实验平台：python为编程语言，sklearn为分类工具，pytorch为神经网络分类工具，matplotlib和graphviz为可视化工具。

预处理：最大最小归一化

一、评价指标

Precision recall accuracy f1\_metric

F1是综合指标，各实验以F1作为对比标准

ROC曲线，面积越大性能越好

损失曲线，下降得越低越平稳越好

二、各方法对比

对主流的机器学习方法进行对比，评价指标为F1，控制变量为测试集占数据集的比例。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 30% | 50% | 70% | 90% |
| DecisionTree | 0.960 | 0.941 | 0.928 | 0.950 |
| GaussianNB | 0.966 | 0.974 | 0.960 | 0.936 |
| MultinomialNB | 0.880 | 0.878 | 0.829 | 0.732 |
| BernoulliNB | 0.982 | 0.984 | 0.977 | 0.975 |
| RandomForest | 0.995 | 0.997 | 0.981 | 0.979 |
| MLP | 0.978 | 0.976 | 0.966 | 0.957 |
| SVM | 0.978 | 0.961 | 0.9921 | 0.795 |
| GradientBoosting | 0.991 | 0.981 | 0.975 | 0.963 |

|  |  |
| --- | --- |
|  |  |
| 30% | 50% |
|  |  |
| 70% | 90% |

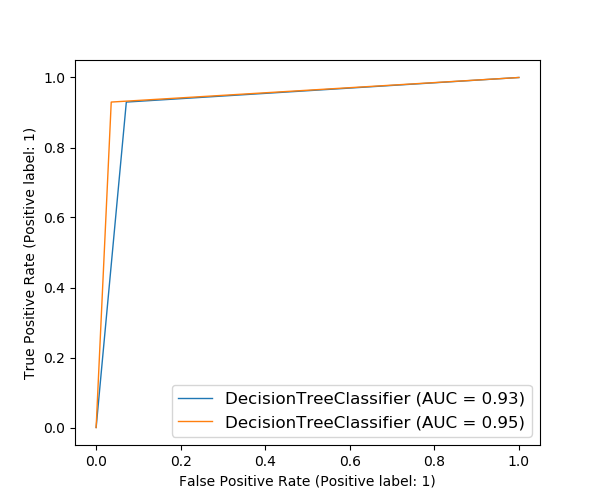
其中，决策树、神经网络、SVM仍然具有较大的提升空间

二、决策树

1、决策树采用了最佳的CART方法，在该基础上对比了gini的方法和entropy的方法

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 metric |
| Gini | 0.929 | 0.946 | 0.929 | 0.938 |
| Entropy | 0.944 | 0.972 | 0.929 | 0.956 |

