说明

从研报与研报论文开始。

讨论一：模型输入的特征是什么?

研报并没有说明输入到几个模型中的特征是什么。通过仔细阅读，从研报的第8页5.2节中的3的描述的一共16个预测信号估计是将4.1节中对Baz文章中的MACD计算指标的值作为输入预测。我们据此做了一个版本的模型。

讨论二：模型与数据预处理?

研报没有指出任何的数据预处理，翻看原文也没有任何的说明。就是用这个几个方法做出了效果。

讨论三：时间维度与样本训练方法?

文中第8页，5.2回测与预测描述中，说，所有所有模型的超参数5年调整一次。策略在每月调仓。这里有几种理解：

1. 使用过去五年的数据，按照月对数据进行训练。训练一次模型对后五年进行预测。每五年更新一次模型。
2. 模型是滑动的，即每次使用前几个月来对后面一个月进行训练。每五年进行一次超参数的调整。

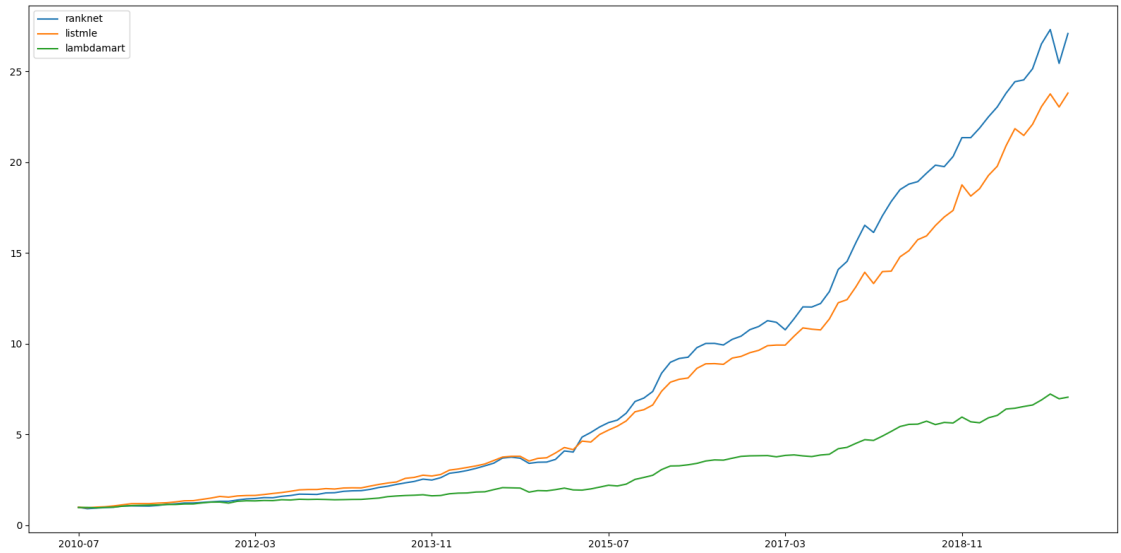
我这边的理解是第二种，其原因在于近期的数据训练出来的模型更可以反映新的情况，模型也比较动态。第二是因为pairwise方法会对数据达成平方级别的扩种。即一个月数据为N的情况下，六个月数据量为$36N^2$的数据量。五年就是$3600N^2$。从训练的角度来看，时间增加100倍。训练困难，且没有展现出动态性。

关于超参数问题，因为巨大计算量。比如一次N个数据的训练，pairwise数据量变成N^2，再加上文中参数搜寻过程，一次计算量变成了5^5N^2量级。这种计算量不是很合理。所以，本文没有进行超参数搜索。

