# **【总】样本筛选&训练方案研究**

@王中元

# **Summary**

在金融机器学习中，由于pattern的不稳定性，过拟合的减轻几乎是最需要考虑的问题。

针对不同区间数据pattern（风格），研究在不同条件下对于训练数据与验证数据的选择，尽量使得模型的训练区间模式与预测区间模式保持一致，优化模型在预测区间的表现，简单来说就是通过训练样本的选择和分割来优化过拟合情况

# **任务拆分**

## **训练样本选择**

* 基于风险因子
* 基于行业因子
* 基于票池
* 基于其他低频长周期模式判断
* 基于标签值（降噪）
* 基于历史统计（例如训练区间与预测区间的时间间隔）
* 基于聚类

由于该部分为重难点仍需更多基础思考

## **训练方案**

* 滚动时间训练：需要考虑训练集长度与验证集的选择
* 历史数据筛选训练：需要同时考虑训练集筛选与验证集筛选
* 预训练+微调：基于长时间的历史数据并引入定制化交叉验证手段，试图得到一较为普适的预训练模型，并以一定时间频率加入最新数据进行微调（类似在线增量学习）