金融大数据课程个人报告

2024100924 李宇杰

1. 个人贡献

#### 1. 多因子模型构建

在本次课程任务中，我主要负责多因子模型的构建，包括多个回归模型的设计与实现，具体涉及ElasticNet+Huber回归、RandomForest、GBRT+Huber以及Neural\_Network。

**ElasticNet+Huber：**

我设计了一个结合Huber Loss与Elastic Net正则化的回归模型——**HuberElasticNet**。由于scikit-learn中的Elastic Net默认采用MSE损失，并未提供直接的Huber Loss实现，我通过继承sklearn.base.BaseEstimator和RegressorMixin类扩展了新的模型。在实现过程中，我使用了梯度下降法进行参数更新，在\_compute\_gradient方法中首先计算损失部分的梯度，然后计算L1和L2正则化部分的梯度，并将其合并进行参数更新。

**RandomForest：**

在实现随机森林时，最初使用了sklearn.ensemble.RandomForestRegressor，但在面对大规模和高维数据时，模型的计算速度和内存占用表现不理想。为了解决这一问题，我转向了LightGBM框架，并通过将boosting\_type参数设置为'**rf**'，模拟随机森林的构建方式。

**GBRT+Huber：**

在GBRT模型的实现中，我同样使用了LightGBM框架，并将关键参数设置为boosting\_type='gbrt'和objective='huber'，以便同时使用梯度提升树结构并直接应用Huber损失。这种设计确保了模型在面对高噪音和异常值时能够保持较强的鲁棒性。

**Neural\_Network：**

在神经网络部分，我设计了全连接神经网络模型框架，并通过设置n\_layer参数控制全连接层的数量。然而，由于GPU资源有限以及A股数据噪音较大，本次实验中我仅实现了两层的神经网络。具体原因将在模型提升部分详细说明。

#### 2. 模型提升部分

**LGBM模型的优化：**

在LightGBM模型的基础上，我通过调参进一步提升了模型表现。引入了subsample和feature\_fraction等参数，通过随机采样数据和特征来减少模型对噪音的过拟合，增加偏置，提升泛化能力。同时，我增大了n\_estimators并减小了学习率，使得模型在训练时能够更精细地调整，从而提高预测精度。后续，我将模型迁移至XGBoost，并通过GPU加速和精细的特征切分机制进一步提升了训练速度和预测效果。

**神经网络优化：**

对于神经网络的优化，我调整了学习率，采用了AdamW优化器，并规范了正则化使用方式，缓解了鞍点问题。我还通过初始化权重和限制梯度的方法，避免了梯度爆炸现象，使得神经网络能够更稳定地进行训练，提升了在复杂数据上的学习效果。

#### 3. 回测结果分析与阐释

我负责了回测结果的分析与阐释工作，主要是对不同模型在回测中的表现进行对比和分析。通过对模型在不同数据频率（如日频和月频）上的回测结果进行综合评估，我发现XGBoost在年化收益率上表现优于LightGBM，而在日频数据上，二者表现相似。XGBoost可能因其Level-wise的分裂策略，在面对特征复杂的数据时不够灵活，导致拟合不足。

此外，我还发现，所有基于梯度下降的模型在回测中的表现均不如树模型、随机森林等传统模型。这主要是因为梯度下降法容易受到噪音干扰，陷入优化瓶颈，尤其在高频数据中容易遇到局部极值或梯度爆炸问题，导致模型未能有效学习到数据中的真实信号。相比之下，树模型和随机森林由于其较为稳健的参数更新策略，能更好地应对数据中的噪音和复杂性。

1. 对课程的思考

在本课程的学习过程中，我不仅掌握了多因子模型的构建与优化技术，也更加深入地理解了机器学习方法在金融数据中的应用。课程内容十分丰富，特别是在模型实现与理论方面，但个人感觉若增加一些更具实操性的项目和同学间的互动交流环节，会使课程更加完善。

具体来说，目前课程的主要任务是复现论文，这为我们提供了一个很好的学习基础，让我们能够熟悉模型的实现过程、参数调优和回测分析。但在实际的量化投资过程中，单纯的模型复现远不能满足实际需求。为了增强课程的实践性，我建议课程可以引入更多基于真实数据的项目，比如组织一个**量化投资打榜比赛**，让同学们在真实或模拟的市场环境中，使用所学知识设计模型并进行优化。这不仅能帮助我们加深对所学技术的理解，还能锻炼团队合作、项目管理、数据分析等多方面的能力。

通过这种方式，课程将更注重实际操作，同时加强同学间的互动和讨论，促进彼此的思维碰撞。打榜比赛也能让我们体验到如何处理大规模数据集、如何应对金融市场中的随机性和复杂性，以及如何在实际应用中调优模型性能，这对我们未来从事量化投资或数据科学工作将产生非常积极的影响。