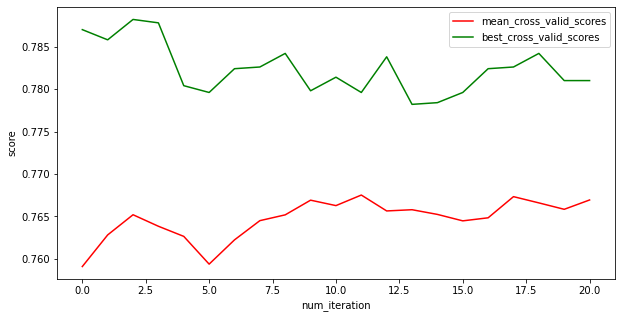
**Ex4 Report**

自硕21 崔晏菲 2021210976

**1. 特征提取**

我选择了遗传算法用于特征提取。因为找不到现成的包，所以自己手写的。分类器我选择的是logistic分类器，为了方便对比，我之后几题也都用的logistic分类器。

因为遗传算法时间复杂度很大，所以我只进行了20轮迭代，可能远远没有迭代到稳态，迭代后的结果如下：



可以看到，在迭代的过程中，平均的交叉验证得分在稳步上升，说明遗传算法确实起到了效果，但是每一代最优的得分都不如一开始，这是显而易见的，因为一开始有的数据根本没有进行特征全责，将全部特征都用于logistic分类，所以达不到是很正常的。最后遗传算法选取的特征为

