法律声明

- 课程详情请咨询
  - 微信公众号：北风教育
  - 官方网址：*http://www.ibeifeng.com/*

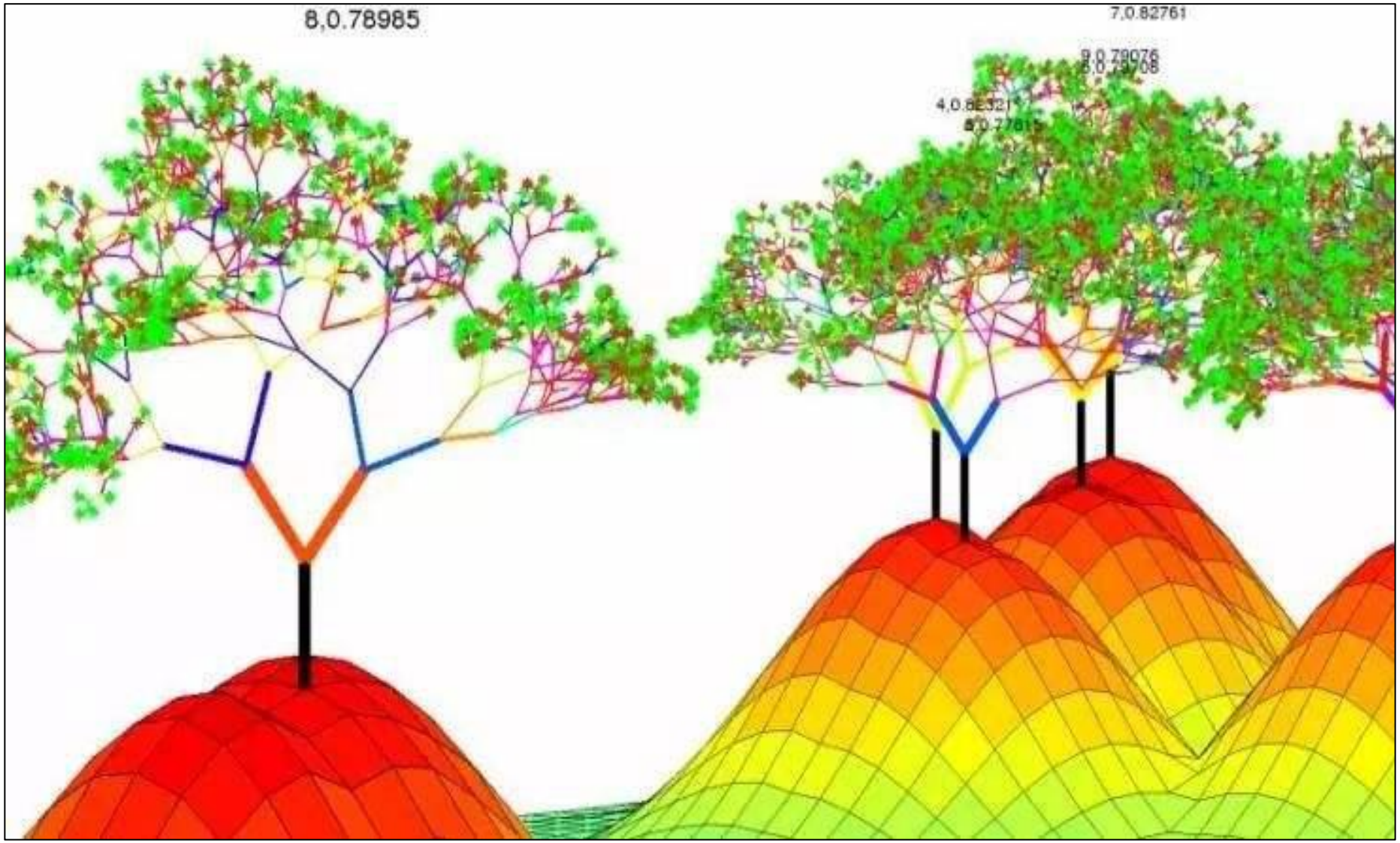
人工智能之机器学习

随机森林（Random Forest）

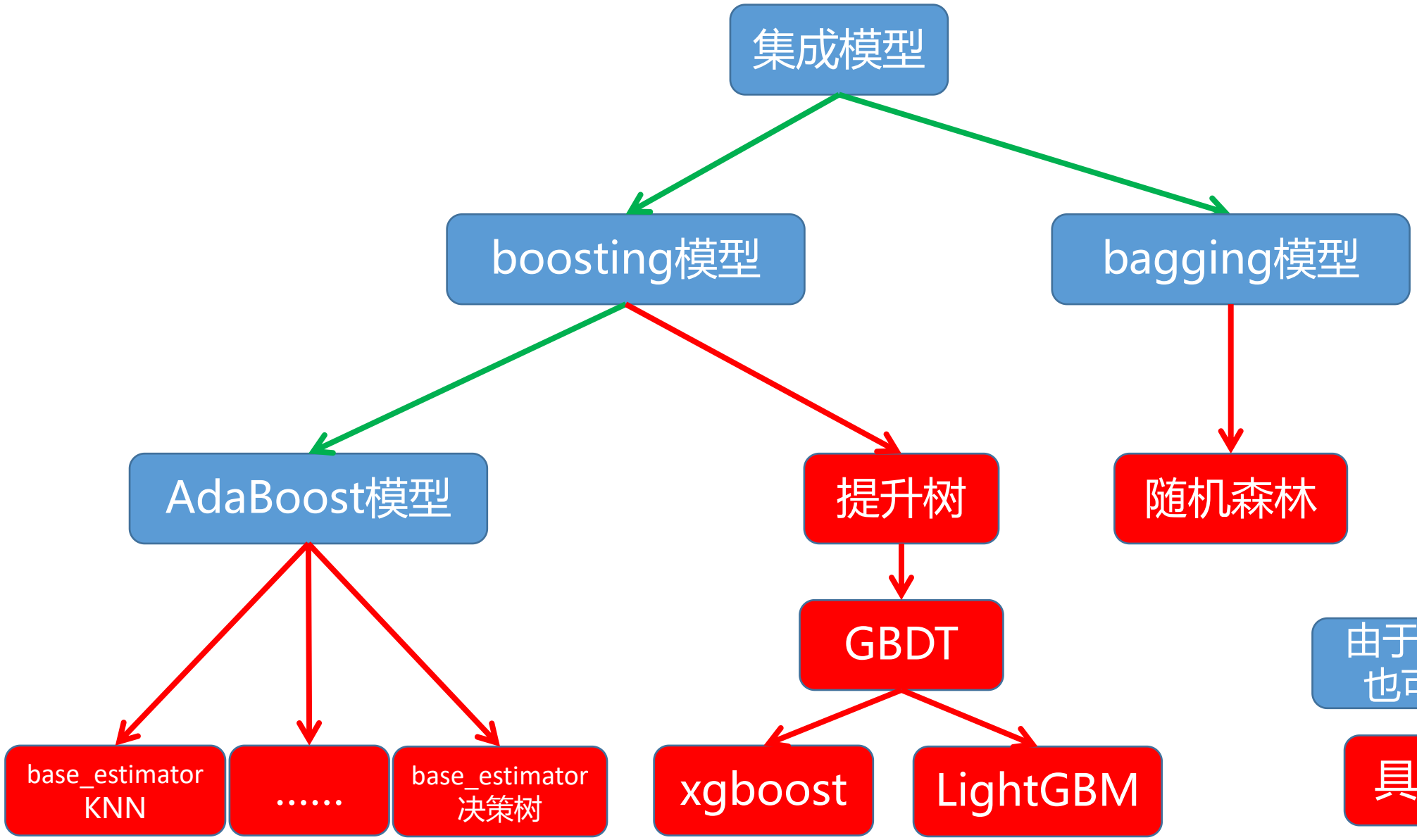
主讲人：赵翌臣

上海育创网络科技有限公司

走进森林，参天大树一棵棵相继出现



集成模型一览

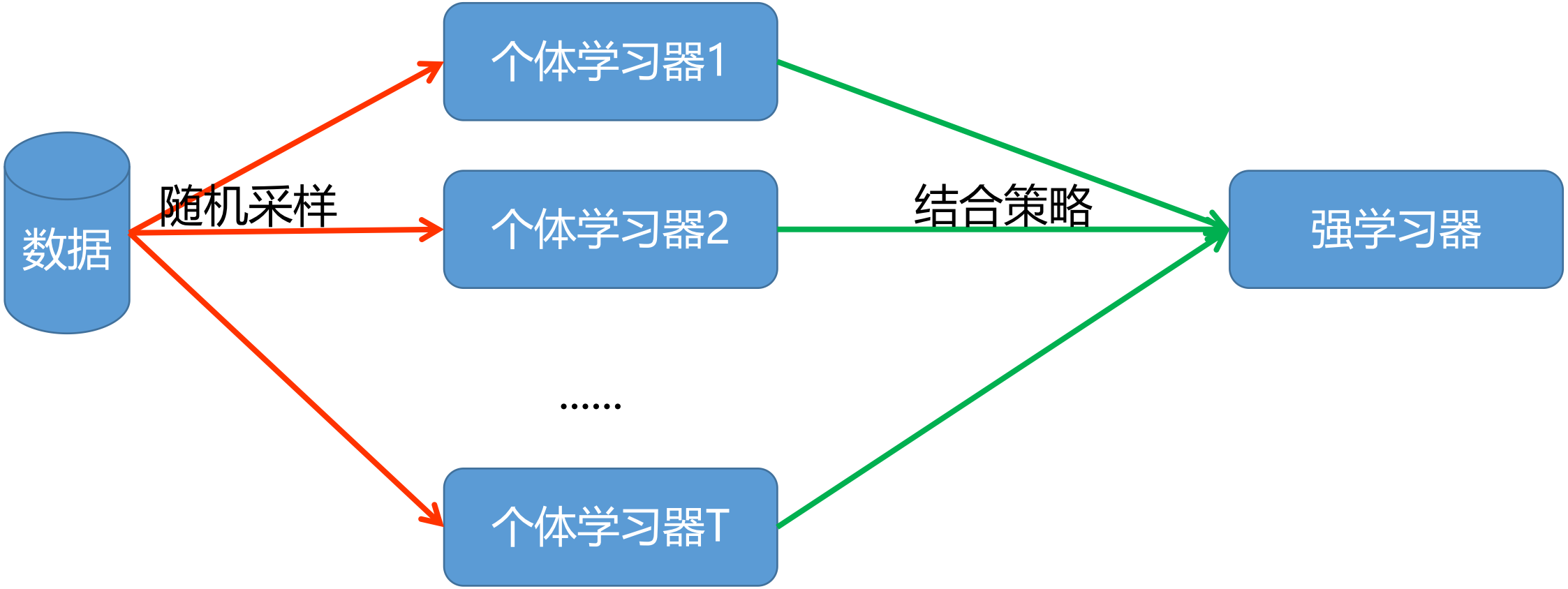


由于是抽象的，
也可以叫思想

具体实现

集成学习——Bagging思想

Bagging是Bootstrap Aggregating的缩写，通过并行地构造多个个体分类器，然后以一定的方式将它们组合成一个强学习器

