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

集成学习（Ensemble Learning）

主讲人：赵翌臣

上海育创网络科技有限公司

集成学习直观理解




+1

+1

+1

-1

sign（0.3*哆啦A梦(x)+0.4*静香(x)+0.2*小夫(x)+0.1*胖虎(x)）

新数据x

输出y

分类集成模型

集成学习直观理解




98

89

78

20

0.3*哆啦A梦(x)+0.4*静香(x)+0.2*小夫(x)+0.1*胖虎(x)

新数据x

输出y

回归集成模型

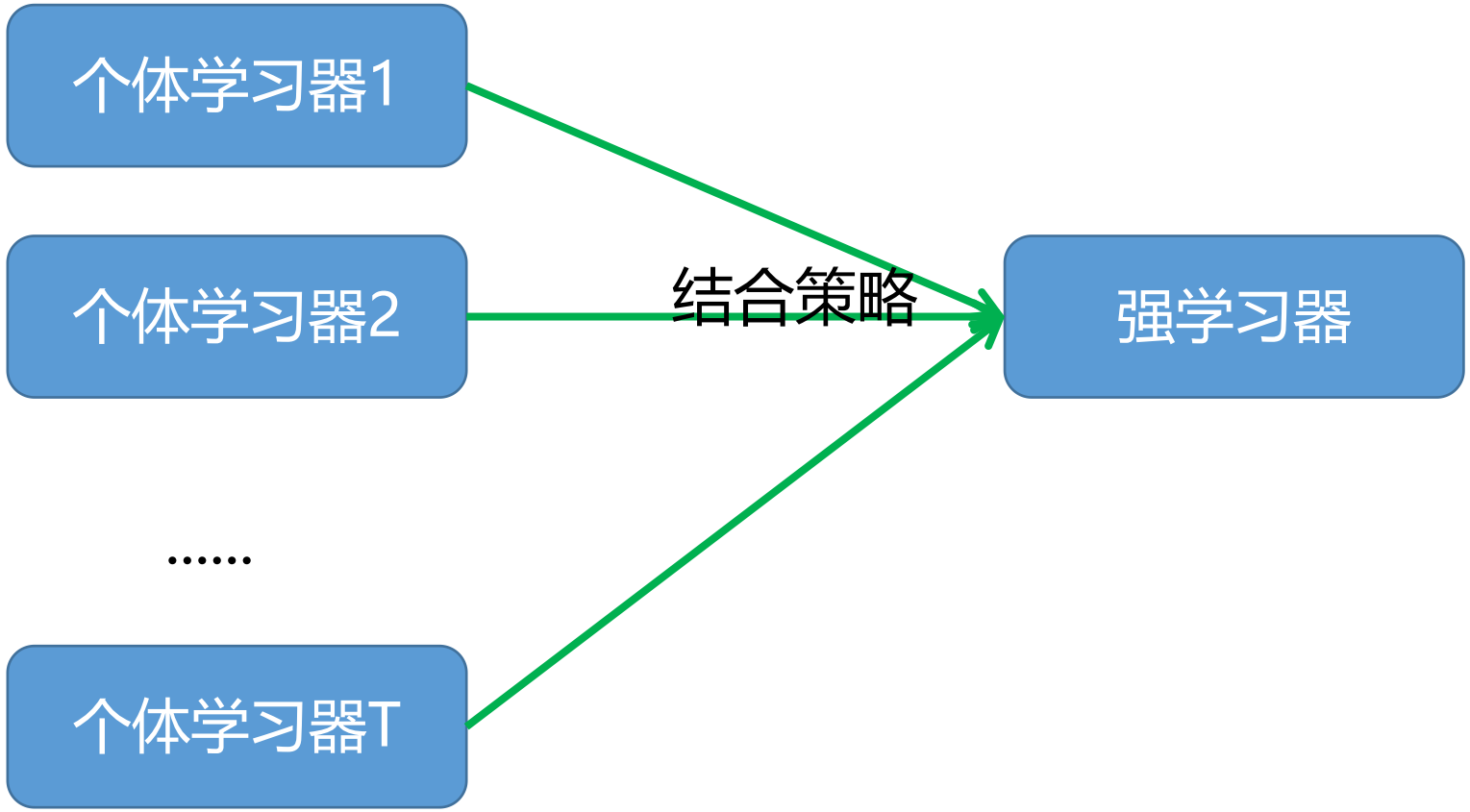
集成学习（Ensemble Learning）

- 介绍
  - 机器学习在生产、科研和生活中有着广泛应用，而集成学习则是机器学习的首要热门方向。
- 思想