age

bmi

elective\_surgery

pre\_icu\_los\_days

arf\_apache

bun\_apache

creatinine\_apache

gcs\_verbal\_apache

glucose\_apache

heart\_rate\_apache

map\_apache

sodium\_apache

ventilated\_apache

wbc\_apache

d1\_diasbp\_max

d1\_heartrate\_min

d1\_mbp\_max

d1\_spo2\_max

d1\_sysbp\_max

d1\_temp\_max

d1\_temp\_min

h1\_diasbp\_min

h1\_mbp\_max

h1\_mbp\_min

h1\_resprate\_max

h1\_spo2\_max

h1\_spo2\_min

h1\_sysbp\_max

h1\_sysbp\_min

d1\_calcium\_max

d1\_glucose\_min

d1\_hemaglobin\_max

d1\_platelets\_max

d1\_potassium\_min

d1\_sodium\_max

apache\_4a\_icu\_death\_prob

aids

cirrhosis

hepatic\_failure

immunosuppression

leukemia

ethnicity\_African American

ethnicity\_Other/Unknown

icu\_type\_CCU-CTICU

icu\_type\_CSICU

icu\_type\_MICU

icu\_type\_SICU

apache\_3j\_bodysystem\_Cardiovascular

apache\_3j\_bodysystem\_Gastrointestinal

apache\_3j\_bodysystem\_Gynecological

apache\_3j\_bodysystem\_Hematological

apache\_3j\_bodysystem\_Metabolic

apache\_3j\_bodysystem\_Musculoskeletal/Skin

apache\_3j\_bodysystem\_Neurological

apache\_3j\_bodysystem\_Respiratory

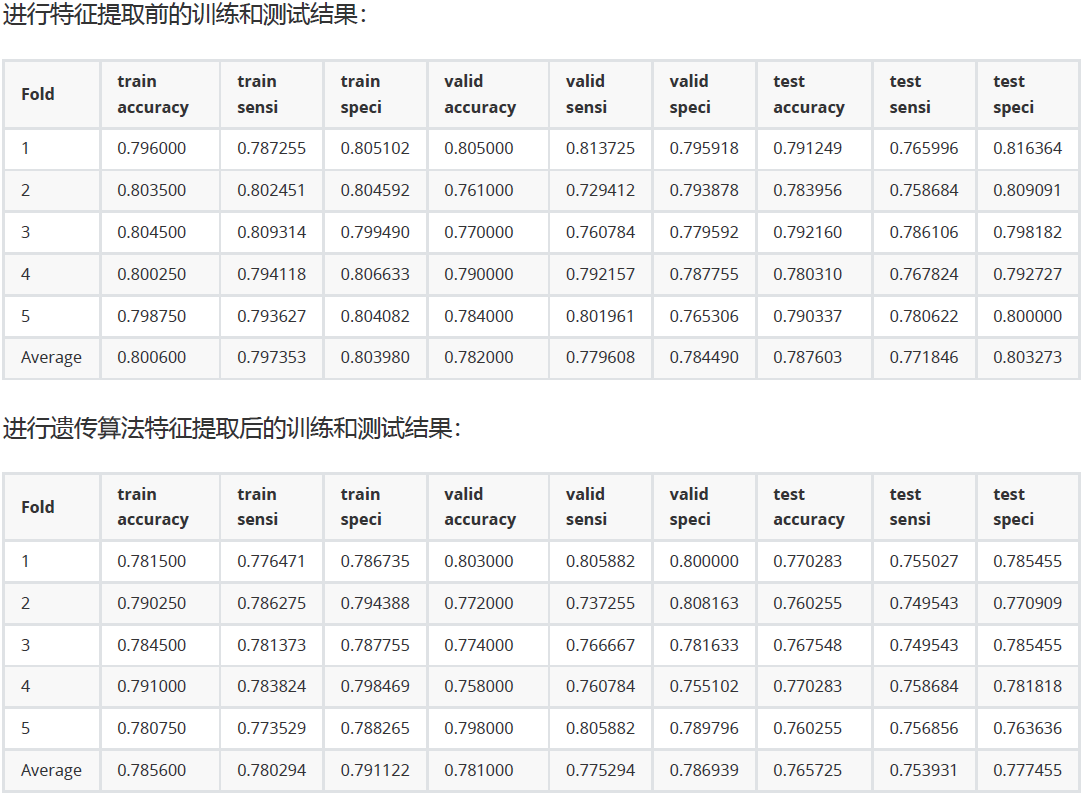
apache\_2\_bodysystem\_Cardiovascular

apache\_2\_bodysystem\_Gastrointestinal

apache\_2\_bodysystem\_Respiratory

apache\_2\_bodysystem\_Trauma

特征提取前后在TestSet-1上的结果为：



可以看到特征提取后训练集表现略微下降，这是可以理解的，但是下降幅度不大。

**2. Lasso特征提取**

Lasso回归提取到的特征为：

age

elective\_surgery

height

pre\_icu\_los\_days

weight

apache\_2\_diagnosis

apache\_3j\_diagnosis

arf\_apache

bun\_apache

creatinine\_apache

gcs\_eyes\_apache

gcs\_motor\_apache

gcs\_verbal\_apache

glucose\_apache

hematocrit\_apache

intubated\_apache

map\_apache

resprate\_apache

sodium\_apache

ventilated\_apache

