金融大数据课程个人报告

2024100908 金威廷

1. 个人贡献

在本次金融大数据课程的期末作业中，我主要负责多因子模型的构建与回测框架的设计与实现。我的贡献可以细分为以下几个方面：

1. 多因子模型构建

参考Gu等（2020），我构建了四种多因子模型，不仅涵盖了经典的线性回归方法，还引入了降维技术和正则化方法，能够适应不同的数据特征和分析需求。具体包括：

OLS（使用Huber损失函数）：通过引入Huber损失函数，增强了模型对异常值的鲁棒性，能够在金融市场数据中更好地处理极端值的影响。

PCR（主成分回归）：利用主成分分析（PCA）对高维数据进行降维，解决了多重共线性问题，同时减少了模型的过拟合风险。

PLS（偏最小二乘法）：在降维过程中同时考虑了自变量和因变量的信息，提升了模型对因变量的解释能力。

GLM（广义线性模型）：结合样条基函数和Group Lasso正则化，捕捉了非线性关系，并通过分组正则化增强了模型的稀疏性。

1. 回测框架设计与实现

我设计并实现了整个项目的回测框架，具体包括：

**样本分割**：我构建了TimeSeriesSplitter类，用于将时间序列数据划分为训练集、验证集和测试集。划分方法严格遵循时间顺序，确保模型在样本外数据上的测试能力。其中训练集用于模型参数的估计；验证集用于超参数调优，模拟模型的样本外测试；样本外测试集作为真正的样本外数据，用于评估模型的最终性能。

**滚动窗口机制**：通过滚动窗口的方式，逐步向前移动时间窗口，确保模型能够在不同时间段上进行测试，从而更全面地评估其稳定性。

**回测与评估**：我使用backtrader库对所有模型进行了回测，模拟了模型在实际市场环境中的表现，具体包括：

收益统计：计算了模型在测试集上的累计收益、年化收益等关键指标。

R2分数：评估了模型对因变量的解释能力。

信息系数（IC）：计算了模型预测值与实际值之间的相关性，衡量了模型的预测能力。

在回测框架中，我采用模块化设计，用类的形式对模型、分割器等不同部分进行封装，能够灵活支持不同模型的回测需求。并且通过自动化脚本实现了从数据加载、模型训练到回测评估的完整流程，显著提高了回测效率。

1. 对课程的理解与思考
2. 量化投资的核心理念

通过本次课程，我深刻理解了量化投资的核心理念：“数据驱动决策”。量化投资通过对历史数据的分析和建模，挖掘市场中的规律，并基于这些规律制定投资策略。本次作业中，我构建的多因子模型正是量化投资中常用的工具，能够帮助投资者从海量数据中提取有效信息，辅助投资决策。

1. 多因子模型的重要性

多因子模型是量化投资中的重要工具，它通过结合多个因子来解释资产收益率的变动。在本次作业中，我不仅深入理解了模型建立、优化和评估的完整流程。这让我认识到，多因子模型的成功不仅依赖于数学和统计方法，还需要对金融市场有深刻的理解。

1. 对课程的建议
2. 增加实战案例

建议课程增加真实的金融数据分析案例，帮助学生更好地理解理论知识在实际中的应用。

1. 引入更多工具

可以引入更多量化分析工具（如微软的Q-Lib、Backtrader等），丰富学生的技术栈。此外，还可以向学生介绍现在市面上好用的量化分析平台，如JoinQuant和WorldQuant。JoinQuant是一个专注于量化投资的在线平台，提供了丰富的历史数据和实时数据，用户可以在平台上直接编写策略并进行回测和模拟交易。WorldQuant则是一个全球化的量化研究平台，提供了海量的金融数据和强大的研究工具，在WorldQuant上只需要用户构建Alpha表达式，就可以一键回测，成为签约顾问后还会有补贴和绩效。这些工具和平台不仅能够帮助学生更好地理解量化投资的实践流程，还能为他们未来的职业发展提供更多可能性。

1. 总结

通过本次课程作业，我不仅提升了自己的技术能力，还加深了对量化投资的理解。我在小组中负责的多因子模型构建与回测框架设计，为团队任务的顺利完成提供了重要支持。未来，我将继续深入学习量化投资的相关知识，探索更多数据驱动方法在金融领域的应用。