1. 数据采集

舞弊公司财务数据采集

第一步：

依靠搜索引擎找到近年来财务造假的企业的名称、年份、股票代码。

第二步：

挑选出其中股票仍未退市的公司的名称和造假年份。

这里我们检索到了【‘尔康制药’，‘复星医药’，‘康美医药’，‘思创医药’，‘ST太安’，‘亚太’】六家财务造假药企，其中各自财务造假的年份为【【2015-2016】，【2018】，【2016-2018】，【2019-2020】，【2018-2021】，【2016-18】】，同时分别在[**同花顺个股**](https://stockpage.10jqka.com.cn/)中找到他们对应的股票页面。（因为后面的脚本都是按同花顺的页面格式定制，所以务必在同花顺个股页面搜索对应股票）

|  |  |
| --- | --- |
| 尔康 | <https://stockpage.10jqka.com.cn/300267/> |
| 复星 | <https://stockpage.10jqka.com.cn/600196/> |
| 康美 | <https://stockpage.10jqka.com.cn/600518/> |
| 思创 | <https://stockpage.10jqka.com.cn/300078/>， |
| ST泰安 | <https://stockpage.10jqka.com.cn/002433/> |
| 亚太 | <https://stockpage.10jqka.com.cn/002370/> |

第三步：

数据爬取，将对应网址带入进对应脚本中爬取对应公司的财务数据。

数据结构：

\_dic = {

    '行业': [],

    '股票': [],

    '指标': [],

    '时间': [],

    '值': []

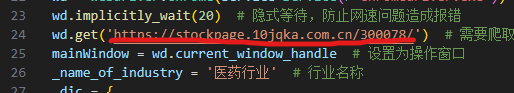
}

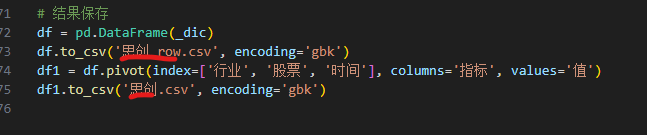
指标选取：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 盈利能力指标 | |  | | --- | | 净资产收益率 | | 净资产收益率-摊薄 | | 基本每股收益 | | 销售净利率 | | 销售毛利率 | |
| 偿债能力指标 | |  | | --- | | 流动比率 | | 速动比率 | | 资产负债率 | |
| 成长能力指标 | |  | | --- | | 净利润同比增长率 | | 营业总收入同比增长率 | |
| 运营能力指标 | |  | | --- | | 营业周期（天） | | 存货周转率（次） | | 存货周转天数（天） | | 应收账款周转次数（天） | |

使用脚本：[违规企业财务数据爬取.py](https://d.docs.live.net/b291d95f9061c8fb/桌面/基于BP神经网络的上市公司审计风险预测/1数据采集/违规企业财务数据爬取.py)

爬取不同的企业要修改两个参数，一个是企业的股票页面网址wd.get(网址)，一个是后面存储的文件的名称





我们需要4只股票的数据就需要运行4次，每次更改不同的网址和输出文档名称

注意，名字带row的csv文件是未从新定义索引的，目的是方便带入SPSS。

未舞弊上次公司财务数据爬取：

使用脚本：[未违规企业财务数据爬取.py](https://d.docs.live.net/b291d95f9061c8fb/桌面/基于BP神经网络的上市公司审计风险预测/1数据采集/未违规企业财务数据爬取.py)

直接运行即可，脚本会爬取医药商业板块的20只股票的所有的有年度记载的分季度财务报告

1. 样本选取

我们选择【复兴，康美，思创】的数据和未违规股票的80%作为训练集，选择【尔康】和未违规股票的20%作为测试集。

有些年份网站没有记录到信息，我们要剔除对应年份的样本。

有些单元格为百分数格式和货币格式，需要将这些格式转化为浮点数，满足tensor张量的转换要求。

运行脚本来执行，运行即可。

[数据集合成.py](https://d.docs.live.net/b291d95f9061c8fb/桌面/基于BP神经网络的上市公司审计风险预测/2数据预处理/数据集合成.py)

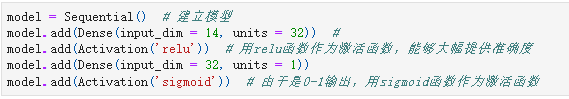
1. 训练神经网络模型

使用tenserflow的Sequential()模型。Sequential()模型即顺序模型，是神经网络的基础模型，全连接神经网络、卷积神经网络CNN、循环神经网络RNN等都可以用此模型构建。

模型通常分为五个步骤：

1.定义模型2.定义优化目标3.输入数据4.训练模型5.评估模型的性能

建立模型：



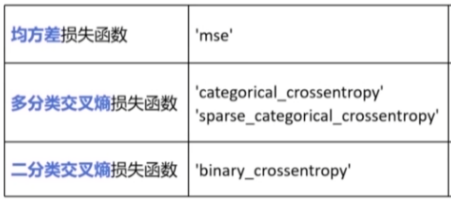
首先建立了Sequential()模型，然后向空的模型容器中加入两个全连接层Dense。

第一个全连接层有32个单元，使用relu函数激活，输入神经元为14个，对应我们选取的14个指标。

第二个全连接层有一个单元，使用sigmiod函数激活。

定义优化目标：



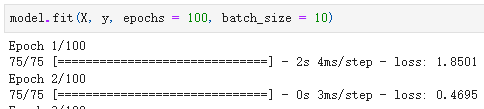
损失函数选择二分类交叉熵损失函数

优化器选择adam

SGD是小批量随机梯度下降算法

Adam是对SGD的扩展，可以代替经典的随机梯度下降算法来更好地更新网络权重

输入数据，训练模型：



输入特征集为X，类别标签为y

训练次数为100次，过多的训练次数可能造成模型的过拟合

批量大小为10，较大的批大小虽然能够达到与较小的批大小相似的训练误差，但往往对测试数据的泛化效果更差

batch size越大：

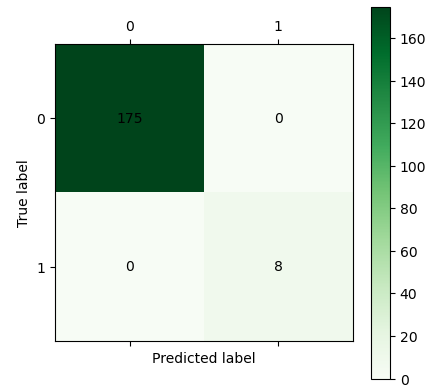
训练损失减少的越慢。

最小验证损失越高。

每个时期训练所需的时间越少。

收敛到最小验证损失所需的 epoch 越多

评估模型的性能：

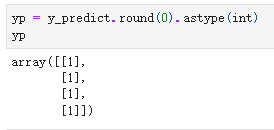
混淆矩阵，模型精度1.0

1. 运用模型进行预测

使用尔康制药2015年的财务数据数据（尔康制药2015.xlsx）

尔康制药在2016年被报道财务造假被证监会处罚，我们看当年的财务数据中的异常能否被模型预测到。



预测结果全部正确