wbc\_apache

d1\_heartrate\_max

d1\_heartrate\_min

d1\_mbp\_max

d1\_mbp\_min

d1\_resprate\_max

d1\_spo2\_min

d1\_sysbp\_max

d1\_sysbp\_min

d1\_temp\_max

d1\_temp\_min

h1\_diasbp\_max

h1\_heartrate\_max

h1\_heartrate\_min

h1\_mbp\_min

h1\_resprate\_max

h1\_resprate\_min

h1\_spo2\_max

h1\_spo2\_min

h1\_sysbp\_min

h1\_temp\_max

d1\_calcium\_max

d1\_calcium\_min

d1\_glucose\_min

d1\_hco3\_max

d1\_hco3\_min

d1\_hemaglobin\_max

d1\_platelets\_max

d1\_potassium\_max

d1\_potassium\_min

apache\_4a\_icu\_death\_prob

aids

cirrhosis

diabetes\_mellitus

hepatic\_failure

immunosuppression

lymphoma

solid\_tumor\_with\_metastasis

ethnicity\_African American

ethnicity\_Hispanic

gender\_F

icu\_type\_CCU-CTICU

icu\_type\_CSICU

icu\_type\_CTICU

icu\_type\_Cardiac ICU

icu\_type\_Med-Surg ICU

icu\_type\_Neuro ICU

apache\_3j\_bodysystem\_Cardiovascular

apache\_3j\_bodysystem\_Gastrointestinal

apache\_3j\_bodysystem\_Metabolic

apache\_3j\_bodysystem\_Musculoskeletal/Skin

apache\_3j\_bodysystem\_Neurological

apache\_3j\_bodysystem\_Respiratory

apache\_2\_bodysystem\_Gastrointestinal

apache\_2\_bodysystem\_Metabolic

apache\_2\_bodysystem\_Neurologic

apache\_2\_bodysystem\_Renal/Genitourinary

apache\_2\_bodysystem\_Respiratory

apache\_2\_bodysystem\_Undefined Diagnoses

apache\_2\_bodysystem\_Undefined diagnoses

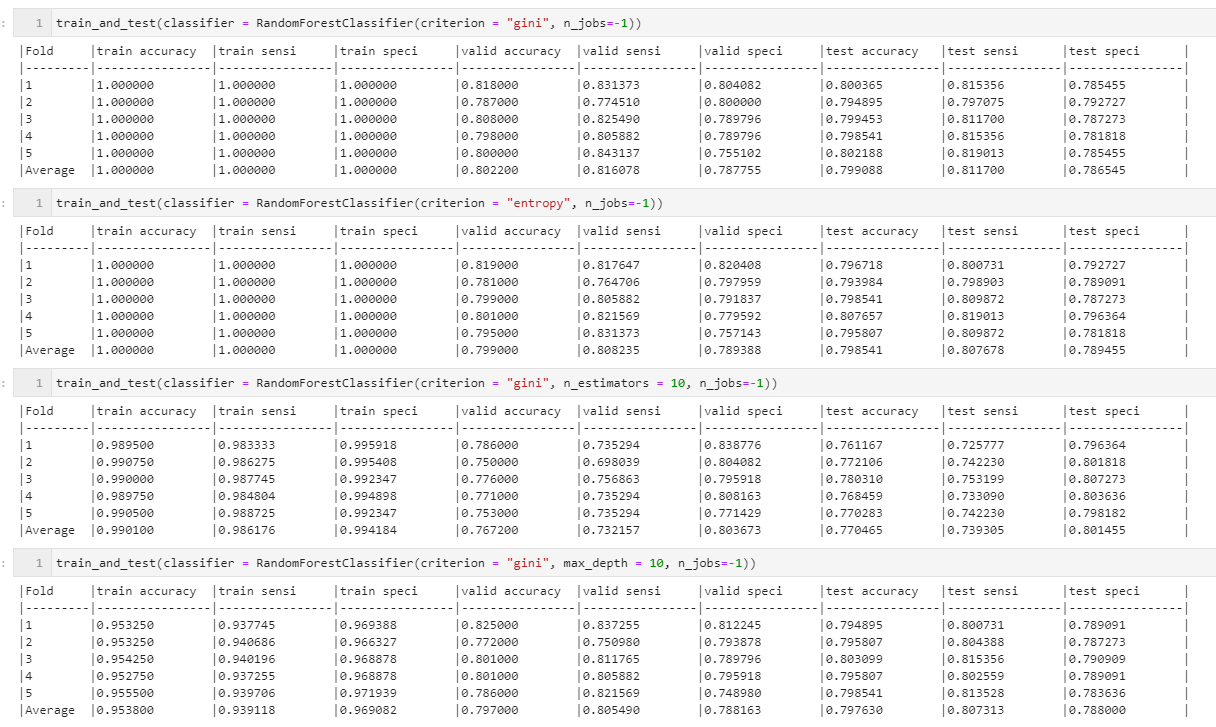
可以看到和遗传算法接近并且有一些差别。



可以看到特征提取后，logistic回归结果略好于遗传算法，低于特征提取前，这是可以理解的，因为我的遗传算法并没有迭代足够的轮次。

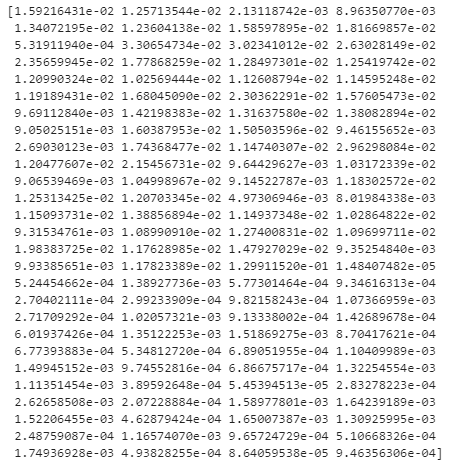
**3. 随机森林**

我主要比较了sklearn随机森林criterion、n\_estimator、和max\_depth这三个参数。结果对比如图：



可以看到，”gini”是优于”entropy”的；随机森林的“树木”的数量会影响到分类的结果，并且数量越多，分类结果越好；并且每个决策树的深度限制也会影响分类性能，降低深度，可以减少过拟合，但是训练集上表现也会稍差。