  - 集成学习是训练一系列学习器，并使用某种结合策略把各个学习结果进行整合，从而获得比单个学习器更好的学习效果的一种方法。如果把单个学习器比作一个决策者的话，集成学习的方法就相当于多个决策者共同进行一项决策。集成模型不是单独的ML模型，而是通过先构建后结合多个ML模型来完成学习任务。

集成学习（Ensemble Learning）

- 如何构建和结合多个学习器
  - 暂略
- 如何决策
  - 对新的实例进行预测的时候，把个体学习器集成起来，通过对多个学习器的结果进行某种组合来决定最终的决策。

集成学习的两个问题



集成学习的两个问题

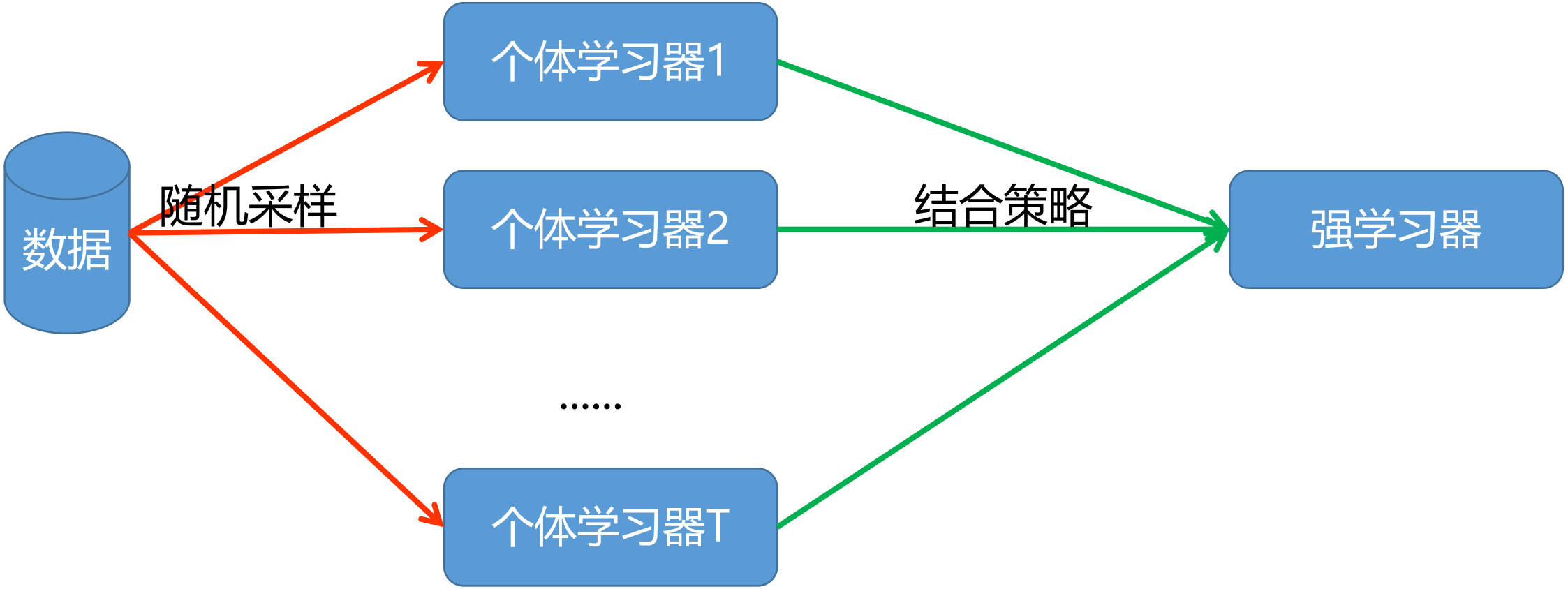
- 先构建：如何得到若干个个体学习器
  - 同质的
  - 异质的
  - 主流
- 后结合：如何选择一种结合策略，将这些个体学习器集合成一个强学习器
  - 回归
    - Boosting：直接叠加、正则后叠加、学习法（Stacking）
    - Bagging ： 平均法、带权平均法、学习法
  - 分类
    - Boosting ： 直接叠加、正则后叠加、学习法
    - Bagging ： 投票法、带权投票法、学习法