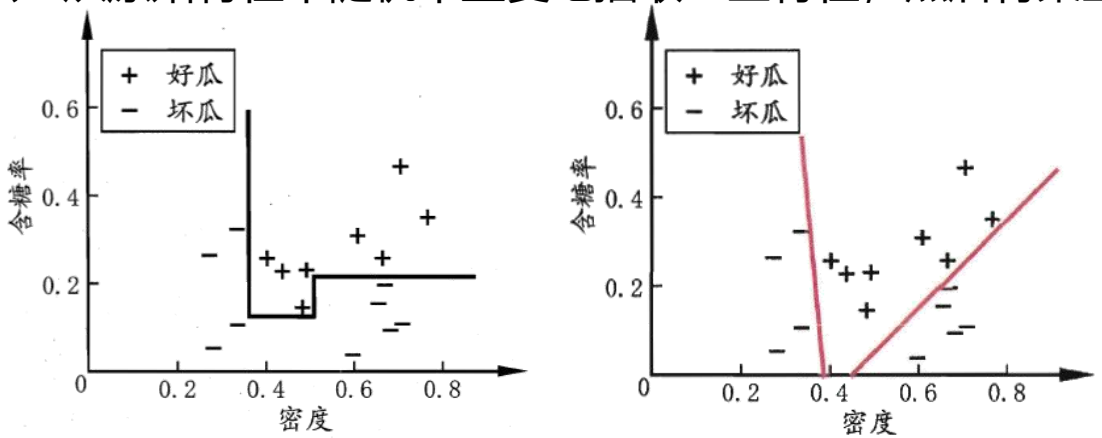


随机森林（Random Forest）

- 介绍
  - RF是基于决策树的集成模型，随机森林是机器学习中最成功的算法之一，他能做二分类、多分类和回归任务。随机森林里集成了很多棵决策树，目的是减小过拟合的风险（减小模型方差）。
- 优点
  - 像决策树一样，RF可以处理类别特征与连续特征，能扩展到多类分类，不需要特征缩放，能捕获非线性关系和特征间的影响
  - 算法可以并行
- 森林
  - 树的集合

随机森林（Random Forest）

- 思考
  - 随机的目的是什么?
- 注入随机性
  - 样本随机：训练每一个决策树使用的都是bootstrapping产生的数据集
  - 特征随机：在每一个树结点上进行结点划分时，考虑特征子空间
