什么是XGBoost

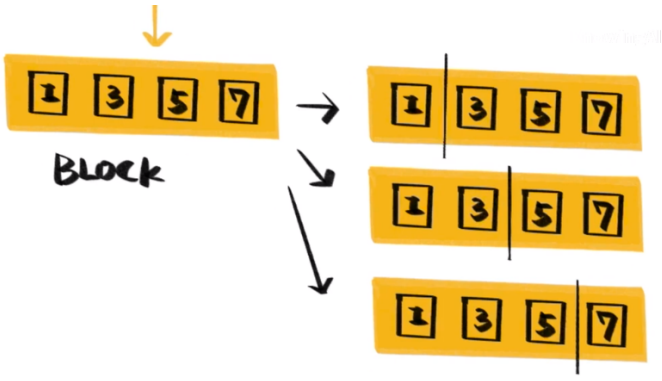
XGBoost是GBDT的优秀版本。XGBoost的整体结构和GBDT一致，都是在训练出一棵树的基础上，再训练下一棵树，预测它与真实分布间的差距，通过不断训练用来弥补差距的树，最终用树的组合实现对真实分布的模拟。

当然，XGBoost有自己的独特之处，我们训练模型通常是定义一个**目标函数**，然后去优化它，XGBoost的**目标函数**包含**损失函数**和正则项两部分。

损失函数代表着**模型拟合数据的程度**，我们通常用它的一阶导数指出梯度下降的方向，XGBoost还计算了它的二阶导数，进一步考虑了梯度变化的趋势，拟合更快，精度更高。

正则项则被用来控制模型的复杂程度，叶子结点越多，模型越大，不仅运算时间长，超过一定限度后还会过拟合，导致分类效果的下降。XGBoost的正则项是一个惩罚机制，叶子结点的数量越多，惩罚力度越大，从而限制他们的数量。

数学原理外，XGBoost最大的改进是大幅提升了计算速度，树的构建中，最耗时的部分是为确定最佳分裂点而进行的**特征值排序**。XGBoost在训练前会将特征进行排序，存储为**Block**结构，此后重复使用这些结构，从而减少计算量。



善于捕捉复杂数据之间的依赖关系，能从大规模数据集中获取有效模型， 在实用性上支持多种系统和语言，这些都是XGBoost的优点。它同样有缺陷，比如在高维系数特征数据集和小规模数据集上表项不是很好。

在XGBoost之后还出现了LightGBM、CatBoost等Boosting方法，在提升运算速度，处理类别型特征等方面各有千秋。

LIGHTGBM → HISTOGRAM

CATBOOST → CATEGORICAL FEATURE