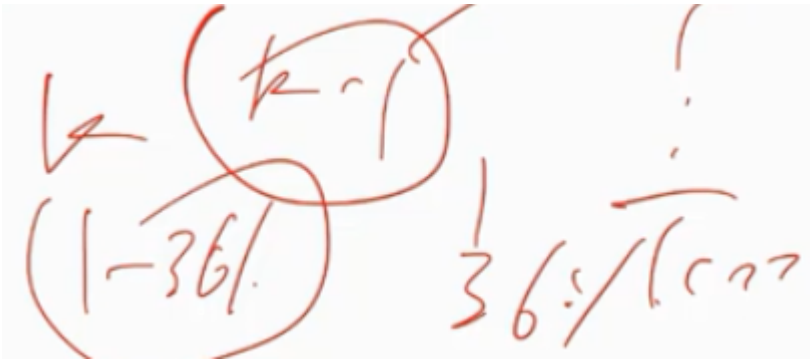
包外误差
训练每一棵决策树时约有部分样本未参与训练。可以在训练时利用这些没有被选中的样本做测
统计它们的预测误差，这称为包外误差
二者都是把样本集切分成多份，轮流用其中的一部分样本进行训练，用剩下的样本进行测试
不同的是交叉验证把样本均匀的切分成份，在训练集中同一个样本不会出现多次；后者在每次bootstrap抽样时同一个样本可能会被选中多次

对于分类问题，包外误差定义为被错分的包外样本数与总包外样本数的比值
对于回归问题，所有包外样本的回归误差和除以包外样本数

前面我们介绍了OOB的概念，



有点像交叉验证，CV，等分成k分，拿k-1份做训练，1做测试。



包外误差和交叉验证

二者都是把样本集切分成多份，轮流用其中的一部分样本进行训练，用剩下的样本进行测试
不同的是交叉验证把样本均匀的切分成份，在训练集中同一个样本不会出现多次；后者在每次bootstrap抽样时同一个样本可能会被选中多次

19.8 计算变量的重要性

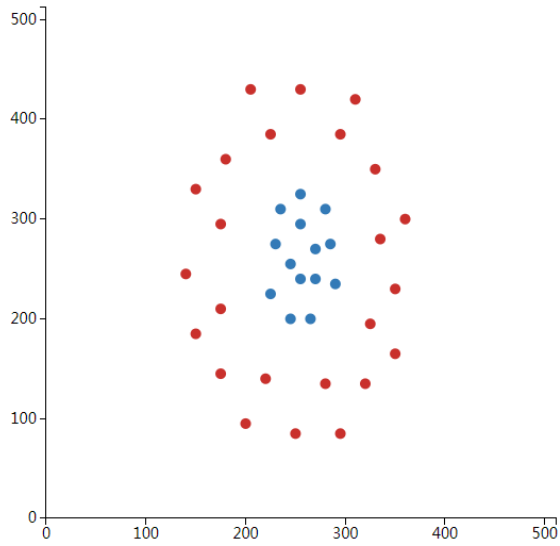
计算变量的重要性
置换法
如果某个特征很重要，那么改变样本的该特征值，该样本的预测结果就容易出现错误
如果一个特征对分类不重要，随便改变它对分类结果没多大影响

