

《数据科学导论》大作业

1.题目：金融数据分析与量化交易。

2.数据集：由老师和助教准备。

10 只股票的 ≥ 15 年的 Daily Price 数据(stock01.csv、stock02.csv、stock03.csv、stock04.csv、stock05.csv、stock06.csv、stock07.csv、stock08.csv、stock09.csv、stock10.csv)，用于模型训练，助教发给同学们。

另外 10 只股票的 ≥ 15 年的 Daily Price 数据 (stock11.csv、stock12.csv、stock13.csv、stock14.csv、stock15.csv、stock16.csv、stock17.csv、stock18.csv、stock19.csv、stock20.csv)，用于评测，助教掌握，不发给同学们。

为了测试自己的模型，同学们可以利用助教发布的数据，适当切割出训练集和测试集。

3. 任务描述：

3.0 简述

本大作业包含 2 个阶段的任务，分别是模型训练阶段和模型测试阶段，描述如下：

(1) 模型训练阶段

同学们拿到 10 只股票的 Daily Price 价格数据。

首先在这些股票价格数据上,进行数据标注,标注合适的买入、卖出、不做动作的 Label。注意这时候，可以针对整个数据集进行处理和分析，以标定合适的 Label。

然后，根据需要从原始数据计算一些导出的特征即 Factor，这些特征可能对上述 Label 具有更强的解释作用。在原始数据以及导出的特征即 Factor 上，加上 Label，建立样本。注意，这时候样本的构建，只能利用到目前为止的数据即历史数据，不能利用未来的数据。

利用样本进行模型训练。模型训练完成后，模型存盘。

该模型具有根据历史上到目前为止的价格数据（以及自行构造的特征即 Factor），做出买入、卖出、不做动作的预测的能力。

(2) 模型测试阶段

同学们修改老师提供的测试代码。

在代码里装载训练好的模型，并且流式地接收和处理依次到达的 Daily Price Bar，利用到目前为止的原始数据计算必要的导出特征即 Factor，然后利用到目前为止的原始数据和导出特征即 Factor（不能利用未来数据），构造新样本，馈入模型，得到买入、卖出、不做动作的输出即 Output。

测试代码根据模型预测的 Output，执行买入、卖出、不做交易等动作。

针对 10 只股票进行测试，得到在这 10 只股票上的利润率、最大回撤、夏普指数、交易次数等指标。

这些测试结果将作为评分依据。

3.1 数据描述

stock01.csv 到 stock10.csv 等文件，其结构是一样的，都包含 Datetime、Open、High、Low、Close、Volume、Adj Close 等字段，表示日期、开盘价、最高价、最低价、收盘价、交易量、临近收盘价等。

3.2 数据标注

同学们可以在 stock01.csv 到 stock10.csv 上进行标注，每个数据点标注买入 1、卖出-1、或者不做动作 0 的 Label。

此环节，老师提供示例代码并在上机课讲解。

同学们应该各展其能，手工或者编写程序自动对数据进行上述标注。

3.3 构造特征

除了 Open、High、Low、Close、Volume、Adj Close 等字段，可以利用 TA-Lib 等 Python 库，构建更多的特征即 Factor；或者根据自己对数据的理解，自行构造特征即 Factor。

此环节，老师提供示例代码并在上机课讲解。

同学们应该各展其能，构造具有解释力的特征。

3.4 构造样本

利用到目前为止的数据以及构造的特征即 Factor，构造训练样本，训练预测模型。该模型根据输入，决定买入、卖出、还是不做动作。

此环节，老师提供示例代码并在上机课讲解。

3.5 训练模型

可以使用从 stock01.csv 到 stock10.csv 等 10 个文件构造的样本，一起训练模型。必要的时候可以对样本进行标准化、规范化等处理（不同股票的价格的绝对值有较大的差别，不利于训练统一模型），以使得模型能够处理不同价位的、不同交易量的股票，即具有较强的泛化能力。

训练好的模型，保存到 Pickle 文件。

此环节，老师提供示例代码并在上机课讲解。

3.6 评测模型

同学们冻结代码。

同学与助教坐下来，运行测试代码，登记运行结果。

测试代码装载训练好的模型，在 10 只股票上进行测试，根据不断到达的 Daily Price Bar，利用目前为止已有的数据以及导出特征即 Factor，进行买入、卖出、不做动作的预测，并且完成交易。最后计算利润率、最大回撤、夏普指数、交易次数等指标。每位同学的测试结果，填写下表的一行。

学号	姓名	Stock 11 结 果	Stock 12 结 果	Stock 13 结 果	Stock 14 结 果	Stock 15 结 果	Stock 16 结 果	Stock 17 结 果	Stock 18 结 果	Stock 19 结 果	Stock 20 结 果

注意，10 只股票，每只股票的结果包括利润率、最大回撤、夏普指数、交易次数等 4 个指标。

3.7 撰写报告

撰写 Word 试验报告，内容包括报告题目、作者信息、摘要、关键字、简述、数据标注、特征构建、模型优化、模型测试结果、总结、参考文献等。

4. 成果提交：




在上机课上，同学们当面和助教一起完成模型评测，登记评测结果，作为评分依据。

把（1）数据与源代码（*.ipynb、*.py）（2）试验报告 Word 文件，一起压缩为 zip 文件，提交到 OBE 网站。代码和报告也作为评分依据。

备注：notebook 的每个代码 Cell，如果有需要加上必要的注释。

5. 参考文献：

请参考压缩包里“参考”目录下的内容。

名称	类型	大小	修改日期
 《数据科学实践》Chapter04智慧金融.pdf	Adobe Acrobat ...	2,162 KB	2024/9/23 15:14
 《数据科学实践》Chapter04智慧金融-PPT.pptx	Microsoft Power...	10,007 KB	2024/7/2 16:25
 Chapter04_Code&Data.zip	360压缩 ZIP 文件	2,308 KB	2024/9/23 15:15

6.评价方法：

优秀率控制为 30%。

（1）根据利润率越高越好，对全班同学进行排序，根据排序情况，对分数 1 进行赋分 75-95。利润率小于 0，直接对分数 1、2、3、4 赋分 60。

（2）根据最大回撤越小越好，对全班同学进行排序，根据排序情况，对分数 2 进行赋分 75-95。

（3）根据夏普指数越大越好，对全班同学进行排序，根据排序情况，对分数 3 进行赋分 75-95。

（4）根据交易次数越多越好，对全班同学进行排序，根据排序情况，对分数 4 进行赋分 75-95。

注意，利润率、最大回撤、夏普指数、交易次数等可能互相冲突，请注意自行折中处理。

（5）计算最后得分，为分数 1、分数 2、分数 3、分数 4 的平均分。