    - 简单做法：从原始特征中随机不重复地抽取一些特征；
    - 延伸做法：从原始特征中随机不重复地抽取一些特征，然后将某些特征线性合并，产生一系列组合特征。



随机森林的训练伪代码

function RandomForest(D,T)

for t=1,2,...,T　//可以并行执行

①request size-**N'** data D_t from bootstrapping

②obtain base g_t by DTree(**D'_t**)

return G=Uniform(g_t)

随机森林的结合策略

- 分类
  - 投票，少数服从多数。每个树的预测结果就是给某个类别投一票，最终随机森林的输出值就是得票最多的类别
- 回归
  - 平均法，每一个树都会输出一个实数，随机森林的输出值就是所有决策树输出值的均值

编程——基于Bagging的回归

例 8.2 已知如表 8.2 所示的训练数据，x 的取值范围为区间[0.5,10.5]，y 的取值范围为区间[5.0,10.0]，学习这个回归问题的提升树模型，考虑只用树桩作为基函数.

表 8.2 训练数据表

x_i	1	2	3	4	5	6	7	8	9	10
y_i	5.56	5.70	5.91	6.40	6.80	7.05	8.90	8.70	9.00	9.05

并行的训练多颗回归树，对有个样本进行预测时，所有回归树同时预测，取均值作为输出

编程——基于Bagging的分类

例 8.1 给定如表 8.1 所示训练数据. 假设弱分类器由 $x<v$ 或 $x>v$ 产生，其阈值 v 使该分类器在训练数据集上分类误差率最低. 试用 AdaBoost 算法学习一个强分类器.

