基于Titanic数据集的决策树算法实践

王铭嵩

2023 年 10 月 23 日

1 数据分析与处理

数据来源为Kaggle的The Complete Titanic Dataset数据集。源文件为一个包含舱位、年龄、性别、是否幸存等信息的csv文件，共有1309条数据。

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2	<null>	St Louis, MO
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	<null>	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.5500	C22 C26	S	<null>	<null>	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.5500	C22 C26	S	<null>	135	Montreal, PQ / Chesterville, ON
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Dani…	female	25	1	2	113781	151.5500	C22 C26	S	<null>	<null>	Montreal, PQ / Chesterville, ON
6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500	E12	S	3	<null>	New York, NY
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583	D7	S	10	<null>	Hudson, NY
8	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0.0000	A36	S	<null>	<null>	Belfast, NI
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamso…	female	53	2	0	11769	51.4792	C101	S	D	<null>	Bayside, Queens, NY
10	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042	<null>	C	<null>	22	Montevideo, Uruguay
11	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.5250	C62 C64	C	<null>	124	New York, NY
12	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge …	female	18	1	0	PC 17757	227.5250	C62 C64	C	4	<null>	New York, NY
13	1	1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69.3000	B35	C	9	<null>	Paris, France
14	1	1	Barber, Miss. Ellen "Nellie"	female	26	0	0	19877	78.8500	<null>	S	6	<null>	<null>
15	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0000	A23	S	B	<null>	Hessle, Yorks
16	1	0	Baumann, Mr. John D	male	<null>	0	0	PC 17318	25.9250	<null>	S	<null>	<null>	New York, NY
17	1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247.5208	B58 B60	C	<null>	<null>	Montreal, PQ
18	1	1	Baxter, Mrs. James (Helene DeLaudeniere Cha…	female	50	0	1	PC 17558	247.5208	B58 B60	C	6	<null>	Montreal, PQ
19	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917	D15	C	8	<null>	<null>
20	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6	C	A	<null>	Winnipeg, MN
21	1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542	D35	S	5	<null>	New York, NY
22	1	1	Beckwith, Mrs. Richard Leonard (Sallie Mony…	female	47	1	1	11751	52.5542	D35	S	5	<null>	New York, NY
23	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30.0000	C148	C	5	<null>	New York, NY
24	1	1	Bidois, Miss. Rosalie	female	42	0	0	PC 17757	227.5250	<null>	C	4	<null>	<null>
25	1	1	Bird, Miss. Ellen	female	29	0	0	PC 17483	221.7792	C97	S	8	<null>	<null>
26	1	0	Birnbaum, Mr. Jakob	male	25	0	0	13905	26.0000	<null>	C	<null>	148	San Francisco, CA
27	1	1	Bishop, Mr. Dickinson H	male	25	1	0	11967	91.0792	B49	C	7	<null>	Dowagiac, MI
28	1	1	Bishop, Mrs. Dickinson H (Helen Walton)	female	19	1	0	11967	91.0792	B49	C	7	<null>	Dowagiac, MI
29	1	1	Bissette, Miss. Amelia	female	35	0	0	PC 17760	135.6333	C99	S	8	<null>	<null>
30	1	1	Bjornstrom-Steffansson, Mr. Mauritz Hakan	male	28	0	0	110564	26.5500	C52	S	D	<null>	Stockholm, Sweden / Washington, DC
31	1	0	Blackwell, Mr. Stephen Weart	male	45	0	0	113784	35.5000	T	S	<null>	<null>	Trenton, NJ
32	1	1	Blank, Mr. Henry	male	40	0	0	112277	31.0000	A31	C	7	<null>	Glen Ridge, NJ
33	1	1	Bonnell, Miss. Caroline	female	30	0	0	36928	164.8667	C7	S	8	<null>	Youngstown, OH
34	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.5500	C103	S	8	<null>	Birkdale, England Cleveland, Ohio
35	1	0	Borebank, Mr. John James	male	42	0	0	110489	26.5500	D22	S	<null>	<null>	London / Winnipeg, MB
36	1	1	Bowen, Miss. Grace Scott	female	45	0	0	PC 17608	262.3750	<null>	C	4	<null>	Cooperstown, NY
37	1	1	Bowerman, Miss. Elsie Edith	female	22	0	1	113505	55.0000	E33	S	6	<null>	St Leonards-on-Sea, England Ohio
38	1	1	Bradley, Mr. George ("George Arthur Brayton…	male	<null>	0	0	111427	26.5500	<null>	S	9	<null>	Los Angeles, CA
39	1	0	Brady, Mr. John Bertram	male	41	0	0	113054	30.5000	A21	S	<null>	<null>	Pomeroy, WA
40	1	0	Brandeis, Mr. Emil	male	48	0	0	PC 17591	50.4958	B10	C	<null>	208	Omaha, NE
41	1	0	Brewe, Dr. Arthur Jackson	male	<null>	0	0	112379	39.6000	<null>	C	<null>	<null>	Philadelphia, PA
42	1	1	Brown, Mrs. James Joseph (Margaret Tobin)	female	44	0	0	PC 17610	27.7208	B4	C	6	<null>	Denver, CO
43	1	1	Brown, Mrs. John Murray (Caroline Lane Lams…	female	59	2	0	11769	51.4792	C101	S	D	<null>	Belmont, MA
44	1	1	Bucknell, Mrs. William Robert (Emma Eliza W…	female	60	0	0	11813	76.2917	D15	C	8	<null>	Philadelphia, PA
45	1	1	Burns, Miss. Elizabeth Margaret	female	41	0	0	16966	134.5000	E40	C	3	<null>	<null>
46	1	0	Butt, Major. Archibald Willingham	male	45	0	0	113050	26.5500	B38	S	<null>	<null>	Washington, DC

