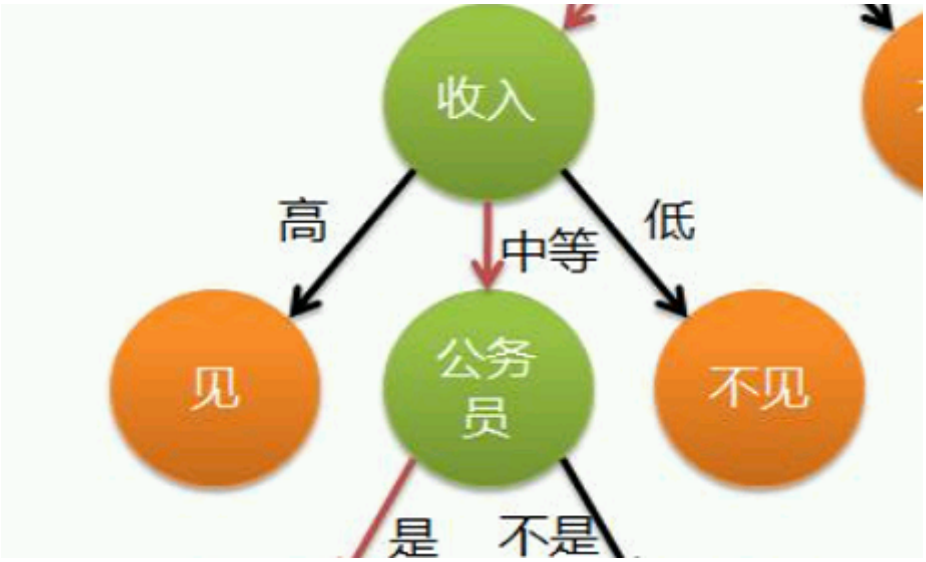
决策树

- 选择哪个属性，哪个根节点，哪个在前

  信息争议：熵（描述信息的混乱程度，未知的可能性，信息的不确定性）变化后的差值。

  优先选择信息争议最大的。
- 按照什么标准分
    - 连续型：如年龄【均方差最小分分支】
    - 离散型（类别型）【有几个级别几个分支】
    - 连续 -> 离散 （连续数据离散化）：
        - 划分区间（变成离散值）

离散型：



变成二叉树：

剪枝

- 预剪枝
- 后剪枝

目标函数

- 构建一个函数，把特征放进去得出结果和预测的吻合。
- 拟合

损失函数

- 尽可能小，为0时，拟合的函数与实际最吻合。

凸优化

- 凸函数，有全局最优解，圆锥曲线开口向上或向下
- 梯度下降
- 求导斜率判断梯度方向
- 一阶导数体现趋势，二阶导体现速率

最优化

- 求极值，

过拟合，欠拟合

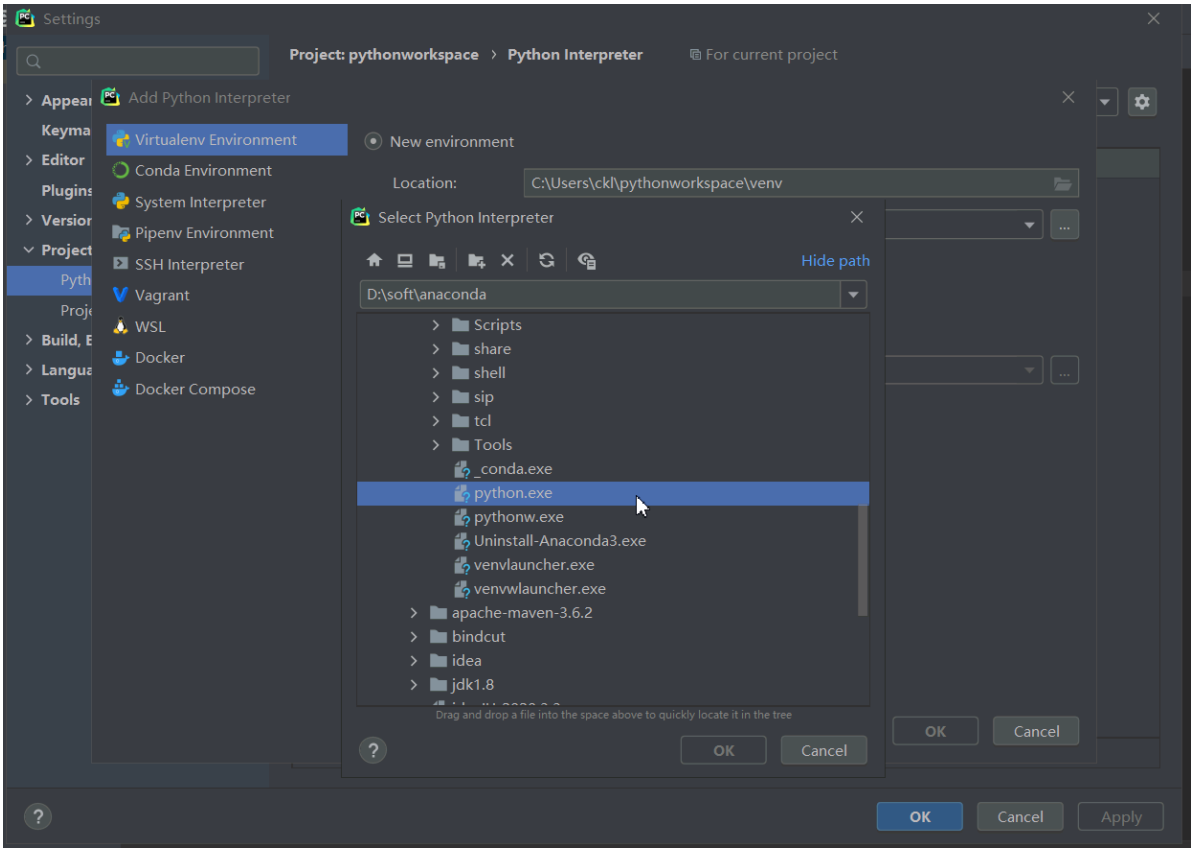
泛化能力

LR无法拟合非线性场景

数据科学家足够了解业务，来挖掘高阶特征，组合发掘LR

出现GBDT后

python中下载用pip 相当于yum



预测可以用nginx等用c底层调用

pmml比较中庸，不快不慢

https://xgboost.readthedocs.io/en/latest/build.html