表 8.1 训练数据表

序号	1	2	3	4	5	6	7	8	9	10
x	0	1	2	3	4	5	6	7	8	9
y	1	1	1	-1	-1	-1	1	1	1	-1

编程——RF综合案例之森林植被类型预测

- 数据集：
  - https://archive.ics.uci.edu/ml/datasets/covertype
- 解释：
  - 该数据集记录了美国科罗拉多州不同地块的森林植被类型。每个样本包含了描述每块土地的若干特征，包括海拔、坡度、到水源的距离、遮阳情况和土壤类型，并且随同给出了地块的已知森林植被类型。我们需要总共54 个特征中的其余各项来预测森林植被类型

Data Set Characteristics:	Multivariate	Number of Instances:	581012	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	54	Date Donated	1998-08-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	185453

编程——RF回归案例之共享单车租赁数量预测

- This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.
  - 数据下载 http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Data Set Characteristics:	Univariate	Number of Instances:	17389	Area:	Social
Attribute Characteristics:	Integer, Real	Number of Attributes:	16	Date Donated	2013-12-20
Associated Tasks:	Regression	Missing Values?	N/A	Number of Web Hits:	232895

用随机森林做特征选择

- 为什么要特征选择?
  - 答
- 优点
  - 高效：更简单的分割平面、更短的训练预测时间
  - 泛化能力增强：无用特征被移除
  - 可解释性增强
- 缺点
  - 计算代价
  - 选出了不好的特征的话，会影响模型精度

特征选择的思考

- 如果可以计算出每个特征的重要性，即 importance(k) for k = 1, 2, ... ,d。那就能将不重要的特征舍弃，达到降维的效果
- 那么如何考量每个特征的重要性呢?
  - 线性模型
  - 非线性模型

置换检验

- 介绍
  - 置换检验是统计学中显著性检测的一种。
- 思想
  - 如果特征k是重要的，那么用随机的值将该特征破坏，重新训练和评估，计算模型泛化能力的退化程度，即，importance(k) = performance(G) - performance(G')，这个退化程度就可以度量特征k的重要性
- 采用什么样的随机数
  - 均匀分布，高斯分布，...
  - 置换

置换检验

- 置换检验效率问题:
  - importance(k) = performance(G) - performance(G')
  - performance(G') 需要重新训练和验证，耗时耗力
- 如何避免呢?

$$importance(k) = E(G) - E(G')$$

$$importance(k) = E(G) - E'(G)$$

OOB

没有用来训练 g_t 的样本——称为 g_t 的out-of-bag(OOB) 集合中的样本

如果 N' = N

(x_i, y_i) 属于 g_t 的OOB集合的概率: $(1-\frac{1}{N})^N$

如果 N 很大 :

$$(1-\frac{1}{N})^N = \frac{1}{(\frac{N}{N-1})^N} = \frac{1}{(1+\frac{1}{N})^N} \approx \frac{1}{e} \approx 0.368$$

g_t 的OOB 集合中的元素个数 $\approx \frac{1}{e}N$

OOB集合可以用来对模型进行验证

	g_1	g_2	g_3	$\cdots$	g_T
(x_1, y_1)		*			
(x_2, y_2)	*	*			
(x_3, y_3)	*		*		
$\cdots$					
(x_N, y_N)		*	*		*

*意味着该样本属于g的OOB集合

使用OOB完成Bagging的自我验证

如何做自我验证：

用 * 来验证 g_t ？
能做，但很少需要这样做，因为 g_t 不是最终关心的，最终关心的是G

	g_1	g_2	g_3	$\cdots$	g_T
(x_1, y_1)		*			
(x_2, y_2)	*	*			
(x_3, y_3)	*		*		
$\cdots$					
(x_N, y_N)		*	*		*

用 * 来验证G
假设 x_i 同时属于某些 g_t 的OOB集合，我们将这些 g_t 组合成一个G，而不是用所有的 g_t ，记为 G_i^-，如：$G_3^-(\mathbf{x}) = \text{uniform}(g_1, g_3, \cdots)$

那么，G的误差近似表示为：

$$E_{oob}(G) = \frac{1}{N}\sum_{i=1}^{N} err(y_i, G_i^-(x_i))$$

透过OOB误差进行模型选择RF 参数

THANK YOU

上海育创网络科技有限公司