金融机器学习算法

lecture 1

- 金融数据的低信噪比
- 机器学习的适用领域：量化策略、量化交易、贷款审批、债券评级、公司分类、收益预测、风险评估、通胀预报、人才招聘
- 传统机器学习转向金融时的陷阱
    - 单打独斗的西西弗斯范式
    - 依据回测来筛选策略
    - 固定时长标签法
    - 同时决定交易方向与交易规模
    - 每一个示例看作独立同分布iid
    - 交叉验证时有信息泄露
    - 根据真实历史路径进行回测
    - 以回测结果为目标导致过拟合
- 机器学习：人工智能的一个分支，让计算机利用学习算法，从经验数据产生捕捉规律模式的模型，最后利用模型来改进预测的方法体系。
- 假设空间：潜在的真实模式称之为假设，学习过程是在所有的假设组成的空间中进行搜索的过程，找到与训练集匹配的假设。假设的经验对应物就是模型。
- 归纳偏好：机器学习算法在学习过程中对某种类型假设的偏好称为“归纳偏好”，它是学习算法的“价值观”。
- 好模型的两个要求：
    - 与训练集和测试集的匹配程度好（经验）　在平均意义下的损失情况
    - 整个样本空间的适应度高（推断）　在期望意义下的损失情况
- 评价单个样本拟合度的方法：损失函数
- 查全率Sen
- 查准率PPV
- 准确率ACC
- F1值：Sen和PPV的调和平均值
- 正则化方法：调整算法，在学习算法中增加控制模型复杂度的惩罚项，该方法增加了算法的超参个数，也会引入过拟合。
- 外部交叉验证：只能检验算法的性能，不能用于确定模型，综合评价在超参数个数固定时，学习算法复杂度导致的过拟合性。
- 偏差-方差分解
    - 偏差：度量期望预测与真实结果的偏离程度，表示拟合能力
    - 方差：同样大小的训练集变动导致模型预测变化，表示模型对数据的敏感性
    - 噪声：数据中真实规律起作用的程度，表示学习的难度
    - 偏差与方差权衡：偏差和方差此消彼长，不能同时消减，模型复杂度提高，偏差下降但是方差上升。

lecture 2

- 结构化数据：以二维表结构来实现逻辑表达的数据。每一行表示一次观测，每一列表示包含的变量。尽量保证每一行观测都包含独立的信息。
- 标准结构化法：将一系列以不规律频率观测的数据，化为通过规律性采样获得的衍生均匀序列。
    - 等时长采样（时长把）：以固定的时长间隔来记录数据的结构化方法
        - OHLC，open high low close价格
        - volume 总成交量
        - 时间加权平均价 TWAP
        - 成交量加权平均价 VWAP
        - 统计学特性差，非高斯，非i.i.d
        - 相同周期内金融市场的信息含量有显著不同
    - 等时点采样（时点把）：以固定的报价变动数量（每10次。。。）来记录数据的结构化方法
        - 获得接近于i.i.d正态分布的回报
    - 等量采样（成交量把）：以固定成交量来记录数据的结构化方法
        - 比等时点采样获得接近于i.i.d正态分布的回报
    - 等额采样（成交额把）：以固定成交金额来记录数据的结构化方法
        - 当价格有大幅波动时，采样的回报更加平稳，容易记录趋势；屏蔽股本变化带来的影响
- 信息驱动的结构化方法：使得新信息进入市场时能够获得更加频繁的采样，在信息匮乏时采样频率相应放慢，从而将采样和信息的流动同步，获得等**信息量**的衍生序列。
    - 时点平衡性采样：将累积买卖平衡度在一定范围内的tick数据进行汇总
    - 成交量平衡度采样：将累积的成交量平衡度限定在一定范围内的tick数据进行汇总
    - 成交额平衡性采样：将累积的成交额平衡度限定在一定范围内的tick数据进行汇总
    - 时点游程采样：将**单向的**买卖持续度限定在一定范围内的tick数据进行汇总
    - 交易量游程采样：将**单向**的成交量持续度限定在一定范围内的tick数据进行汇总
    - 交易额游程采样：将**单向的**交易额持续度限定在一定范围内的tick数据进行汇总
- 多产品序列的结构化处理
    - 动机1:需要对权重动态变化的资产组合进行建模
    - 动机2:需要对不定期支付息票或股息的金融产品进行处理
    - 动机3:需要考虑时间序列的结构变化
    - 目标：将任何复杂的多产品数据转化为一个回报类似ETF的单一数据（ETF技巧）
- 数据已经结构化，为什么还要继续进行结构化抽样(Sampling)?
    - 很多机器学习算法无法支持规模很大的数据（例如SVM）
    - 只有数据具有足够信息量时，机器学习算法才能形成具有较高预测精度的学习器。
- 结构化抽样
    - 缩减抽样（Sampling for reduction）
    - 事件驱动抽样（Event-based Sampling）：CUSUM过滤法
        - 重要的事件往往带有较强预测能力的信息
        - 事件可能与某些宏观经济统计数据的发布、波动率飙升、曲线差值与平衡水平的显著偏离等有关
        - 通过对特征识别来推定有信息量的事件
            - 机制转换（Regime shift）或结构变化（St ructural break）
            - 熵变化（Entropy Change）
            - 市场微观结构（Microstructural Information）