接着，我使用默认参数的随机森林，对各个特征的重要程度进行了打分，得到的结果如下：



观察得到，重要性在1e-4这个量级的特征相对不是很重要，因此以此为界，我得到了特征选择的结果如下：

age

bmi

elective\_surgery

height

pre\_icu\_los\_days

weight

apache\_2\_diagnosis

apache\_3j\_diagnosis

bun\_apache

creatinine\_apache

gcs\_eyes\_apache

gcs\_motor\_apache

gcs\_verbal\_apache

glucose\_apache

heart\_rate\_apache

hematocrit\_apache

intubated\_apache

map\_apache

resprate\_apache

sodium\_apache

temp\_apache

ventilated\_apache

wbc\_apache

d1\_diasbp\_max

d1\_diasbp\_min

d1\_heartrate\_max

d1\_heartrate\_min

d1\_mbp\_max

d1\_mbp\_min

d1\_resprate\_max

d1\_resprate\_min

d1\_spo2\_max

d1\_spo2\_min

d1\_sysbp\_max

d1\_sysbp\_min

d1\_temp\_max

d1\_temp\_min

h1\_diasbp\_max

h1\_diasbp\_min

h1\_heartrate\_max

h1\_heartrate\_min

h1\_mbp\_max

h1\_mbp\_min

h1\_resprate\_max

h1\_resprate\_min

h1\_spo2\_max

h1\_spo2\_min

h1\_sysbp\_max

h1\_sysbp\_min

h1\_temp\_max

h1\_temp\_min

d1\_calcium\_max

d1\_calcium\_min

d1\_glucose\_min

d1\_hco3\_max

d1\_hco3\_min

d1\_hemaglobin\_max

d1\_platelets\_max

d1\_potassium\_max

d1\_potassium\_min

d1\_sodium\_max

apache\_4a\_icu\_death\_prob

diabetes\_mellitus

ethnicity\_African American

ethnicity\_Caucasian

gender\_F

gender\_M

icu\_type\_MICU

icu\_type\_Med-Surg ICU

apache\_3j\_bodysystem\_Cardiovascular

apache\_3j\_bodysystem\_Gastrointestinal

apache\_3j\_bodysystem\_Metabolic

apache\_3j\_bodysystem\_Neurological

apache\_3j\_bodysystem\_Respiratory

apache\_3j\_bodysystem\_Sepsis

apache\_2\_bodysystem\_Cardiovascular

apache\_2\_bodysystem\_Gastrointestinal

apache\_2\_bodysystem\_Metabolic

apache\_2\_bodysystem\_Respiratory

接着我将这个特征选择结果用于logistic分类器，发现结果较好，和lasso回归的表现类似，都优于遗传算法，主要是因为遗传算法迭代次数不够多。