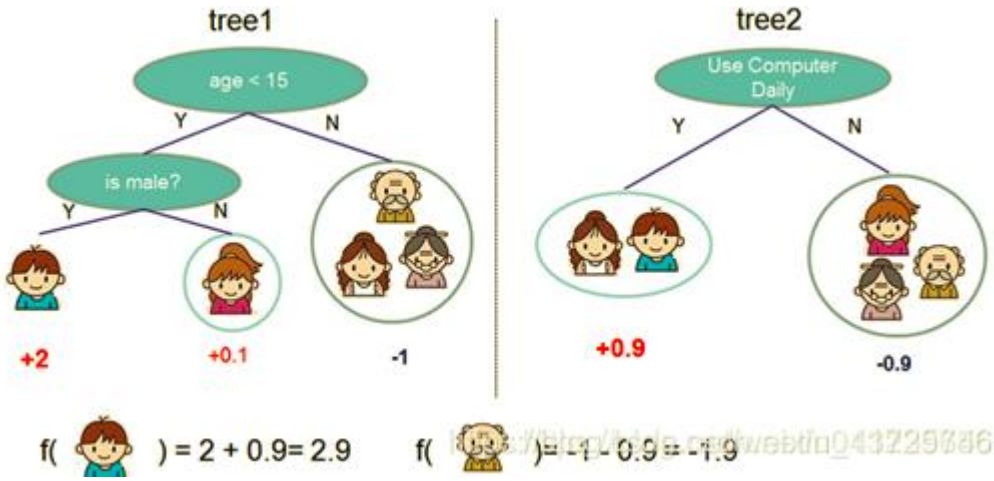
极度梯度提升树 Xgboost

eXtreme Gradient Boosting

XGBoost(eXtreme Gradient Boosting)又叫极度梯度提升树，是 boosting 算法的一种实现方式。针对分类或回归问题，效果非常好。在各种数据竞赛中大放异彩，而且在工业界也是应用广泛，主要是因为其效果优异，使用简单，速度快等优点。

XGBoost 是陈天奇提出的一个端对端的梯度提升树系统，该算法在 GBDT 的基础之上，在算法层面和系统设计层面都做了一些创新性的改进，可以把 XGBoost 看作是 GBDT 更好更快的实现。

Xgboost 是 Boosting 算法的一种实现方式，主要是降低偏差，也就是降低模型的误差。因此它是采用多个基学习器，每个基学习器都比较简单，避免过拟合，下一个学习器是学习前面基学习器的结果和实际值的差值，通过多个学习器的学习，不断降低模型值和实际值的差。



Xgboost 在算法层面的一些创新性改进包括：

(1) 在 GBDT 目标函数的基础上，在对优化目标求解的时候使用了二阶导数的信息，因此会使优化目标的定义更加精确，训练速度会更快；此外，XGBoost 在优化目标函数中加入了正则项，这会限制模型的复杂度，防止过拟合的发生。

(2) 提出了一种处理缺失值的算法，XGBoost 能够自动对缺失值进行处理

(3) 在树生成选择特征划分节点的时候，通过加权分位数草图算法，预先对每个特征建立候选的划分节点，不在使用原先的贪婪算法（遍历每个特征所有取值），从而大大加快训练速度.

XGBoost 在许多机器学习以及数据挖掘的任务中表现惊艳，2015 年，kaggle 竞赛平台上发布了 29 个挑战获胜的解决方案，其中 17 个解决方案用了 XGBoost。