集成学习的两种思想

- Boosting思想
  - 个体学习器之间存在强依赖关系，一系列个体学习器基本都需要串行生成，然后使用组合策略，得到最终的集成模型，这就是boosting的思想
- Bagging思想（ Bootstrap AGGregatING ）
  - 个体学习器之间不存在强依赖关系，一系列个体学习器可以并行生成，然后使用组合策略，得到最终的集成模型，这就是Bagging的思想

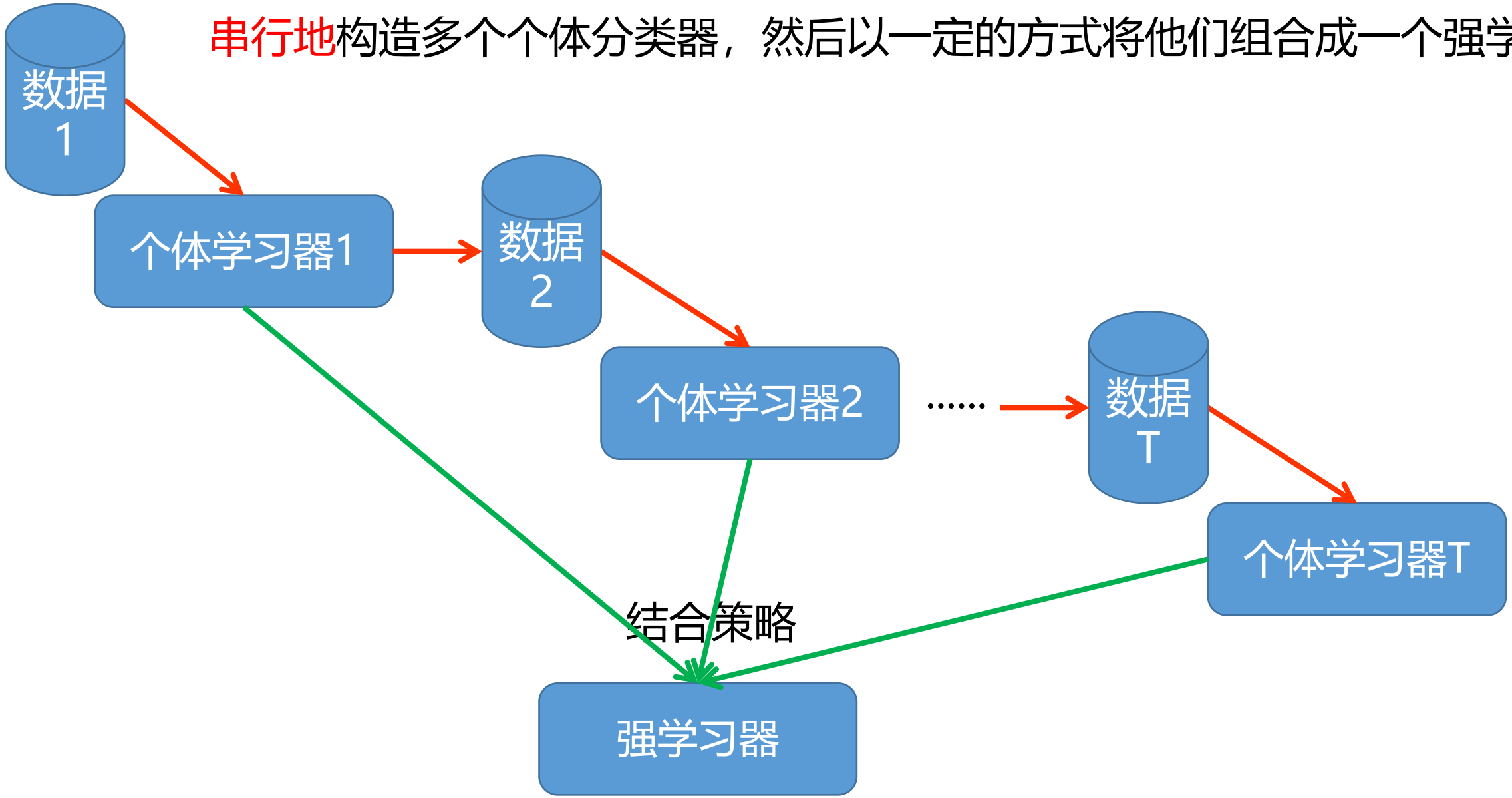
集成学习——Bagging

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



集成学习——Boosting

串行地构造多个个体分类器，然后以一定的方式将他们组合成一个强学习器



编程1——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

并行地训练多颗回归树，对样本进行预测时，所有回归树同时预测，取均值作为输出

编程2——基于Boosting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

用 $f_1(x)$ 拟合训练数据的残差见表 8.4，表中 $r_{2i}=y_i-f_1(x_i)$，$i=1,2,\cdots,10$.

表 8.4 残差表

x_i	1	2	3	4	5	6	7	8	9	10
r_{2i}	–0.68	–0.54	–0.33	0.16	0.56	0.81	–0.01	–0.21	0.09	0.14

每一轮的训练集发生变化（标签变为了残差），即下一个模型要基于新训练集进行学习

编程2——基于Boosting的回归

学习完毕后，将所有模型简单叠加，就得到了最终模型

$$f_6(x) = f_5(x) + T_6(x) = T_1(x) + \cdots + T_5(x) + T_6(x)$$

$$= \begin{cases} 5.63, & x < 2.5 \\ 5.82, & 2.5 \leqslant x < 3.5 \\ 6.56, & 3.5 \leqslant x < 4.5 \\ 6.83, & 4.5 \leqslant x < 6.5 \\ 8.95, & x \geqslant 6.5 \end{cases}$$

用 $f_6(x)$ 拟合训练数据的平方损失误差是

$$L(y, f_6(x)) = \sum_{i=1}^{10} (y_i - f_6(x_i))^2 = 0.17$$

假设此时已满足误差要求，那么 $f(x) = f_6(x)$ 即为所求提升树.

