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/*

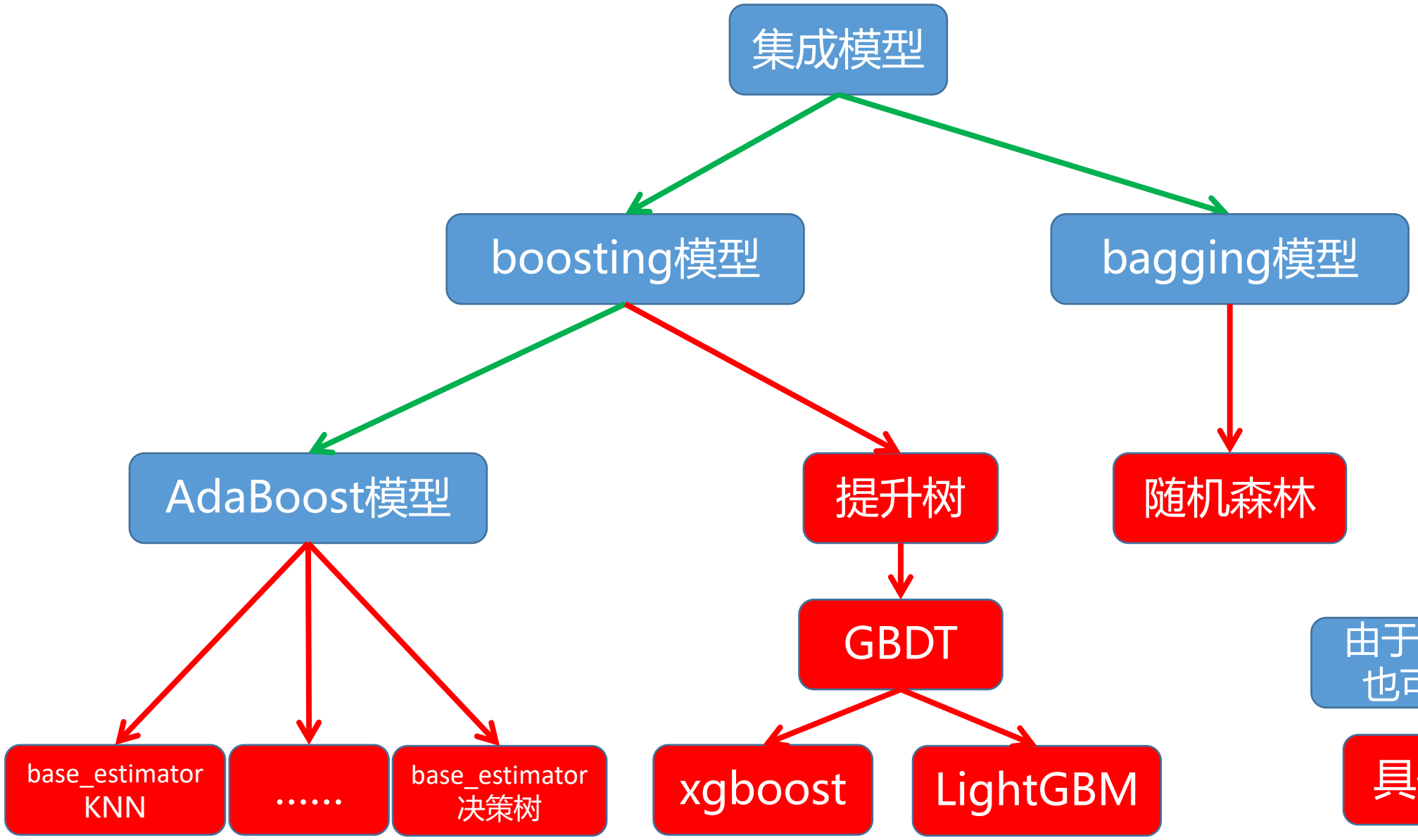
人工智能之机器学习

极限梯度提升模型（XGBoost）

主讲人：赵翌臣

上海育创网络科技有限公司

集成模型一览

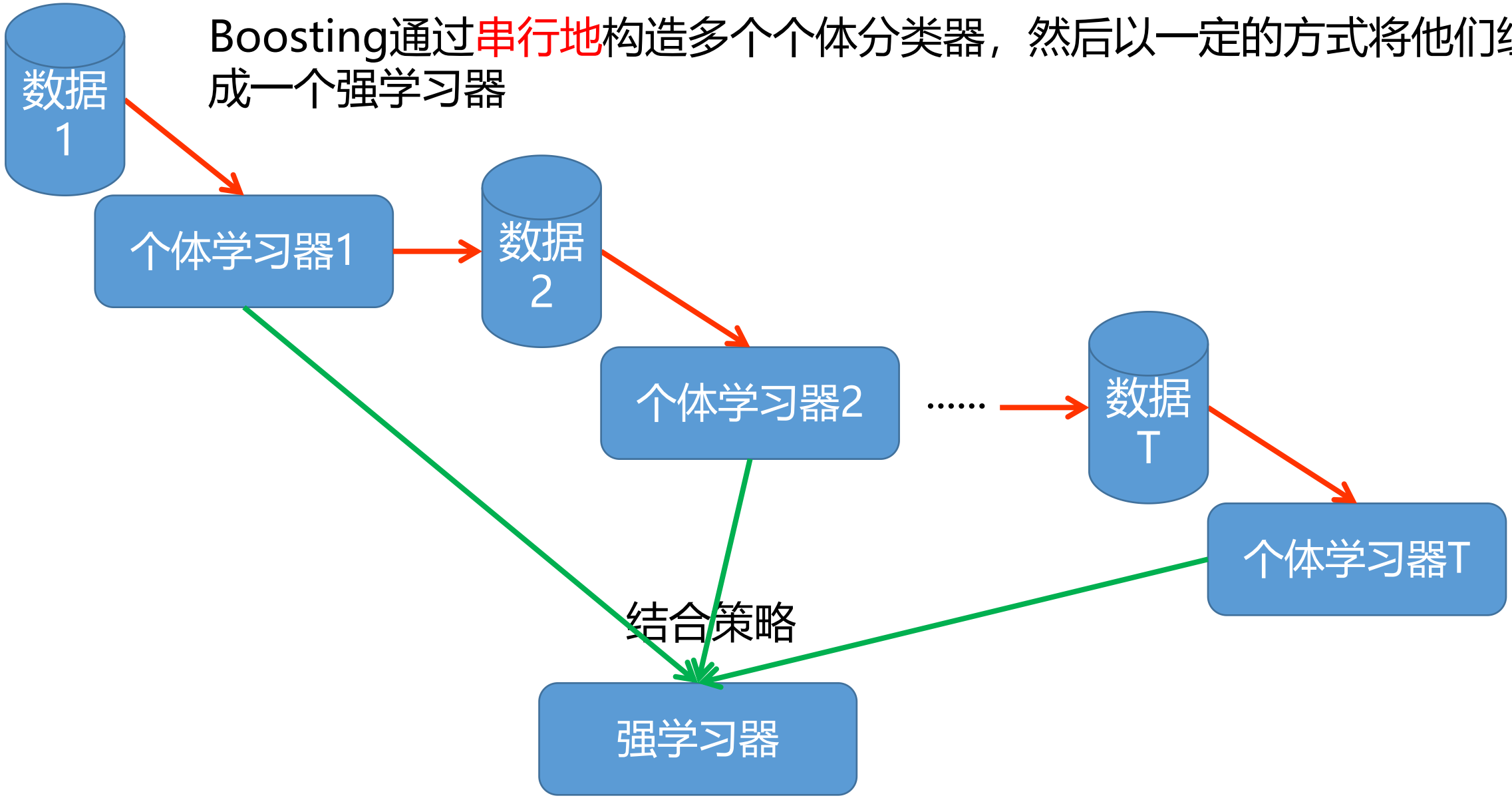


由于是抽象的，
也可以叫思想

具体实现

集成学习——Boosting思想

Boosting通过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



极限梯度提升模型（XGBoost）

- 介绍
  - XGBoost （eXtreme Gradient Boosting），是一种将Boosting做到“极致” 的方法
  - 陈天奇受AdaBoost、GBDT的启发于2014年“搞了些事情”
  - 数据竞赛表现优秀
- 作者
  - 陈天奇
  - 论文+PPT

GBDT VS XGBoost

- 区别1（构造下一轮模型的方式不同）：
  - GBDT：每一次训练构建的树是CART回归树，去拟合损失函数在当前模型的负梯度
  - XGBoost：构建目标函数（代价函数+正则化项），使用目标函数的二阶泰勒展开作为目标函数的替代，OBJ可以看做“不纯度”，Gain看做“信息增益”，利用Gain构建当前轮的回归树