两种方法ROC曲线对比

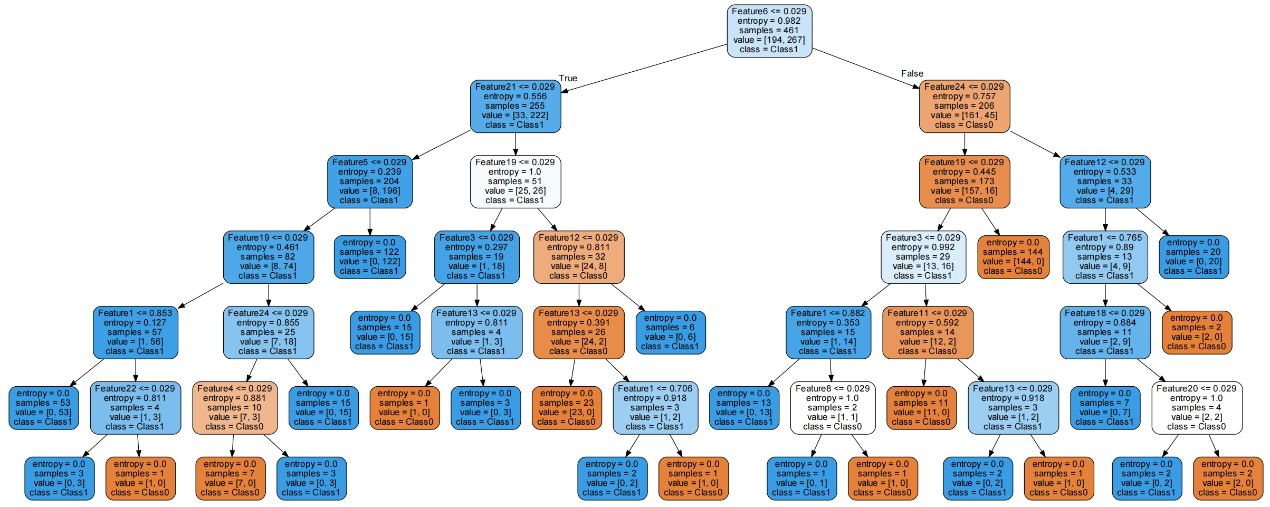


黄色为entropy，蓝色为Gini。可见entropy方法更优

2、特征重要性

|  |  |  |
| --- | --- | --- |
|  | Gini | Entropy |
| 性别 | 0.0% | 0.0% |
| GerdQ | 3.1% | 3.1% |
| 反酸 | 0.0% | 0.0% |
| 烧心 | 5.3% | 3.9% |
| 呃逆 | 2.2% | 1.9% |
| 嗳气 | 0.2% | 2.4% |
| 口苦 | 43.1% | 34.2% |
| 胃胀满 | 1.5% | 0.0% |
| 心烦易怒 | 0.0% | 0.4% |
| 胁肋胀满 | 0.8% | 0.0% |
| 胸骨后灼痛 | 0.0% | 0.0% |
| 口干 | 0.9% | 1.2% |
| 咽干 | 0.4% | 4.8% |
| 食少纳呆 | 0.7% | 3.0% |
| 寐差 | 0.0% | 0.0% |
| 神疲乏力 | 0.9% | 0.0% |
| 大便溏薄 | 0.0% | 0.0% |
| 大便秘结 | 0.0% | 0.0% |
| 舌淡红 | 0.0% | 0.8% |
| 舌红 | 12.4% | 16.9% |
| 舌齿痕 | 0.4% | 0.9% |
| 舌苔薄黄 | 7.4% | 9.3% |
| 舌苔黄腻 | 0.0% | 0.7% |
| 脉弦 | 0.0% | 0.0% |
| 脉滑 | 20.7% | 16.4% |
| 脉数 | 0.0% | 0.0% |
| 脉缓 | 0.0% | 0.0% |

3、entropy的树



（更清晰的见pdf，gini的见pdf）

二、神经网络

固定迭代最大迭代数均为200，批次大小均为32

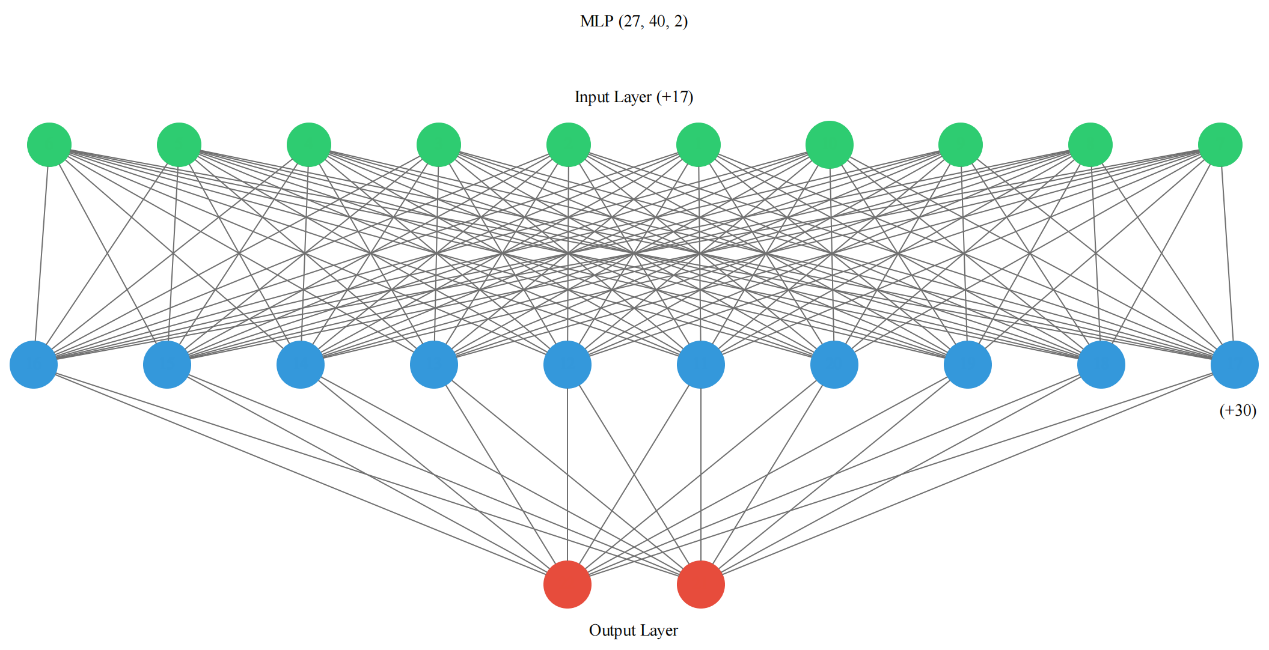
1、多层感知机，各层神经元数为（27, 27, 2）的情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 激活函数 | Acc | P | R | F1 |
| Tanh | 0.969 | 0.965 | 0.982 | 0.973 |
| Logistic | 0.924 | 0.889 | 0.991 | 0.937 |
| ReLU | 0.969 | 0.965 | 0.982 | 0.974 |