读入数据后清除了年龄、配偶兄弟姐妹数目、父母子女数目信息缺失的条目，并选择舱位、性别、年龄、配偶兄弟姐妹数目、父母子女数目作为特征，是否幸存作为预测目标。

```
# data loading
titanic_data=pd.read_csv('data/titanic3.csv')


# data cleaning
titanic_data=titanic_data.dropna(subset=["age","sibsp","parch"])


# select relevant features and target
X=titanic_data[["pclass", "sex", "age", "sibsp", "parch"]]
Y=titanic_data["survived"]
```

2 决策树原理与代码

决策树通过递进的属性判断对样本进行划分，并采用一定的度量标准来决定每一次划分所基于的属性。度量标准一般包括信息增益、信息增益率以及基尼指数，对应ID3决策树，C4.5决策树和CART决策树算法。

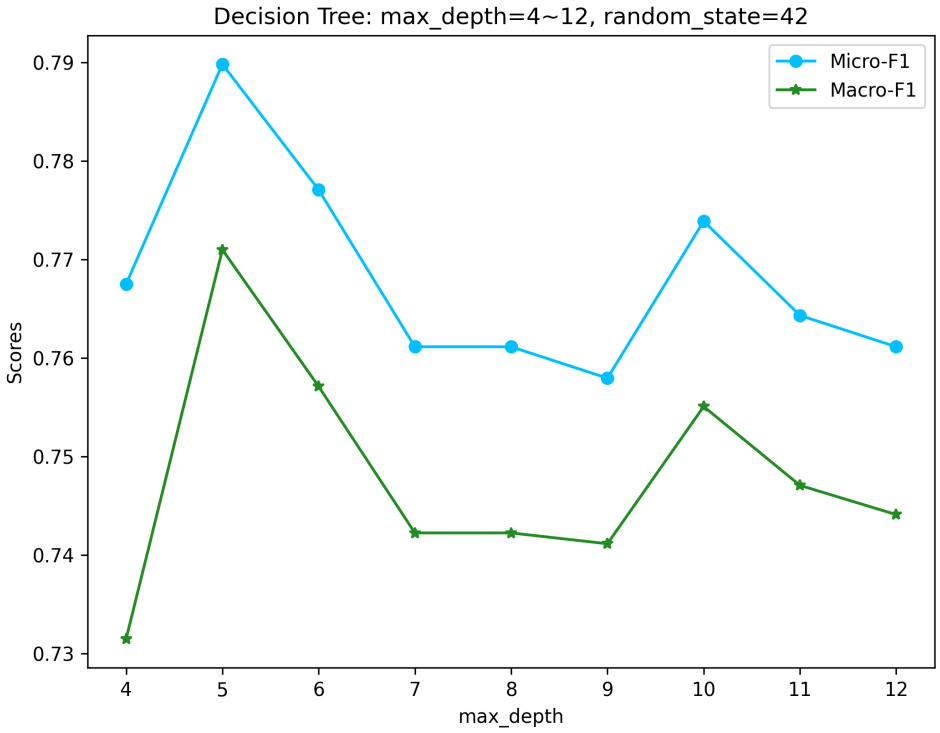
叶节点采用递归方式生成，达到最大深度或者样本数目小于阈值时停止递归。

代码如下，采用scikit-learn库实现。用信息增益作为纯度指标。

```
'''decision tree implementation'''
# creating classifier
classifier=DecisionTreeClassifier(criterion="entropy",random_state=random_state,max_depth=max_depth)
# fitting data
classifier.fit(X_train,Y_train)
print('Done generating DTree. \n')
# predicting
y_pred=classifier.predict(X_test)
```