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant$$

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + constant$$

$$Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma$$

GBDT VS XGBoost

- 区别2：
  - GBDT：只使用到了一阶导数信息，支持自定义损失函数，只要一阶可导即可
  - XGBoost：同时使用了一阶导数和二阶导数信息，支持自定义损失函数，只要一阶、二阶可导即可
- 区别3：
  - GBDT：只有Shrinkage防止过拟合
  - XGBoost：构建算法中就有考虑正则，融合时也有Shrinkage防止过拟合

GBDT VS XGBoost

- 区别4：
  - GBDT：只支持样本抽样（sklearn中支持了列抽样）
  - XGBoost：支持样本抽样和列抽样，借鉴了随机森林的做法，支持列抽样，不仅能降低过拟合，还能减少计算，这也是xgboost异于传统gbdt的一个特性。
- 区别5：
  - GBDT：串行算法
  - XGBoost：并行算法，？？粒度上实现并行，各个特征的增益开多线程进行计算。

Xgboost优缺点

- 优点
  - 正则化防止过拟合效果好
  - 特征粒度上并行
  - 使用泰勒公式，提升了代价函数的灵活性
  - 内置交叉验证：XGBoost 允许在每一轮 Boosting 迭代中使用交叉验证
- 缺点
  - 算法参数过多
  - 只适合处理结构化数据
  - 不适合处理超高维特征数据

XGBoost安装

- 下载xgboost离线包
  - https://www.lfd.uci.edu/~gohlke/pythonlibs/
  - cd到whl的目录，打开控制台执行pip install xxx.whl
- 文档
  - http://xgboost.readthedocs.io/en/latest/python/

编程——XGBoost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——XGBoost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

LightGBM

LightGBM面试要点

- 介绍
  - GBDT是很流行的机器学习算法，他有一些高效的实现，比如 XGBoost和LightGBM
  - LightGBM是2017年微软公司推出的，解决了GBDT在大规模训练时非常耗时的问题。
- 创新点1
  - 在寻找最优划分点时，使用预排序算法需要更多的时空成本，使用直方图算法对连续特征进行分桶可以将时间复杂度从O(#data × #feature)降为 O(#bin × #feature)，空间也会节省很多，因为一个桶中的样本在该特征上都相同

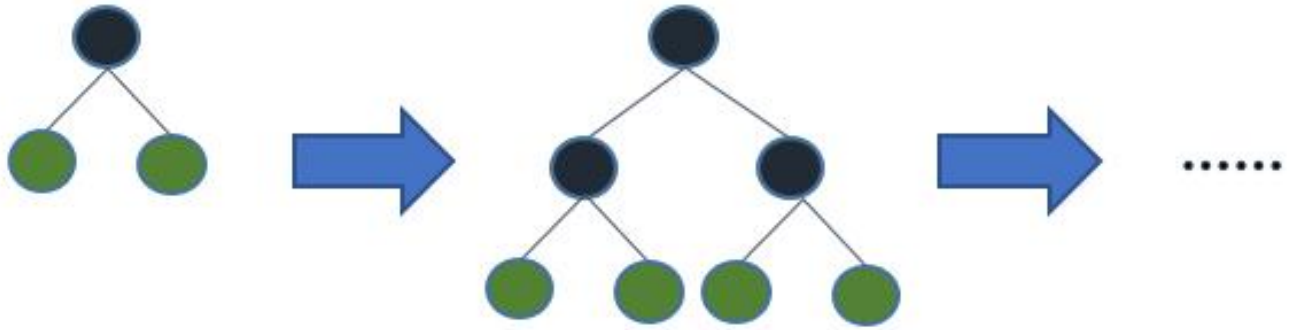
LightGBM面试要点

- 创新点2
  - 方法论：①减少样本数量②样本维度都可以加快计算
  - ①梯度单边采样（GOSS）：
    - 梯度值越小的样本，可以认为他已经训练好了，能带给我们的信息量也就越小，因此可以抛弃掉那些梯度值小的样本。GOSS先对样本按照梯度值排序，取前a%的样本，再在其余的样本中随机选择b%的样本，在计算信息增益的时候需要给那些小权重的样本以补偿，即*(1-a)/b
  - ②互斥特征合并（EFB）：
    - 可以理解为一种降维方法

