数据科学导论 实验报告

一、队伍信息

赛题：乘用车细分市场销量预测

队名：写代码像蔡徐坤

姓名	学号	分工
金哲欣	PB17111663	MLP、协同过滤、长尾效应、规则、融合、可视化分析
许世晨	PB17030846	数据预处理、残差神经网络、DFM、CNN、LSTM、SVM
李纯羽	PB17111618	LGB模型、特征工程、数据预处理

二、比赛成果

- A榜：第 196 名
- B榜：第 145 名（本赛题下排名最靠前的科大队伍）
- 自主实现了MLP、协同过滤、DFM、CNN、LSTM、SVM等模型，其中MLP的输出结果经过长尾处理之后，得到的分数超过了当时所有开源的 Baseline ，对已开源的 Baseline 模型，包括LGB、XGB、规则模型进行学习和改进，并受其启发，改进自主实现的模型。

三、进展概要

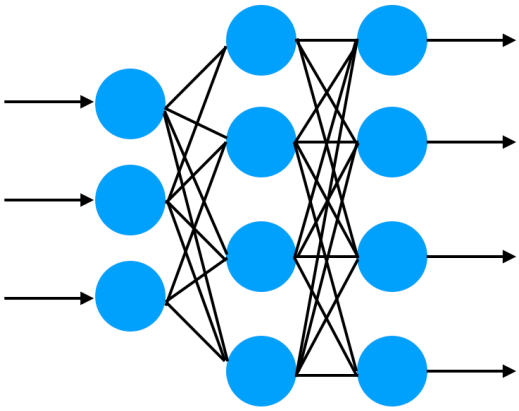
四、模型介绍

1. 多层感知机 MLP

朴素的多层感知机

概念介绍

全连接层是最简单的神经网络结构之一，它的每个神经元都与上一层的所有神经元相连接。许多全连接层首尾相连，便构成了一个多层感知机，我们称第一层为输入层，最后一层为输出层，中间的层都称为隐藏层。下图是一个输入纬度是3，输出维度是4，隐藏层神经元个数为4的全连接神经网络：



神经网络可以视作一个函数，接收输入x，获得输出y，通过训练集中真实的y与神经网络输出的y进行对比，使用反向传播算法修正神经网络所代表的函数，最终拟合出一个接近真实的函数。

数据划分

由于我们获得的训练集仅仅只有两个年，所以如果仅拿上一年的1～12月预测下一年的1～4月，那么下一年的5～12月的数据就无法被利用，这不利于神经网络的训练。

所以我们决定使用前12个月预测后4个月的方法，比如用2016.1～2016.12的数据预测2017.1～2017.4，用2016.2～2017.1的数据预测2017.2～2017.6。这样做既可以在一定程度上保留一年12个月的周期性特征，也大大增加了训练数据的条目数。