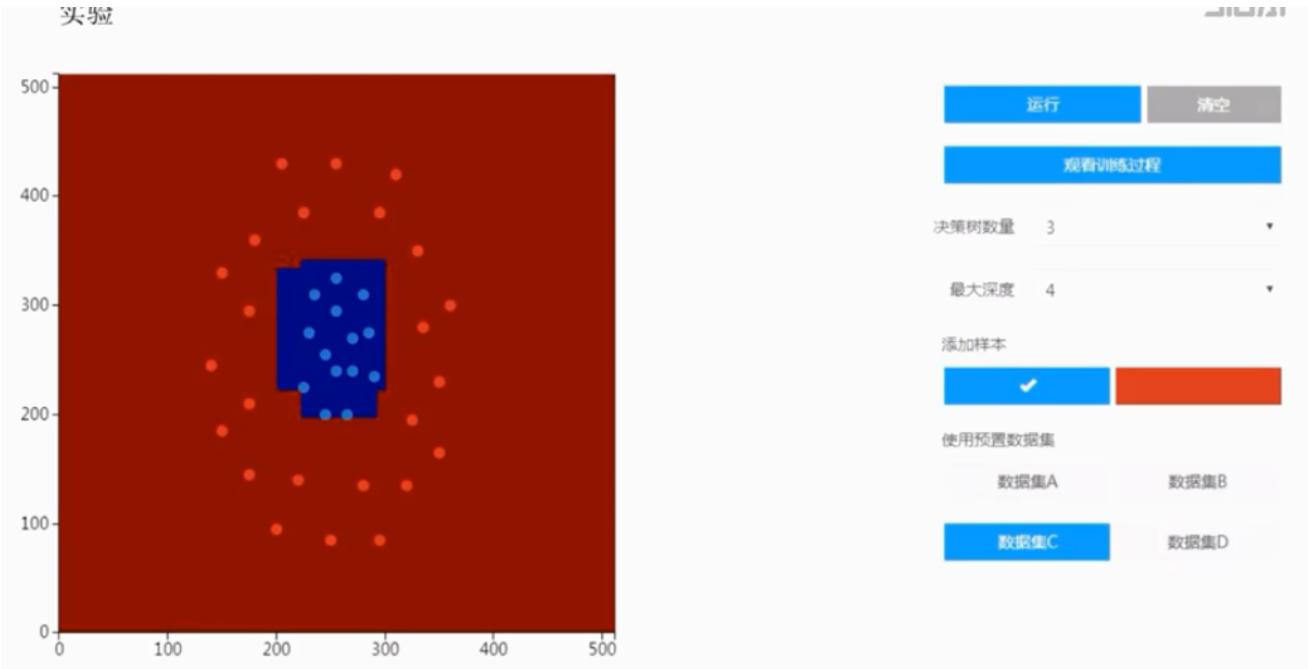
$$v = \frac{n_{y=y^*} - n_{y=y_\pi^*}}{|oob|}$$