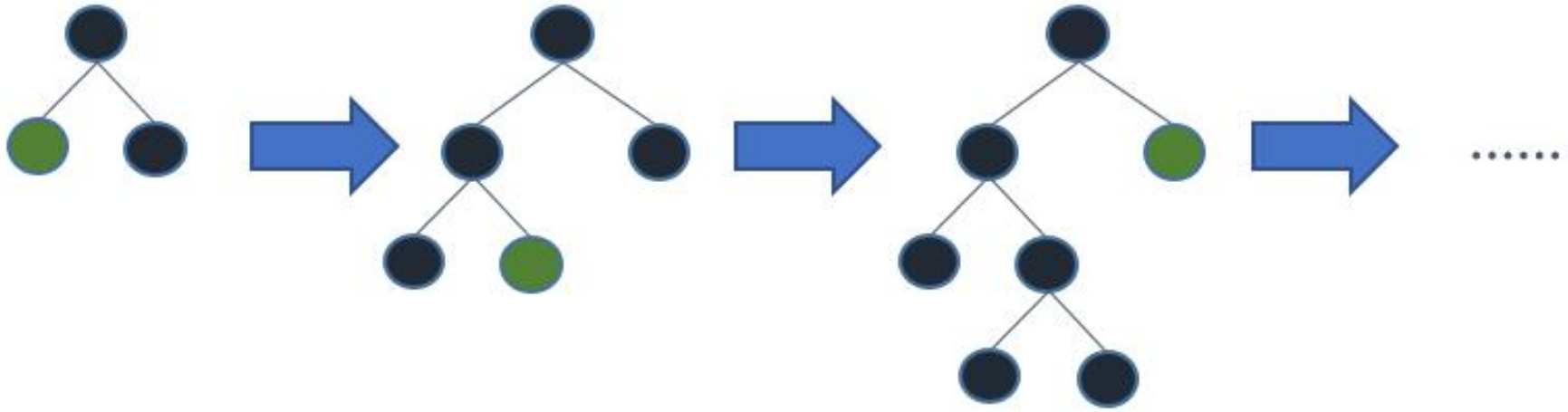
LightGBM面试要点

- 创新点3
  - 带深度限制的Leaf-wise的叶子生长策略
  - Level-wise过一次数据可以同时分裂同一层的叶子，容易进行多线程优化，也好控制模型复杂度，不容易过拟合。但实际上Level-wise是一种低效算法，因为它不加区分的对待同一层的叶子，带来了很多没必要的开销，因为实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。
  - Leaf-wise则是一种更为高效的策略：每次从当前所有叶子中，找到分裂增益最大的一个叶子，然后分裂，如此循环。因此同Level-wise相比，在分裂次数相同的情况下，Leaf-wise可以降低更多的误差，得到更好的精度。
  - Leaf-wise的缺点：可能会长出比较深的决策树，产生过拟合。因此LightGBM在Leaf-wise之上增加了一个最大深度限制，在保证高效率的同时防止过拟合。

LightGBM面试要点



Level-wise tree growth



Leaf-wise tree growth

LightGBM论文分析《A Highly Efficient Gradient Boosting Decision Tree》

LightGBM安装

- 在线安装
  - pip install lightgbm
- 文档
  - https://lightgbm.readthedocs.io/en/latest/Python-API.html#scikit-learn-api

编程——LightGBM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——LightGBM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

THANK YOU

上海育创网络科技有限公司

Histogram optimization

Convert feature value to bin before training

Use bin to index histogram, not need to sort

Reduce computation cost of split gain

Algorithm: **FindBestSplitByHistogram**

Input: Training data X **, Current Model** $T_{c-1}(X)$

First order gradient G, second order gradient H

For all Leaf p in $T_{c-1}(X)$:

For all f in X.Features:

▷ construct histogram

H = new Histogram()

For i in (0, num_of_row) //go through all the data row

$H[f.bins[i]].g \mathrel{+}= g_i; H[f.bins[i]].n \mathrel{+}= 1$

▷ find best split from histogram

For i in (0,len(H)): //go through all the bins

$S_L \mathrel{+}= H[i].g; n_L \mathrel{+}= H[i].n$

$S_R = S_P - S_L; n_R = n_P - n_L$

$\Delta loss = \frac{S_L^2}{n_L} + \frac{S_R^2}{n_R} - \frac{S_P^2}{n_P}$

if $\Delta loss > \Delta loss(p_m, f_m, v_m)$:

$(p_m, f_m, v_m) = (p, f, H[i].value)$