人工智能导论 分类实践

周雨豪 2018013399 软件92

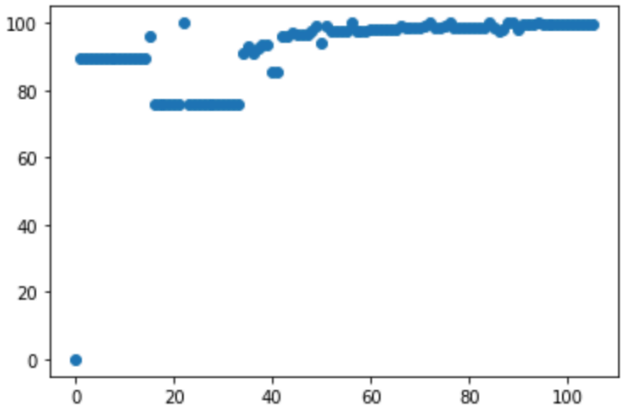
简介

根据患者数据集预测诊断结果（包含 4 项指标）。流程大致为数据读取、处理 - 模型选择、训练 - 模型评估。

包含两个 Jupyter Notebook 文件，`classification` 预测指标 **SARS-Cov-2 exam result**，训练和测试集为随机 2 : 1 划分；`Task2` 预测指标 **Patient addmited to regular ward**，测试集为编号 5001-5645 的样本。除此以外两项任务流程一致。

数据处理

处理数据前根据观察可知数据大范围为缺失值（大部分缺失 95% 以上），106 项属性的缺失比例如图所示



用常数填充缺失值（const = -999），之后的降维操作会淹没这些缺失值的影响。

特征工程使用主成分分析（PCA）对输入特征进行降维，其原理是先将样本归一化处理，计算协方差矩阵，然后对协方差矩阵做特征值分解，取最大的前 n 个特征值对应的特征向量为新的特征空间，调用 `sklearn.decomposition.PCA` 方法实现，将特征维数降低至 20。

模型

选择的三种模型为 random forest, Linear SVC 和 KNN。

对每种模型分别使用两个版本，一个为默认参数的模型，另一个为使用 k 折交叉验证（k=10）调参选择的模型（调用 `sklearn.model_selection.GridSearchCV`，评判指标为 accuracy）。参数及调参结果如下

Random Forest

```
params = {
    'max_depth': [10, 50, None],
    'min_samples_split': [2, 4],
    'min_samples_leaf': [1, 2],
    'bootstrap': [True, False] }
```

> **best params:** max_depth = 10, min_samples_split = 2, min_samples_leaf = 2, bootstrap = True

SVC

```
params = {
    'penalty': ['l1', 'l2'],
    'loss': ['hinge', 'squared_hinge'],
    'dual': [True, False],
    'max_iter': [100, 1000, None] }
```

> **best params:** penalty = l1, loss = squared_hinge, dual = False, max_iter = 100

KNN

```
params = {
    'n_neighbors': [2, 5, 10, 50],
    'leaf_size': [10, 30, 100] }
```

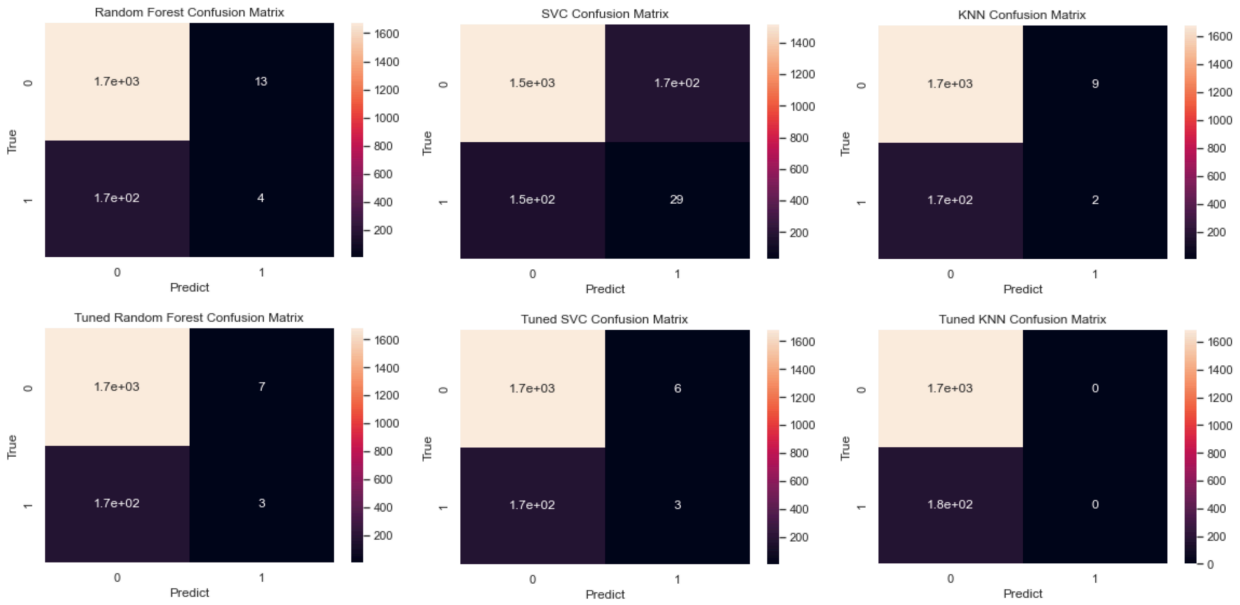
> **best params:** n_neighbors = 50, leaf_size = 10

评估

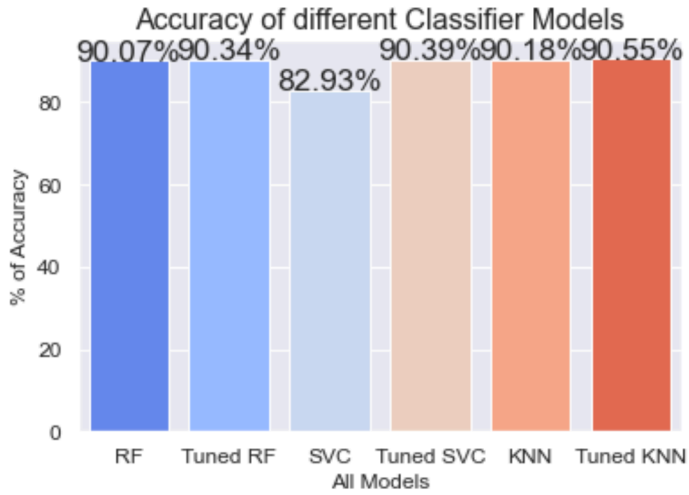
模型训练完后调用 `model_assess()` 评估其性能，性能评估包括 accuracy，precision，f1_score，roc_auc_score 和混淆矩阵等等，以任务`classification` 中的结果为例（详细数据见 ipynb 或 html 文件）

	Accuracy	ROC AUC
RF	90.07%	0.596
Tuned RF	90.34%	0.603
SVC	82.93%	-
Tuned SVC	90.39%	-
KNN	90.18%	0.515
Tuned KNN	90.55%	0.599

混淆矩阵如下



Accuracy 对比如下



根据 Accuracy 指标，使用`GridSearchCV`调参的模型预测结果均优于默认参数的模型，三种模型间对比 KNN 结果最好而默认的 SVC 结果最差，运行时发现 RF 调参的时间开销较高，远大于其他两种算法，原因应该在于需要训练大量子决策树。