3 验证集评估结果

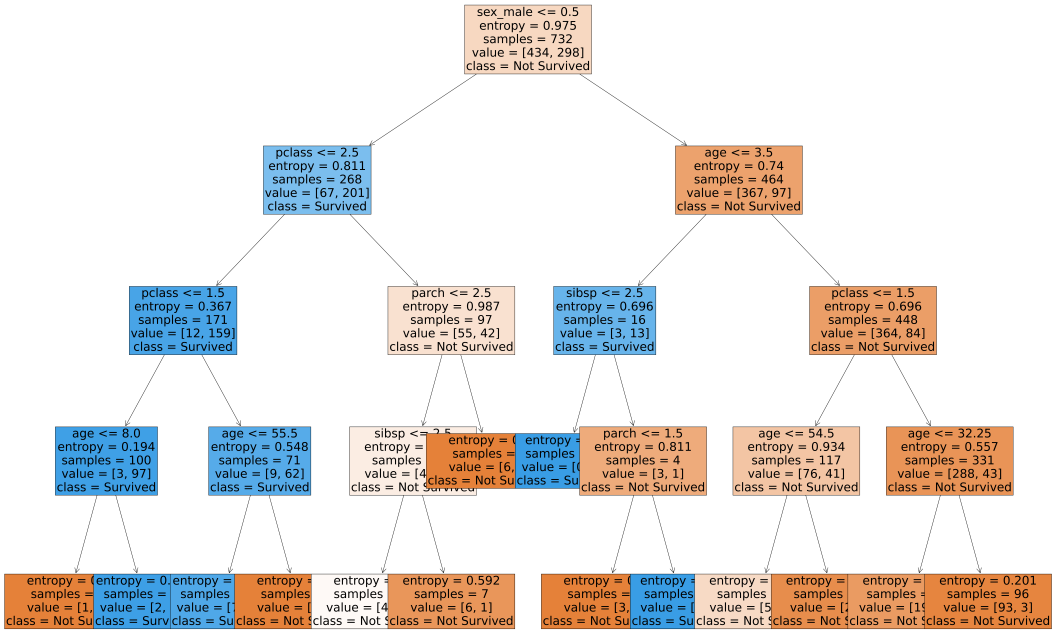
分别在最大树高度4到12的情况下对决策树进行评估。验证集为F1-Micro和F1-Macro。以下为各最大树高条件下的各验证集测量值。

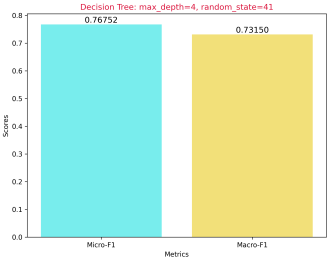


4 决策树可视化

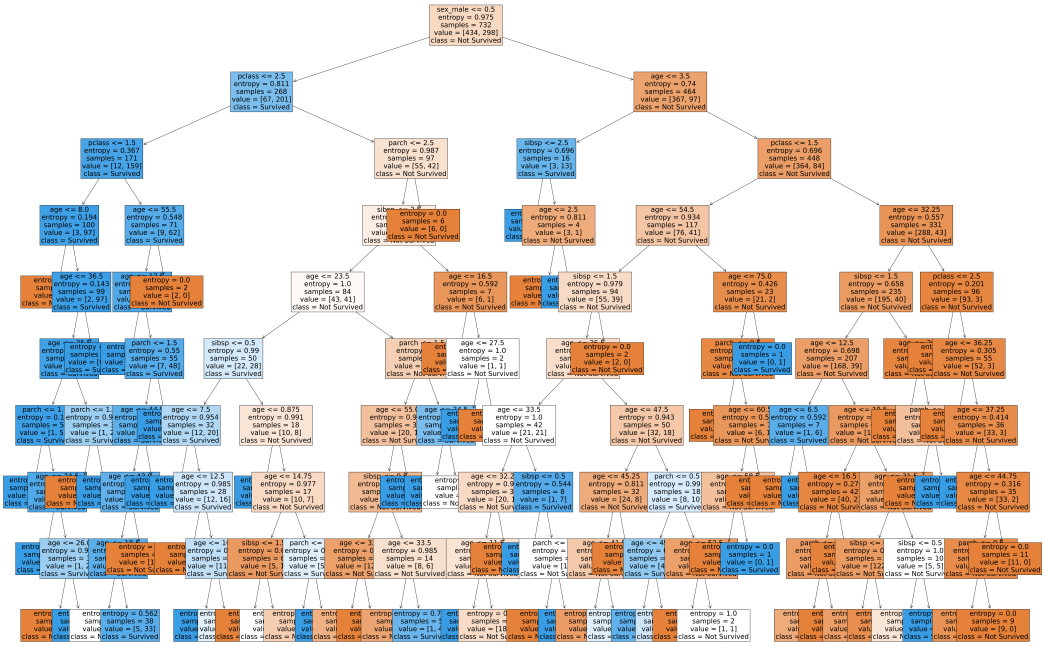
仅显示部分最大树高度下的可视化。完整可视化图片可在../python/result目录下查看。

