向量化回测

——基于机器学习开发流程

策略开发流程（开发人员前期工作）：

1、明确业务问题: 定下策略的核心逻辑

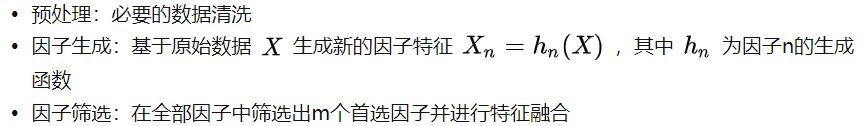
2、明确特征集: 搜集特征集、目标集数据，即x,y

3、基本计量分析: 对于每个特征标签，均进行统计分析，确认数据质量、了解其数据分布及特点

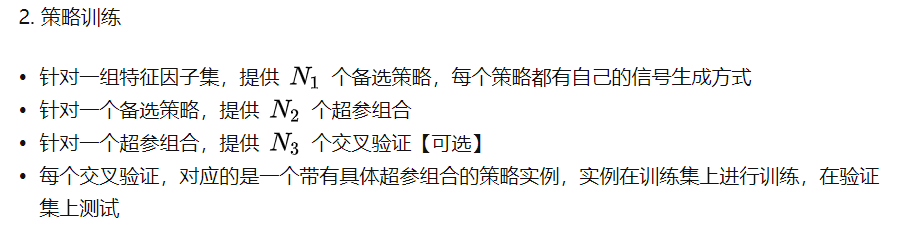
4、制作最小价值模型: 做一个简易的可行模型，例如建立线性回归模型,作为对复杂模型的效果对比

自动化流程：

1、特征工程



上述筛选因子的方式不一定采用，需要进一步验证这种方式是否更加有效，初步目标为：进行基础数据处理，生成我们需要的因子（如通过遗传规划筛选出合适的因子）



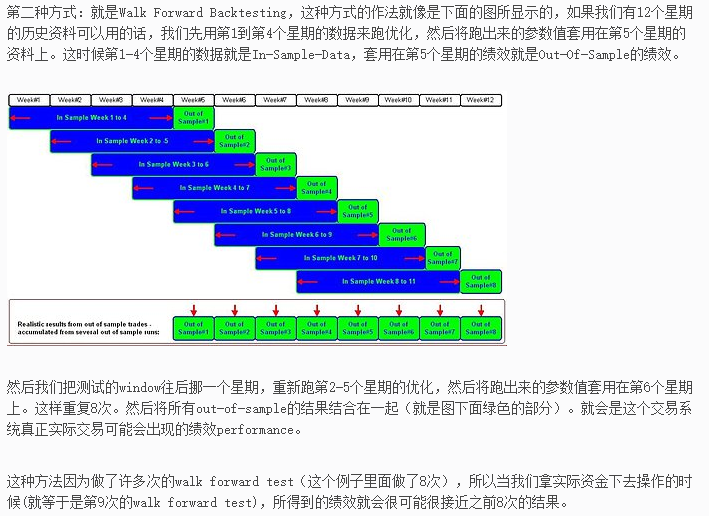
备选策略：提供多种交易逻辑，例如策略1：震荡策略，策略2：趋势策略，等等

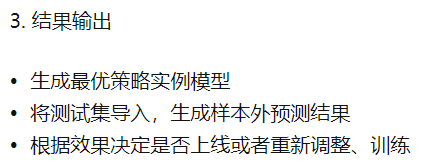
超参组合：例如学习率、参数优化步长、k均值聚类中的簇数等等的设置

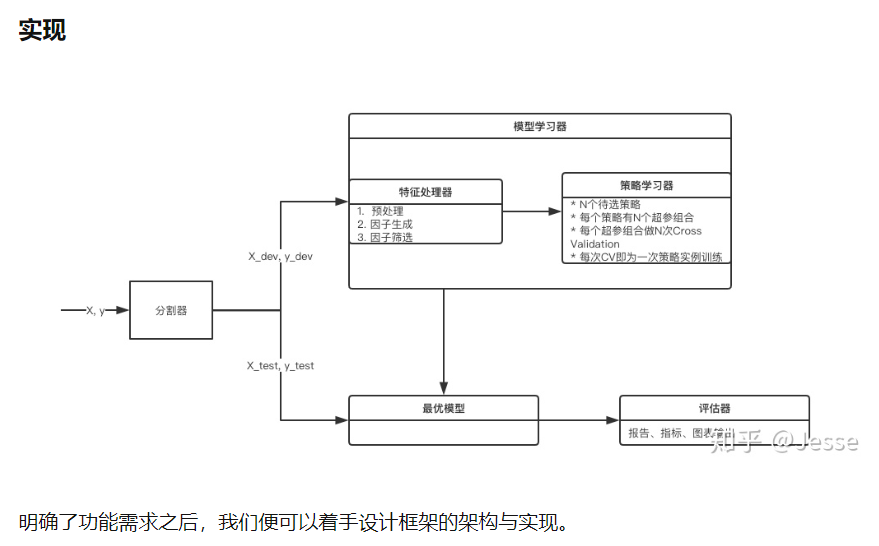
交叉验证：

（1）机器学习中推荐：K-fold Cross Validation(记为K-CV)：

将原始数据分成K组(一般是均分),将每个子集数据分别做一次验证集,其余的K-1组子集数据作为训练集,这样会得到K个模型,用这K个模型最终的验证集的分类准确率的平均数作为此K-CV下分类器的性能指标.K一般大于等于2,实际操作时一般从3开始取,只有在原始数据集合数据量小的时候才会尝试取2.K-CV可以有效的避免过学习以及欠学习状态的发生,最后得到的结果也比较具有说服性.

（2）也可以采用前向分析法： 





资料来源：<https://zhuanlan.zhihu.com/p/55302594>

回测框架设计：

因子数据(X)结构：三维数据，如下图：

股票名称

日 期

因子特征

时间序列数据如宏观宏观数据本身是二维数据，可以根据股票（期货）转化为三维数据

行情（收益率）数据（Y）（用于计算策略净值、绩效等指标）数据结构：二维

日 期

股票收益率

对原始因子数据X按时间序列（如天）抽取形成二维因子表：

股票名称

因子1

因子2

因子n

基于每天（月、周、分、tick）的交易逻辑，对当期因子表做处理，形成一维持仓表：

股票名称

持仓比例

基于每天的持仓表，形成二维持仓表：

日 期

股票1持仓比例

股票n持仓比例

基于持仓表与行情（收益率）数据计算每日组合收益率、净值及其他绩效数据。

向量化回测基本逻辑如上，需要解决的难点：

1. 是否能够将交易的规则映射到每天的因子表中（尽量避免采用for循环的形式）；
2. 当交易规则复杂时，如：当日涨停，则不交易，放入下一日成交；类似规则能否实现。