上面定义的是单棵决策树的变量重要性，计算出每棵树的变量重要性之后，对该值取平均就得到随机森林的变量重要性

19.9 实验

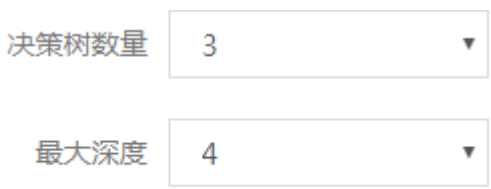
数据集C



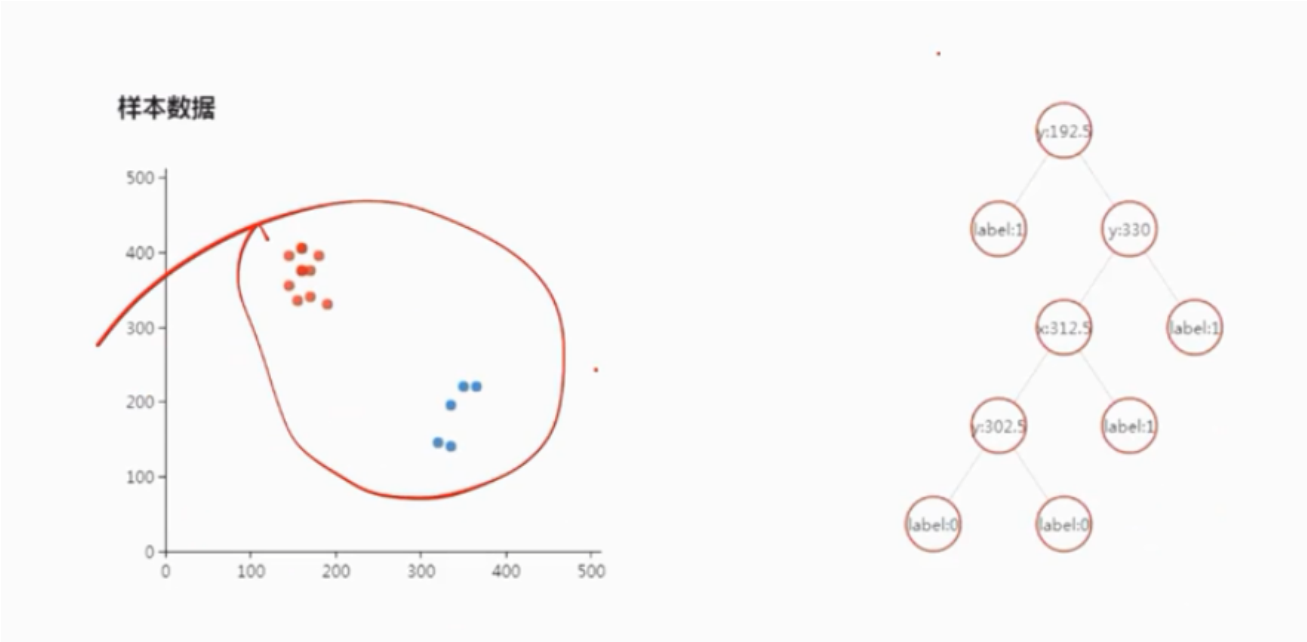


甚至两个参数

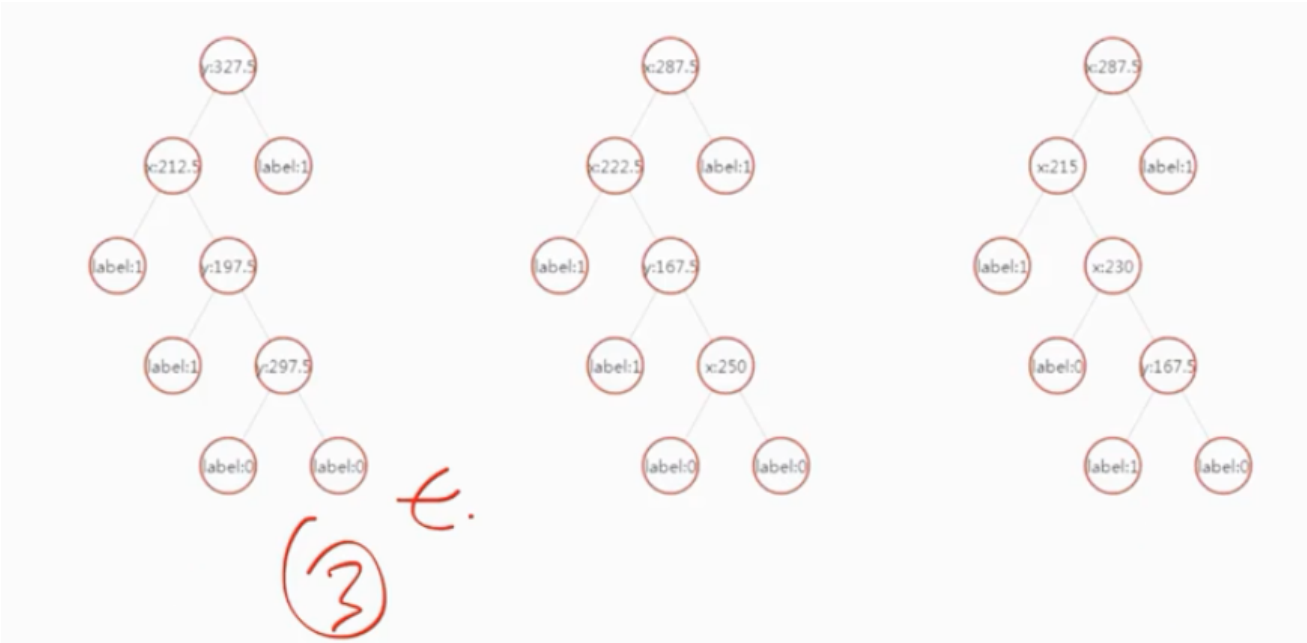
- 决策树的数量
- 最大深度



设置深度



设置决策树数量为3



19.10 实际应用

实际应用

[1] Jisoo Ham, Yangchi Chen, Melba M Crawford, Joydeep Ghosh. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing. 2005.
[2] M Pal. Random forest classifier for remote sensing classification.International Journal of Remote Sensing. 2005.
[3] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, Jian Sun. Joint Cascade Face Detection and Alignment. european conference on computer vision. 2014.

随机森林虽然看起来很简单，但是性能是得到明显的提升的。看看应用。

应用领域，

1. 人脸识别
2. JDA
3. 很多互联网公司拿来做预测，核心的还是特征工程，把特征值做得好。

19.11 本集总结

1. 集成
2. Boostrap
3. 随机森林
4. OOB
5. 包外误差
6. 计算变量的重要性