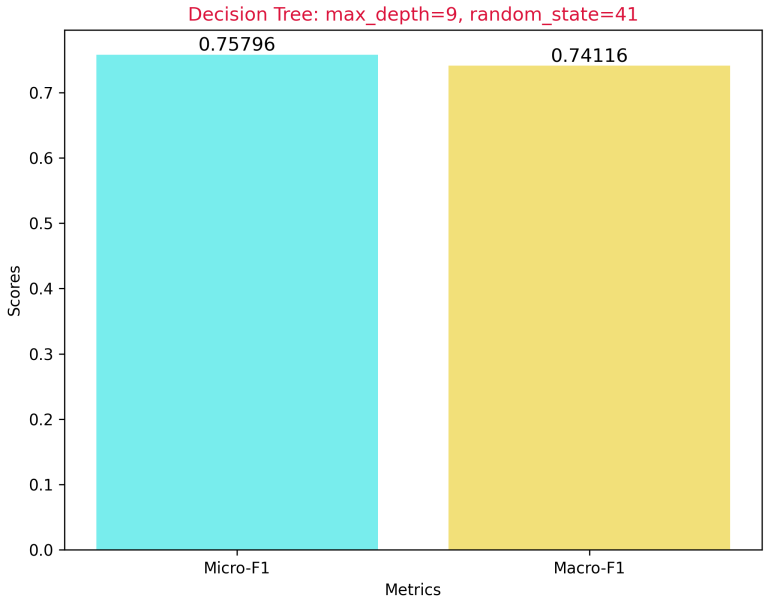
4.1 最大树高度为4



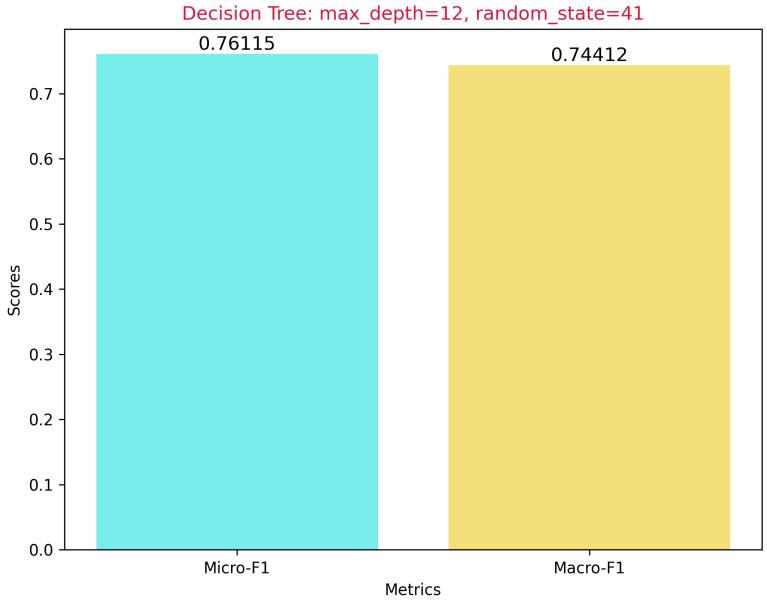


4.2 最大树高度为9


sex_male <= 0.5
entropy = 0.975
samples = 732
value = [434, 298]
class = Not Survived
pclass <= 2.5
entropy = 0.811
samples = 268
value = [67, 201]
class = Survived
age <= 3.5
entropy = 0.74
samples = 464
value = [367, 97]
class = Not Survived


Decision Tree: max_depth=9, random_state=41
0.75796
0.74116
Scores
Micro-F1
Macro-F1
Metrics

4.3 最大树高度为12


Decision Tree: max_depth=12, random_state=41
0.76115
0.74412
Scores
Metrics
Micro-F1
Macro-F1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7