决策树

ID3

- 利用信息增益来选择特征
  - 信息增益是减法计算，H(D)-H(D/A)
  - 利用信息增益很可能会选有更多属性的特征
  - sklearn.tree.DecisionTreeClassifier(entropy)

C4.5

- 利用信息增益比来选择特征
  - 信息增益比是除法运算 （H(D)-H(D/A))/splitinformation（D,A）
  - 惩罚了有更多属性的特征，弥补ID3
  - sklearn.tree.DecisionTreeClassifier(entropy)

CART

- 对回归用MSE，分类用GINI
  - GINI指数越大，样本的不确定性越大
  - sklearn.tree.DecisionTreeClassifier(gini)
  - 信息增益最大~信息增益比最大~基尼指数最小

注意

- 当样本数少于特征数时，容易过拟合，最好先用PCA,LDA等提取主成分
- 如果类别分布非常不均匀，就要考虑用class_weight
- 如果样本中有string，可以用labelEncoder、OneHotEncoder或者DictVectorizer
- 如果想把连续型数据类别化，可以用分桶的思想，或者二值化
- 基于树的计算都不需要对数据做标准化处理

sklearn.tree.DecisionTreeClassifier参数

- criterion：gini/entropy 一般选gini
- splitter:样本不大选best，反之random 全局最优和局部最优的关系
- max_features：None考虑全部特征，其它的为考虑N的对数、平方根，看样本大小
- max_depth：一般不超过100
- min_samples_split：一般10万数据控制在10左右
- min_samples_leaf：一般10万数据在5个左右
- min_weight_fraction_leaf：如果想给某个叶子节点特定的权重
- max_leaf_nodes：最大叶子节点数
- class_weight：类别权重,balanced 样本少的类别会赋予跟高的权重
- min_impurity_split：不纯度小于这个阈值，则该节点不再生成子节点。即为叶子节点 。
- presor：数据是否预排序，一般false

总概：决策树DT是一种分类和回归的方法。

- 决策树可以认为是一种if-then的集合，即互斥且穷尽。
- 可以看做是一种条件概率分布，基于特征的类划分。
- 3个步骤：特征选择、决策树生成、减枝