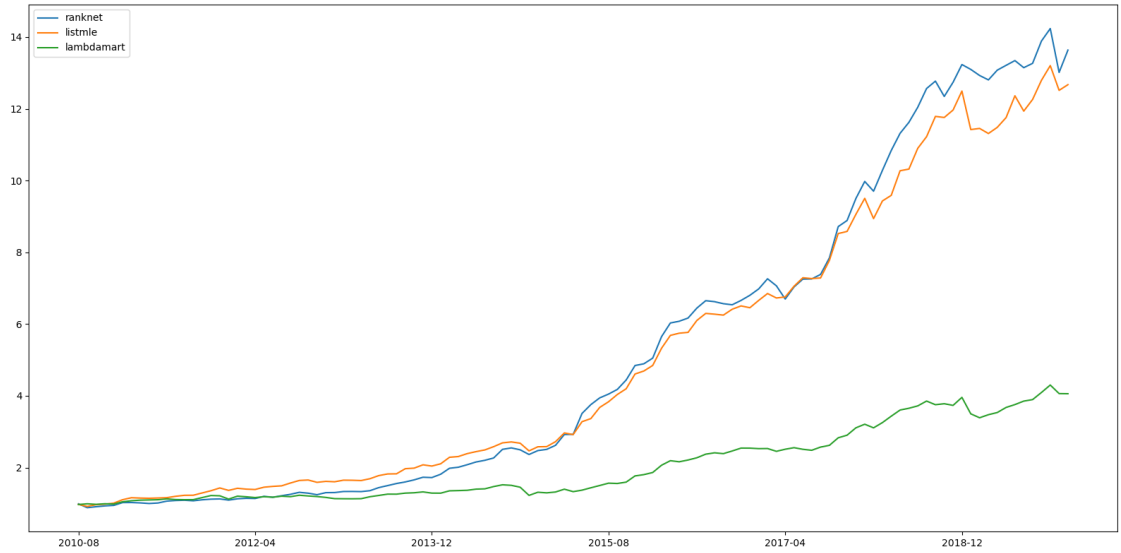
数据使用

本文使用的数据是之前给的wind数据库数据，且计算上述16个特征。最终效果不好。其原因在于，MACD作为一种动量策略（Baz 2015），在美国市场效果不错。国外的动量在国内效果是不尽人意的，关于国内的A股的动量因子说明可以寻找一些类似的文章。

本次最终的结果使用的是Adaboost那次项目提供的A股市场的月度因子数据。



关于波动率调整的部分没有实现，没实现的原因是使用的数据中没有波动率数据。所幸，大量文献指出，波动率与市值的开根号有正比关系。我们用市值对波动进行近似来得到波动调整的收益。由于没有办法设置15%阈值，所有，我们还是将波动作为一种加权方法。即对波动大的股票给予小的权重，波动小的给予较大权重。其思想与原文的波动调整是一致的。以下为波动调整后结果。



Ranknet里面的一些创新细节

ranknet中，我们在选择样本时，做了一些大量改进。因为我们并不是要对比所有的样本，于是我们在pair样本选择时，做的是

1. 当前排名与当前排名后100名进行学习。这样可以避免以下一种情况，当两个排名比较接近时，一个股票x收益为0.11，另一个为y为0.108。此时，区别其实很小，但是算法依然将两者当成较大区别。所以，大的跳步可以一定程度避免这种情况。
2. 我们需要算的是前100与后100分数的差别，中间排名200与排名300是不是能学好其实并不重要，所以，构建pairdata时，排除了中段排名可能对模型造成的影响。
3. ndcg指标注重头部排名，而我们的算法其实也很注重尾部排名，此时ndcg指标作为早停指标就不是很合适。因此，在ranknet中，我们使用了准确率来做早停的指标。
4. ranknet数据量大，因此，对应的学习率也应该更小对应一个epoch的数据量大的问题。

代码说明

本次项目分为两部分代码，s部分为对应原论文的方法。效果不佳。不建议跑。

但还是附上跑代码的顺序：s1 -> s2_main -> s3_main -> s4_main 跑完有对应的s2_eval_**.py文件查看结果。

N部分为上面展示效果的方法，该部分为该项目主要部分。

跑代码的顺序为：

N2-> N3 -> N4 -> N5

其中，有两部分数据需要自定义路径：

S项目的数据为：Beta dispersion and market timing项目中的new_stk_data文件夹中数据，请自行定义好绝对路径。路径位置对应s1.py中第8行base_dir。

N项目的数据为：Adaboost项目中的month_data数据文件夹，请将该数据集的绝对路径对应到每一个N_.py文件的month_list变量中。