可见，采用ReLU综合表现更佳。下面均为ReLU的情况下进行实验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 各层神经元数 | Acc | P | R | F1 |
| （27, 27, 2） | 0.969 | 0.965 | 0.982 | 0.974 |
| （27, 40, 2） | 0.979 | 0.974 | 0.991 | 0.982 |
| （27, 70, 2） | 0.964 | 0.965 | 0.973 | 0.969 |
| （27, 100, 2） | 0.979 | 0.974 | 0.991 | 0.982 |
| （27, 32, 16， 8， 32，2） | 0.984 | 0.991 | 0.982 | 0.986 |

其中，神经网络的结构以(27,40,2)为例，可视化如下



可见，增加隐藏层的神经元数可以提高性能，而且增加其层数可以进一步提升分类效果。现在对（27, 32, 16， 8， 32，2）的神经网络增加其他更改，对比实验结果如下。同时，对比了卷积神经网络的方法，具体采用restnet34结构，并将其二维卷积和二维池化结构更改为一维卷积和一维池化，以适应本文数据。

由于神经元数极大增加，为了保证模型不欠拟合，顺利收敛，下述实验将迭代数从200增加到1000.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Acc | P | R | F1 |
| MLP | 0.984 | 0.991 | 0.982 | 0.986 |
| MLP +  Mish Activation | 0.974 | 0.957 | 1.0 | 0.978 |
| MLP + Skip  Connection | 0.989 | 1.0 | 0.982 | 0.991 |
| ResNet34 | 1.0 | 1.0 | 1.0 | 1.0 |

可见，增加Mish激活和增加跳连接的改进均能提高MLP的分类精度。而且，更为复杂的卷积神经网络ResNet34在测试集中实现了完全分类正确。下面展示各方法在训练期间的损失收敛曲线。

|  |  |
| --- | --- |
|  |  |
| MLP | MLP + Mish Activation |
|  |  |
| MLP + Skip Connection | ResNet34 |

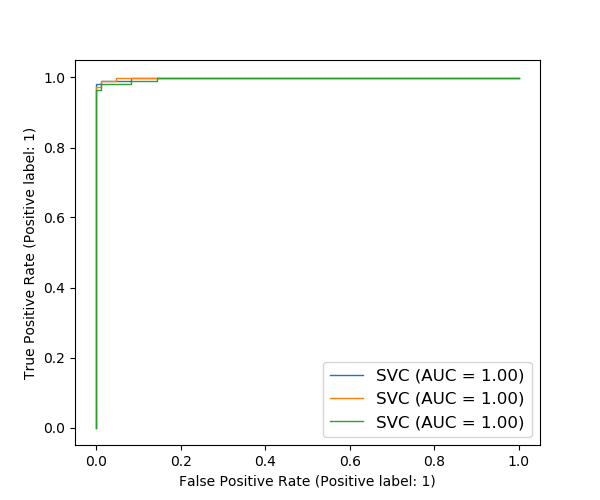
可见，上述神经网络在约等1/3次最大迭代（约300次）时开始收敛，结果正常。

三、支持向量机

有三种核函数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 核函数 | Acc | P | R | F1 |
| Poly | 0.964 | 0.949 | 0.991 | 0.969 |
| Rbf | 0.974 | 0.965 | 0.991 | 0.979 |
| Linear | 0.974 | 0.973 | 0.982 | 0.978 |

可见，rbf拥有最佳精度。



从ROC曲线上看，三种核函数的差异不大

现对rbf-svm进行进一步调试其惩罚参数C

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| C | Acc | P | R | F1 |
| 0.5 | 0.979 | 0.974 | 0.991 | 0.982 |
| 1 | 0.974 | 0.965 | 0.991 | 0.979 |
| 2 | 0.989 | 0.982 | 1.0 | 0.991 |
| 3 | 0.989 | 0.982 | 1.0 | 0.991 |
| 4 | 0.989 | 0.982 | 1.0 | 0.991 |

可见，增加C可以获得收益，但当C>2时，模型已无显著变化