决策树 实验记录

11月4日

特征提取的思考：
离散特征时，使用matplot画图，寻找更加的中间值（如有不同的label）（尚未完成）

建决策树之前，先定义一些用来计算的函数：

1. 计算权重
2. 计算众数
3. 计算信息熵
4. 计算信息增益
5. 计算信息增益率
6. 计算Gini值

先试着建一棵能跑的简单树（没有剪枝）
第一次用的是Gini，输出数据有329个无效的，827个0，820个1，准确率0.5354。
将无效数据预测为0后，正确率为0.6169。（及格了，狗头.jpg）

Total: 1976

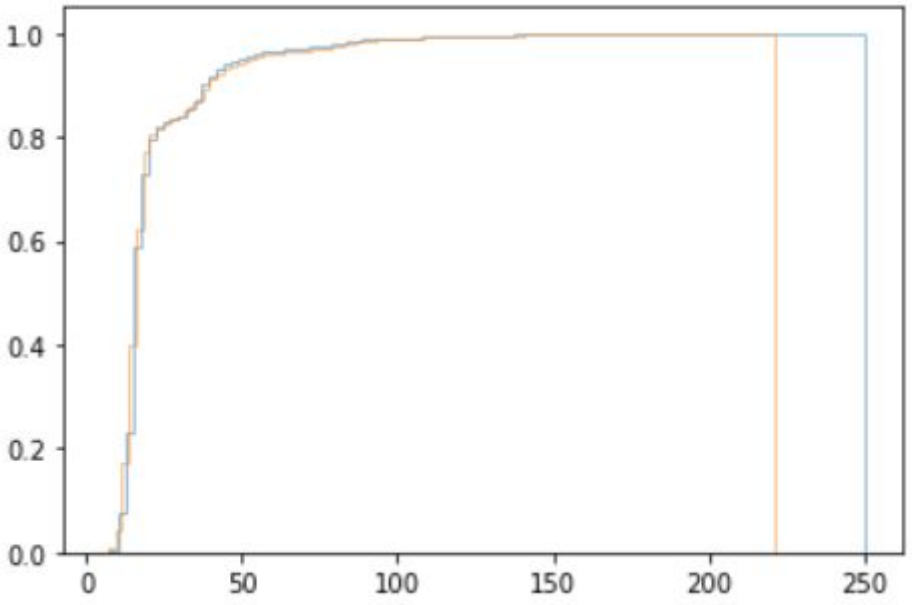
method	accuracy	# invalid	1	0	correct	Accuracy without invalid（让自己开心开心）
Gini	0.5354	329	820	827	1058	0.6423
Info Gain	0.5481	308	742	926	1083	0.6493
Gain Ratio	0.5481	308	742	926	1083	0.6493

需要改善的问题：0log0 = 0
特征值离散改进
pre/post pruning

11月5日

关于特征值离散改进：
每个特征根据label分为两类，plot cumulative graph
C = e ^(1 - cwin/(close+1))/e 但不知道怎么写
取的参数：cwin或者close 0,0.001,0.01,0.05,0.1,0.2, 0.5, 0.75, 0.9, 0.99, 1
或者：去掉极值后（即 cwin<1%/close>99%, 或者异常数值），中间部分等分划分（这个看上去比较简单？）
实现方法：sort list, 取list[int(0.01 * len(list))] / list[int(0.99 * len(list))], 中间分15？份

redKills, redDeaths 和blueDeaths, blueKills重复，删除。
可能需要重新离散的数据：blueWardsPlaced （这个和输赢好像关系不大）



可调参数：lowerbound, upperbound,step

method	accuracy	# invalid	1	0	correct	Accuracy without invalid（好像也没有很开心）	Invalid -> 0
Gini（original）	0.5354	329	820	827	1058	0.6423	0.6169
Gini（	0.5506	312	812	852	1089	0.6544	0.6203

new）							

设置了max_depth = 20 和 min_samples_split = 10 后：（稍微开心一点了）

Gini（new2）	0.6204	109	846	1021	1226	0.6567	0.6442

思考：要调整参数的话，需要从train里分一个val组出来
为什么还有那么多invalid？1. 特征离散的时候，有些组数据太少，树上没有对应的子叶（可以考虑每个节点增加一个default子叶）；2.树是不是过拟合了？调整pre pruning；增加post pruning。
如何寻找更优的离散（寻找特征）方案？
考虑增加每个特征值的权重？考量特征值（x）和label（y）之间的关系，maybe r?
解决方案:

1. 分出val组（training:7112, val:791, test:1976）
2. 调整离散方案：
    a. 去极值后，从等分改为等比？可调参数：极值的定义，区间大小 Accuracy为0.5967，不理想。
3. 调整返回的值，从众数，改成1的概率，同时修改predict_函数（增加用来计算概率和根据概率预测1或0的函数）Accuracy: 0.6235
4. 调整返回的值，当prob1（y）> 0.9 返回1, <0.1返回0。Accuracy: 0.6280

准备剪枝：
用划分好的training先生成一棵“枝繁叶茂”的大树

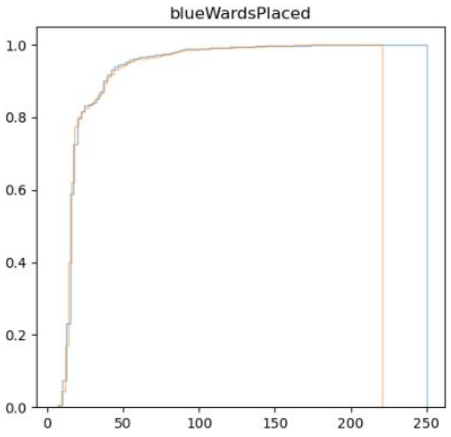
先定义一些函数：
计算在val上的准确率

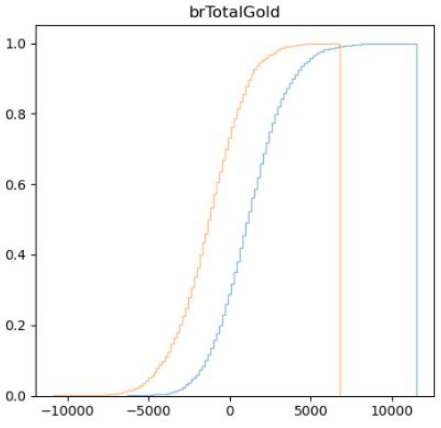
考虑问题：
怎么判断到达叶子：用for value in dict1.values(): isinstance(value, int or float)
怎么建新树：1. 在老树上找到下一个要尝试剪的结点（每次返回一个新树）

后剪枝遇到问题：为什么越剪越差了？
于是去调了一下pre的参数，当最大深度为2时，accuracy = 0.7171。（手动狗头，所以之前在干嘛？？）

11月6日

进一步细化特征离散
画累积直方图，考量每个特征和标签的关联性。
参数：bins = 100, alpha = 0.5, histtype='step', cumulative = True,density=True
发现：有些特征和标签关联性不强（可以舍去），有些关联性较强





舍去的特征包括：
'blueWardsPlaced', 'blueWardsDestroyed', 'blueTotalJungleMinionsKilled', 'redWardsPlaced', 'redWardsDestroyed', 'redTowersDestroyed', 'redTotalJungleMinionsKilled', 'brWardsPlaced'
剩余特征数：33
离散之后，再画一次图，
等比或等区间划分，感觉离散效果差不多。

Train : val : test = 7112 : 791 : 1976 约为 7 : 1 : 2

当label值不同时，用众数还是1的概率的思考：
在此次实验中，输赢的概率各占一半
假设这个节点1的概率为0.8：(用概率的准确率，众数的准确率)
实际概率为0.8：0.8*0.8 + 0.2*0.2 = 0.68 < 0.8；
实际概率为0.5：0.8*0.5 + 0.2*0.5 = 0.5 = 0.5；
实际概率为0.2：0.8*0.2 + 0.2*0.8 = 0.32 > 0.2。
由此可见当实际概率大于0.5时，即接近模型预测的概率时，使用众数效果较好。
基于对自己建的模型的自信（不是），我决定用众数。

