## ****5. XGBoost的参数****

XGBoost的作者把所有的参数分成了三类：

1. 通用参数：宏观函数控制。
2. Booster参数：控制每一步的booster(tree/regression)。
3. 学习目标参数：控制训练目标的表现。

**5.1 通用参数**

这些参数用来控制XGBoost的宏观功能。

**1、booster[默认gbtree]**

* 选择每次迭代的模型，有两种选择：   
  gbtree：基于树的模型   
  gbliner：线性模型

**2、silent[默认0]**

* 当这个参数值为1时，静默模式开启，不会输出任何信息。
* 一般这个参数就保持默认的0，因为这样能帮我们更好地理解模型。

**3、nthread[默认值为最大可能的线程数]**

* 这个参数用来进行多线程控制，应当输入系统的核数。
* 如果你希望使用CPU全部的核，那就不要输入这个参数，算法会自动检测它。

还有两个参数，XGBoost会自动设置，目前你不用管它。接下来咱们一起看booster参数。

**5.2 booster参数**

尽管有两种booster可供选择，我这里只介绍**tree booster**，因为它的表现远远胜过**linear booster**，所以linear booster很少用到。

**1、eta[默认0.3]**

* 和GBM中的 learning rate 参数类似。
* 通过减少每一步的权重，可以提高模型的鲁棒性。
* 典型值为0.01-0.2。

**2、min\_child\_weight[默认1]**

* 决定最小叶子节点样本权重和。
* 和GBM的 min\_child\_leaf 参数类似，但不完全一样。XGBoost的这个参数是最小*样本权重的和*，而GBM参数是最小*样本总数*。
* 这个参数用于避免过拟合。当它的值较大时，可以避免模型学习到局部的特殊样本。
* 但是如果这个值过高，会导致欠拟合。这个参数需要使用CV来调整。

**3、max\_depth[默认6]**

* 和GBM中的参数相同，这个值为树的最大深度。
* 这个值也是用来避免过拟合的。max\_depth越大，模型会学到更具体更局部的样本。
* 需要使用CV函数来进行调优。
* 典型值：3-10

**4、max\_leaf\_nodes**

* 树上最大的节点或叶子的数量。
* 可以替代max\_depth的作用。因为如果生成的是二叉树，一个深度为n的树最多生成n2个叶子。
* 如果定义了这个参数，GBM会忽略max\_depth参数。

**5、gamma[默认0]**

* 在节点分裂时，只有分裂后损失函数的值下降了，才会分裂这个节点。Gamma指定了节点分裂所需的最小损失函数下降值。
* 这个参数的值越大，算法越保守。这个参数的值和损失函数息息相关，所以是需要调整的。

**6、max\_delta\_step[默认0]**

* 这参数限制每棵树权重改变的最大步长。如果这个参数的值为0，那就意味着没有约束。如果它被赋予了某个正值，那么它会让这个算法更加保守。
* 通常，这个参数不需要设置。但是当各类别的样本十分不平衡时，它对逻辑回归是很有帮助的。
* 这个参数一般用不到，但是你可以挖掘出来它更多的用处。

**7、subsample[默认1]**

* 和GBM中的subsample参数一模一样。这个参数控制对于每棵树，随机采样的比例。
* 减小这个参数的值，算法会更加保守，避免过拟合。但是，如果这个值设置得过小，它可能会导致欠拟合。
* 典型值：0.5-1

**8、colsample\_bytree[默认1]**

* 和GBM里面的max\_features参数类似。用来控制每棵随机采样的列数的占比(每一列是一个特征)。
* 典型值：0.5-1

**9、colsample\_bylevel[默认1]**

* 用来控制树的每一级的每一次分裂，对列数的采样的占比。
* 我个人一般不太用这个参数，因为subsample参数和colsample\_bytree参数可以起到相同的作用。但是如果感兴趣，可以挖掘这个参数更多的用处。

**10、lambda[默认1]**

* 权重的L2正则化项。(和Ridge regression类似)。
* 这个参数是用来控制XGBoost的正则化部分的。虽然大部分数据科学家很少用到这个参数，但是这个参数在减少过拟合上还是可以挖掘出更多用处的。

**11、alpha[默认1]**

* 权重的L1正则化项。(和Lasso regression类似)。
* 可以应用在很高维度的情况下，使得算法的速度更快。

**12、scale\_pos\_weight[默认1]**

* 在各类别样本十分不平衡时，把这个参数设定为一个正值，可以使算法更快收敛。

**5.3学习目标参数**

这个参数用来控制理想的优化目标和每一步结果的度量方法。

**1、objective[默认reg:linear]**

* 这个参数定义需要被最小化的损失函数。最常用的值有：   
  + binary:logistic 二分类的逻辑回归，返回预测的概率(不是类别)。
  + multi:softmax 使用softmax的多分类器，返回预测的类别(不是概率)。   
    - 在这种情况下，你还需要多设一个参数：num\_class(类别数目)。
  + multi:softprob 和multi:softmax参数一样，但是返回的是每个数据属于各个类别的概率。

**2、eval\_metric[默认值取决于objective参数的取值]**

* 对于有效数据的度量方法。
* 对于回归问题，默认值是rmse，对于分类问题，默认值是error。
* 典型值有：   
  + rmse 均方根误差(∑Ni=1ϵ2N−−−−−√)
  + mae 平均绝对误差(∑Ni=1|ϵ|N)
  + logloss 负对数似然函数值
  + error 二分类错误率(阈值为0.5)
  + merror 多分类错误率
  + mlogloss 多分类logloss损失函数
  + auc 曲线下面积

**3、seed(默认0)**

* 随机数的种子
* 设置它可以复现随机数据的结果，也可以用于调整参数