2023年4月1日 星期六

Titanic Survivor's Prediction

Decision Tree

- DT的參數特性：
  - random_state & splitter 增加整體隨機性
  - max_depth: 限制 tree 的最大高度，可以避免 overfitting
  - min_samples_leaf & min_samples_split 必須包含一定數量的訓練樣本才會分支
  - max_feature & min_impurity_decrease 作為調完 max_depth 後的精修
- Data preprocessing:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

- Operation:

```
#資料預處理

#刪除缺失過多和不重要的列
data.drop(["Cabin","Name","Ticket"],inplace=True,axis=1)

#Age補缺失值
data["Age"] = data["Age"].fillna(data["Age"].mean())
data = data.dropna()

#將 True / False 轉成數值變量
data["Sex"] = (data["Sex"]=="male").astype("int")

#轉成數值變量
labels = data["Embarked"].unique().tolist()
data["Embarked"] = data["Embarked"].apply(lambda x: labels.index(x)) #分類變量轉成index變量

#資料預處理結果
data.head()
```

- Result:

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0	3	1	22.0	1	0	7.2500	0
2	1	1	0	38.0	1	0	71.2833	1
3	1	3	0	26.0	0	0	7.9250	0
4	1	1	0	35.0	1	0	53.1000	0
5	0	3	1	35.0	0	0	8.0500	0

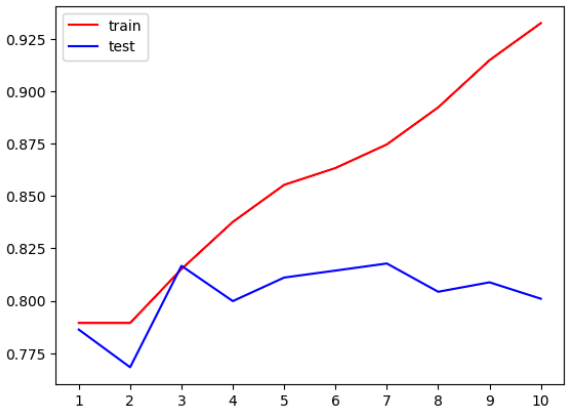
- Parameter:
  - 利用學習曲線找出最佳的 max_depth 參數 3

```python
#利用學習曲線調max_depth參數

tr = []
te = []
for i in range(10):
    clf = DecisionTreeClassifier(random_state=25
                                 ,max_depth=i+1
                                 ,criterion="entropy")
    clf = clf.fit(Xtrain,Ytrain)
    score_tr = clf.score(Xtrain,Ytrain)
    score_te = cross_val_score(clf,X,y,cv=10).mean()
    tr.append(score_tr)
    te.append(score_te)

print(max(te))
plt.plot(range(1,11),tr,color="red",label="train")
plt.plot(range(1,11),te,color="blue",label="test")
plt.xticks(range(1,11))
plt.legend()
plt.show()

#在 max_depth = 3 時，測試集和訓練集最接近
```



  - 利用網格搜索調參

```python
#利用網格搜索調參數
import numpy as np
gini_thresholds = np.linspace(0,0.5,20)

parameters = {'splitter':('best','random')
              ,'criterion':("gini","entropy")
              ,"max_depth":[*range(1,10)]
              ,"min_samples_leaf":[*range(1,20,5)]
              ,"min_impurity_decrease":[*np.linspace(0,0.5,20)]}

clf = DecisionTreeClassifier(random_state=25)
GS = GridSearchCV(clf,parameters,cv=10)
GS.fit(Xtrain,Ytrain)

GS.best_params_
```

- Result:
  - 學習曲線 Score : 0.817786
  - 網格搜索 Score : 0.824756