#寻找最佳深度
depth: 1 accuracy:0.7219
invalid 0 illusion accuracy:0.7219
depth: 2 accuracy:0.7130
invalid 0 illusion accuracy:0.7130
depth: 3 accuracy:0.6726

invalid 5 illusion accuracy:0.6776
depth: 4 accuracy:0.6397
invalid 18 illusion accuracy:0.6498
depth: 5 accuracy:0.6384
invalid 20 illusion accuracy:0.6498
depth: 6 accuracy:0.6384
invalid 20 illusion accuracy:0.6498
depth: 7 accuracy:0.6384
invalid 20 illusion accuracy:0.6498
在树长到4层以后，因为预剪枝的效果，不再继续生长。

Error reduction pruning
两层树后剪枝前: 0.7130，（测试集：0.7100), 剪枝后在验证集上的accuracy: 0.7231（测试集：0.7100）
三层树后剪之前：0.6726，（测试集：0.6888）；剪枝后在验证集上的accuracy：0.7408 （测试集：0.6943）
四层树后剪之前：0.6397，（测试集：0.6579）；剪枝后在验证集上的accuracy：0.7826（测试集：0.6705）
三层树/四层树剪枝有效，但可能在验证集上过拟合了。

想要尝试一下规则后剪枝，但是，对于怎么写代码毫无头绪，于是决定，手动剪枝。
准备：离散的时候，划分区间数减少为一共4个区间试试。
生成的树：{'#': 28,
 0: 0,
 1: {'#': 24, 0: 0, 1: 0, 2: 1, -2: 0, -1: 0},
 2: {'#': 30, 1: 1, 2: 1, 3: 1},
 3: 1}
化简后：{'#': 28,
 0: 0,
 1: {'#': 24, 0: 0, 1: 0, 2: 1, -2: 0, -1: 0},
 2: 1
 3: 1}
存在的规则：f[28] f[24]
若先判断 f[28]， 最优树已经生成了，accuracy 在val上0.7320 （test: 0.7120)。
如果先判断 f[24],
DT2.tree = {'#': 24,
        2: 1,
        0:{'#': 28, 0: 0, 1: 0, 2: 0, 3: 1},
        1:{'#': 28, 0: 0, 1: 0, 2: 0, 3: 1},
        -1:{'#': 28, 0: 0, 1: 0, 2: 0, 3: 1},
        -2:{'#': 28, 0: 0, 1: 0, 2: 0, 3: 1}}
然后借用postpruning剪枝，得到：{'#': 24,

2: 1,
0: {'#': 28, 0: 0, 1: 0, 2: 0, 3: 1},
1: 1,
-1: {'#': 28, 0: 0, 1: 0, 2: 0, 3: 1},
-2: 0}
Accuracy在val上为0.6283 （test: 0.6215）
所以选择原本的树。
最终accuracy 0.7120。
1976个测试样本中，有570个被误分类
标准差 S = sqrt(1 * 570/1975) = 0.5372
标准误差 SEM = 0.537222/sqrt(1976) = 0.01208
σ (error) = sqrt((15/52*(1-15/52)/1976) = 0.01019
实际accuracy为0.7115 +-1.96 * σ = 0.7115 +-0.0098的置信度为95%。

7个等区间，总共9个区间。
depth: 1 accuracy:0.7080
invalid 0 illusion accuracy:0.7080
depth: 2 accuracy:0.7332
invalid 0 illusion accuracy:0.7332
depth: 3 accuracy:0.6903
invalid 2 illusion accuracy:0.6915
depth: 4 accuracy:0.6662
invalid 12 illusion accuracy:0.6751

深度2，accuracy：val上0.7320->0.7358, test上：0.7267
深度3, accuracy: val上0.6903 -> 0.7421, test上：0.7120
明天可以继续考虑优化的地方：