此外，我们还对数据的划分进行了其他尝试，比如：输入前8个月输出后4个月、输入前6个月输出后4个月、输入前4个月输出后4个月等。

模型实现

我们使用 Tensorflow 搭建了一个4层的神经网络，隐藏层神经元个数为32，每层输出都经过一个 Sigmoid 激活函数，并使用正态分布初始化权重。

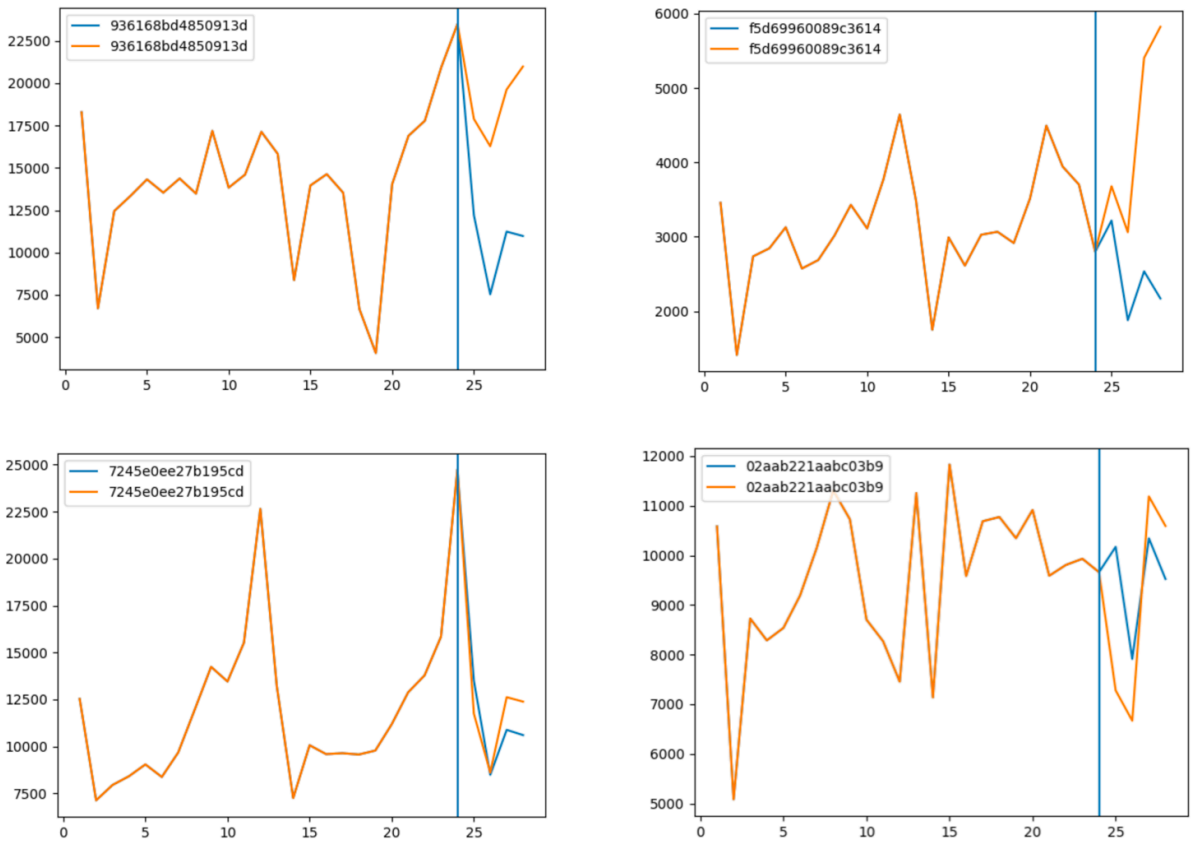
```
model = keras.Sequential([
    layers.Dense(32, activation='sigmoid', kernel_initializer='he_normal',
input_shape=(x_train.shape[1], )),
    layers.Dense(32, activation='sigmoid', kernel_initializer='he_normal'),
    layers.Dense(32, activation='sigmoid', kernel_initializer='he_normal'),
    layers.Dense(4)
])
```

结果分析

最终的训练结果不尽人意，分数大概在0.3～0.4左右，远远低于开源的 baseline 的水平。

在多种数据划分的尝试中，我们发现使用前4个月预测后4个月的效果最佳。这使我们感到诧异，因为使用前4个月预测后4个月，意味着我们放弃了数据中潜在的周期性规律，仅仅寄希望于拟合前4个月销量对后4个月销量的影响。我们认为这种差异有可能是数据量的不同造成的，因为如果使用前12个月预测后4个月，我们对某个车型某个省份只能分割出8个数据条目，总共 $8 \times 22 \times 60 = 10560$ 个数据条目；然而，如果使用前4个月预测后4个月，我们对某个车型某个省份只能分割出16个数据条目，总共 $16 \times 22 \times 60 = 21120$ 个数据条目。

我们将该模型的结果与我们目前得到的最优结果进行可视化对比：（蓝线为最优结果，黄线为当前模型）



可以看到在某些车型上，朴素多层感知机的预测结果与最优结果相去甚远。

残差神经网络

平稳化处理

长尾效应处理

2. 协同过滤

3. DFM

4. LSTM

5. LGB

6. 规则

7. 模型融合

五、心得体会

六、参考文献