编程3——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

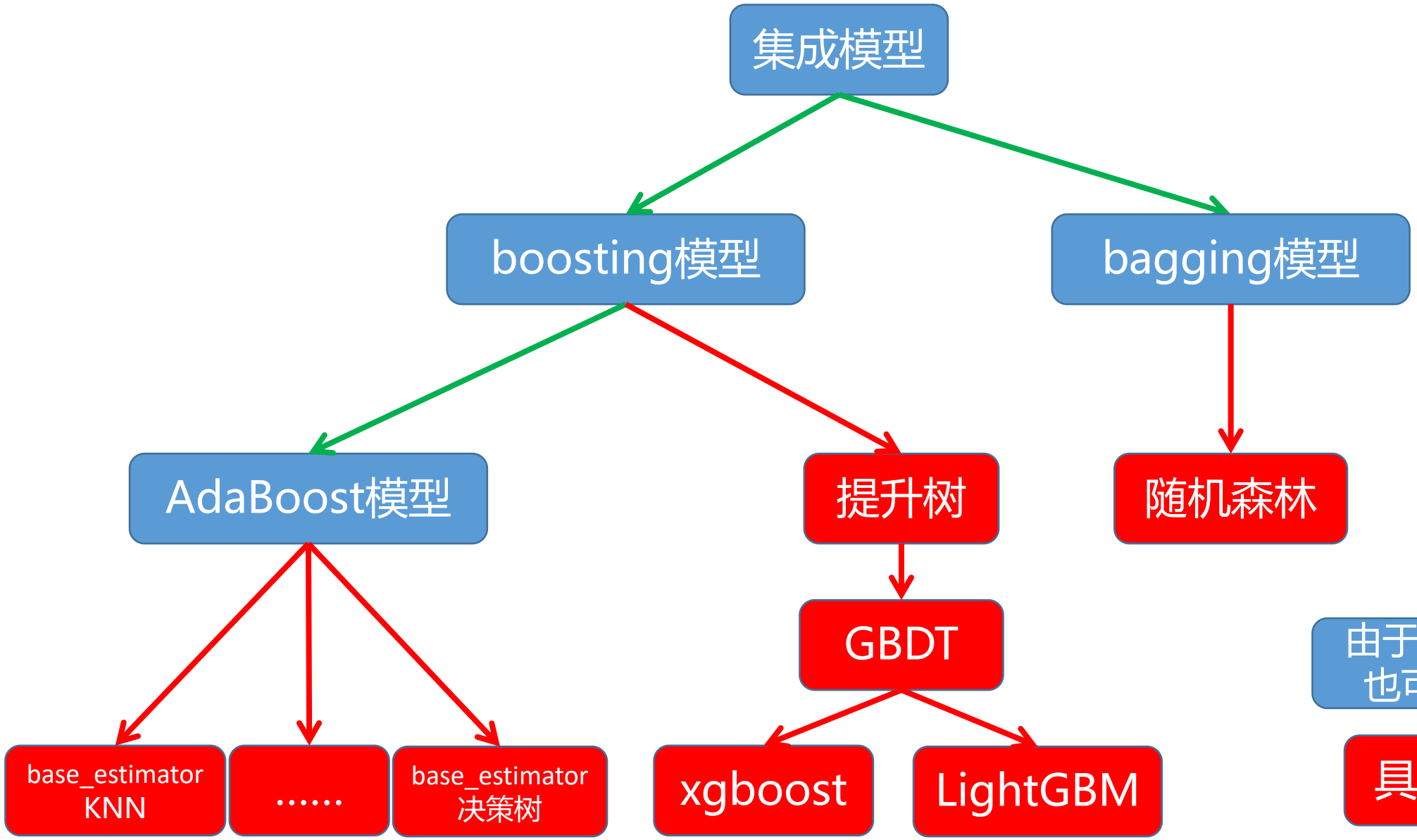
编程4——基于Boosting的分类

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

集成模型一览



由于是抽象的，
也可以叫思想

具体实现

THANK YOU

上海育创网络科技有限公司

Stacking举例：XGBoost+LR融合方案

- 背景
  - facebook于14年提出的一种融合方法，最初用于广告点击率预估（Click-Through-Rate，CTR）
- 架构说明
  - 假设训练了2颗树，一共有5个叶子结点，那么我们可以将这5个叶子结点进行编号，然后用1-k one hot来表示他们的取值，如果x样本在第一颗树中经过映射到达第2个叶子结点，在第二颗树上到达第二棵树上的第一个叶子结点，那么我们就可以得到样本经过变化后的向量为 [0,1,0,1,0] ,这5个数就表示叶子结点的，1对应的就是将样本是否落在了这个叶子结点上。直观来看，我们将一个样本向量，经过变换成了一个0,1的向量。最后我们使用经过变换后的特征再放进任意一个训练器中训练，比如说LR