价格预测PLUS

周沁泓 钮泽平

1 数据集的处理:

1.1 训练集与测试集的划分

- 训练集与测试集按照作业说明中的要求进行划分, `20170703-20170809` 为训练集; `20170810-20170825` 为测试集
- 由于训练集本身已经比较小了,验证集抽取方法为在将测试集标注后随机抽取20%作为验证集

1.2 类别标注

- 标注为三类,与作业要求一致,其中$d_{a,b}(t)$中的 `a` 与 `b` 在 `config.py` 中分别有 `predict_st` , `predict_ed` 对应, 我们取a=5, b=25, θ=0.15%
- 整个时间切成seg_time时间长的小段,每两个训练样本的起始时间的最短间隔为seg_time, 使用5min(300s)的历史数据作为输入向量, 取seg_time=5, 历史数据时间长data_time=300
- a, b, θ, seg_time, data_time都可以在config.py中配置

1.3 数据预处理(特征的提取)

1. 对数据进行平衡, 保留珍贵的上涨/下降数据,丢弃不变类别的数据,使"涨:平:跌=3:4:3"
2. 寄希望于CNN可以提取特征,使用1维CNN进行卷积核自动学习,结果算法根本不收敛(正确率在0.02%与99%之间摆动)
3. 使用"人工卷积核", [-1, 1]进行卷积,之后进行均值池化, 这种卷积相当于相邻两项相减, 即提取出了价格的变化率(导数), 平衡后的训练集正确率可以达到60%
4. 对导数再次卷积[-1, 1], 之后进行均值池化, 相当于得到价格变化率的变化率,即二阶导数,反映了价格曲线的曲率
5. 最终将上述两个序列中的元素交替排列,得到(导数,二阶导,导数,二阶导, ...)的特征向量,这个特征向量保持了时间序列的性质,方便之后使用LSTM进行进一步特征提取

2 模型结构:

2.1 动机

- 考虑到期货A1和A3具有强相关性, 我们可以把A1与A3的历史数据同时作为输入, 对训练集进行拟合, 可是数据维度过高,训练效果不好
- 使用A1的历史数据对A1的走势进行预测,如果采用神经网络模型,最后一层的softmax输出有着清晰的含义:数据属于各个类别的概率
- 如果能够使用A3的历史数据,对A1走势进行预测,和用A1历史数据预测的模型做一个"投票",那么模型的准确率应该是能够得到提高的
- 如何投票呢?使用神经网络学习!

2.2 单一模型

2.2.1 动机

- LSTM可以记忆历史的数据,对于时间序列十分给力
- LSTM输出上再加一个LSTM,或许可以获取更长时间维度上数据的关联

2.2.2 模型结构

1. 输入: 提取出的由价格的导数和二阶导数构成的向量(600x1)
2. 模型结构: Reshape->GRU->LSTM->BatchNorm->Dense->Dropout->Dense

```
Layer (type)                 Output Shape              Param #
=================================================================
reshape (Reshape)            (None, 40, 15)            0
_________________________________________________________________
cu_dnngru (CuDNNGRU)         (None, 40, 30)            4230
_________________________________________________________________
cu_dnnlstm (CuDNNLSTM)       (None, 30)                7440
_________________________________________________________________
batch_normalization          (None, 30)                120
_________________________________________________________________
dense (Dense)                (None, 64)                1984
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense  (Dense)               (None, 3)                 195
=================================================================
```

2.3 组合模型

2.3.1 动机

- “四个专家三个说今天下雨,所以下雨概率是75%”
- “三个臭皮匠赛过诸葛亮”

2.3.2 模型网络结构

1. 预训练好的 用A1历史预测A1走势 和 用A3历史预测A1走势, 对它们的网络权值进行冻结,使其后续不能被训练
2. 将输出的概率进行concat, 作为组合模型的输入, 加入Dense层和softmax层再次进行分类, 实现模型的组合

```
Layer (type)                 Output Shape              Param #
=================================================================
merge_6 (Merge)              (None, 6)                 0
_________________________________________________________________
dense_45 (Dense)             (None, 10)                70
_________________________________________________________________
dense_46 (Dense)             (None, 3)                 33
=================================================================
```

3 测试结果(均在测试集上)

1. 预测A1的模型:

类别	正确率	召回率	随机猜测
上涨	0.0444	0.3122	0.0226
不变	0.9727	0.6639	0.9544
下跌	0.0400	0.3270	0.0228

2. 预测A3的模型:

精确率均值:0.0701
召回率均值:0.3940

类别	正确率	召回率	随机猜测
上涨	0.0738	0.3494	0.0395
不变	0.9639	0.5633	0.9302
下跌	0.0665	0.4386	0.0419

3. 预测B2的模型:

精确率均值:0.0125
召回率均值:0.3513

类别	正确率	召回率	随机猜测
上涨	0.0143	0.4043	0.0081
不变	0.9907	0.5882	0.9851
下跌	0.0108	0.2984	0.0068

4. 预测B3的模型:

精确率均值:0.0280
召回率均值:0.3411

类别	正确率	召回率	随机猜测
上涨	0.0308	0.3963	0.0193
不变	0.9793	0.5748	0.9643
下跌	0.0251	0.2860	0.0164

4 结论

1. 本次实验使用3:4:3平衡后的数据进行训练,最后在高度不平衡的测试集上,正确率高于按照测试集各类别比例随机猜测,可以看出模型成功识别出了上涨/下跌/保持不变三类数据的特征
2. 在A1, B2, B3数据上, 关于上涨和下跌的预测精度均能基本达到随机猜测准确度的两倍,召回率也控制在合理的范围内,模型是十分有效的.对于A3类别,模型也成功达到了随机猜测的准确率之上