1. Plot 区间划分
2. 看一下用信息增益或者信息增益率的效果
3. 把报告写了

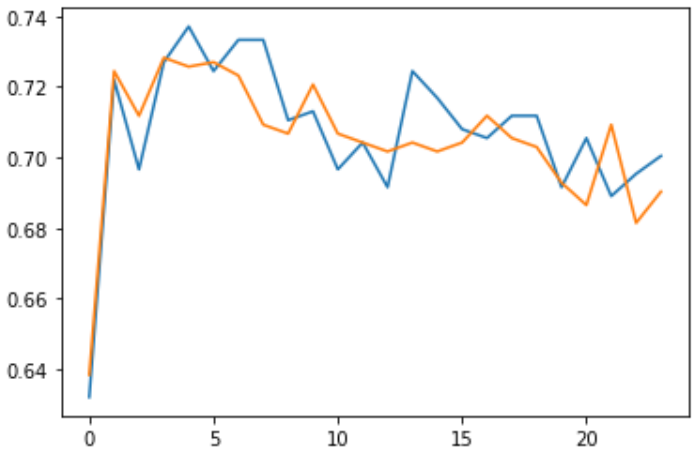
11月7日

区间划分数量对val上accuracy的影响：基于最大深度为2的测试（过拟合可能性更小）

midclass: 1
method: 1 accuracy:0.6321
method 2 accuracy:0.6384
midclass: 2
method: 1 accuracy:0.7219
method 2 accuracy:0.7244
midclass: 3
method: 1 accuracy:0.6966
method 2 accuracy:0.7118
midclass: 4
method: 1 accuracy:0.7269
method 2 accuracy:0.7282
midclass: 5
method: 1 accuracy:0.7370
method 2 accuracy:0.7257
midclass: 6
method: 1 accuracy:0.7244
method 2 accuracy:0.7269
midclass: 7
method: 1 accuracy:0.7332
method 2 accuracy:0.7231
midclass: 8
method: 1 accuracy:0.7332
method 2 accuracy:0.7092
midclass: 9
method: 1 accuracy:0.7105
method 2 accuracy:0.7067
midclass: 10
method: 1 accuracy:0.7130
method 2 accuracy:0.7206
midclass: 11
method: 1 accuracy:0.6966
method 2 accuracy:0.7067
midclass: 12
method: 1 accuracy:0.7042
method 2 accuracy:0.7042
midclass: 13
method: 1 accuracy:0.6915
method 2 accuracy:0.7016
midclass: 14

method: 1 accuracy:0.7244
method 2 accuracy:0.7042
midclass: 15
method: 1 accuracy:0.7168
method 2 accuracy:0.7016
midclass: 16
method: 1 accuracy:0.7080
method 2 accuracy:0.7042
midclass: 17
method: 1 accuracy:0.7054
method 2 accuracy:0.7118
midclass: 18
method: 1 accuracy:0.7118
method 2 accuracy:0.7054
midclass: 19
method: 1 accuracy:0.7118
method 2 accuracy:0.7029
midclass: 20
method: 1 accuracy:0.6915
method 2 accuracy:0.6928
midclass: 21
method: 1 accuracy:0.7054
method 2 accuracy:0.6865
midclass: 22
method: 1 accuracy:0.6890
method 2 accuracy:0.7092
midclass: 23
method: 1 accuracy:0.6953
method 2 accuracy:0.6814
midclass: 24
method: 1 accuracy:0.7004
method 2 accuracy:0.6903

（图片中x轴表示除了两个极值外的分界线数量k，实际组数为其+3）k = midclass -1
（图片中y轴表示accuracy, 蓝线等区间划分，橙线等比划分）



分析：
在验证集（791个样本）95%置信度的区间约为Accuracy+-0.03。
k = 0时，accuracy较低。
k > 1时，对着区间划分数量增多，accuracy呈下降趋势，但是并不显著。
但考虑到当区间大于10个的两层树，平均每个子叶的样本数量少于1%，可能会过拟合。
所以最终选择 k = 6，（midclass = 7), 总区间划分数为9。
结论：置信度为95%的置信区间为0.7267±0.0196。