lecture 3

- 金融机器学习打标签的动机
  - 监督型机器学习需要有因变量集y，因变量的量化通过打标签完成
- 固定时限标签法 Fixed-time horizon labeling, FH
  - 固定时间内对于某个股票，如果其收益高于阈值c，那么被分为正例+1，否则负例-1，在-c和c之间为第三类0
  - 主要问题：c不变，但是价格波动率却随着时间变化
    - 波动率很大时，大量1，很少0
    - 波动率很小时，大量0，少量1
  - 弥补方法
    - 使用等额采样或者等量采样
    - 使用滚动指数加权移动平均（EWMA）来估计动态阈值
- 三限标签法
  - FH法打标签存在致命缺陷，没有考虑价格的路径，无法考虑在时限内提前止损（stop-loss）或止盈（profit-taking）的情况
  - Triple-Barrier,TB，路径依赖的标签法，设立两个价格上水平（horizontal）的界限和一个时间上垂直（vertical）界限
    - 水平界限用于止损止盈，使用历史波动率的函数来量化
    - 垂直界限考虑到时间期限，使用一定数量的采样把数（Bars）来量化
    - 上水平界限先被触及，标注为+1；下水平界限先被触及，标注为-1；垂直界限先被触及，标注为0或者为Sign()
  - 示性三元组[pt, sl, tp]，表示添加哪些界限，（上，下，垂直）
- 元标签法
  - 三限标签法只能训练出预测投资方向（side）的模型，无法确定投资的仓位（size）。可通过在原始模型基础上进行二次标签来训练确定仓位的模型
  - Meta-Labeling：第一次根据三限标签训练初级模型确定投资方向，优化出高查全率的模型。在初级模型确定方向的数据集中重新打出是否入场的二分类标签0，1，训练次级模型，以优化查准率为目标，通过模型给出的概率来确定仓位。
  -  ：当上水平界限先被触及：1，当下水平界限或垂直界限先被触及：0。
  - 意义
    - 仅需次级模型选择机器学习模型即可，初级模型可以使用任意模型。包括：机器学习模型，计量经济学模型，复杂统计模型，基本面模型，技术分析模型，甚至是主观定性观点。
    - 由于大多数的正例情况已经由初级模型捕捉，次级模型相当于再对初级模型的阳性判定一次，这样可以兼顾查全率与查准率。
    - 初级模型可以选择查全率100%的模型，但是全部打一个标签，查全率就100%了，没什么意义
    - 提升了模型的可解读性，先通过简单模型确定头寸方向，随后再使用复杂模型提高预测精度，可以适用于量化基本面的建模
    - 兼顾定性和定量，一定程度上限制了过拟合
    - 将判断方向和确定仓位分开，可生成复杂策略：虽然用同一个次级模型来确定买卖头寸，但是使用完全不同的初级模型确定买卖时机
    - 准确预测的小头寸但大头寸预测不准会叫人破产，开发一个仅针对关键决策（控制规模）准确性的

机器学习算法非常重要

lecture 4

- 确定样本权重的动机：因变量不独立（共享信息）
  - 机器学习无法直接P-A-P（血液检测）
- 样本重叠度和独特度（都是公式，不懂）
  - 样本重叠
  - 样本独特
  - 重叠矩阵
  - 重叠度
  - 独特度
  - 平均独特度
- **去重序贯抽样算法**
  -

lecture 5

- 进行分数差分的动机
  - 金融数据通常信噪比很低
  - 金融预测同时依赖于信号的时序记忆性和平稳性
  - 平稳性对于预测的意义：监督学习算法要求y标签的变量是平稳的，否则无法将未知新观测对应到历史已知观测
  - 整数差分造成了过度差分，为保证平稳性，但进一步降低了时序记忆性，信噪比更低
  - 缺乏时序记忆性时，使用再复杂的统计技术都无法完成预测，只能贡献错误发现
  - 分数差分能够同时兼顾记忆性和平稳性
- F-差分
- 固定窗口F-差分
  - 获得平稳的时间序列
  - 分布可能不再是正态分布
  - 分布可能具有一定的偏度
  - 分布可能具有一定的峰度
  - 选取合适的d值可以得到平稳序列
- ADF检验

lecture 6

- 金融机器学习的模型集成
  - 把一群算法相似的弱学习模型组织起来，生成一个强学习模型
- 模型集成视角下的偏差-方差分解
  - 偏差：度量拟合能力，不现实的假设导致此值偏大。机器学习算法未能识别特征和结果之间的重要关系时贡献偏差。
  - 方差：表示模型对数据的敏感性。算法错误地将噪声误认为信号，而非建模训练集中的一般模式时贡献方差
  - 噪声：表示学习的难度。这是无法由任何模型解释的不可减少的

    - 模型集成视角：进行模型集成可以有效的减小偏差或者方差
- 装袋式集成
    - 通过有放回的随机抽样生成N个训练数据集。
    - 使用N个估计器，分别从一个训练集中拟合一个估计器。这些估计器是相互独立的，因此可以并行拟合模型。
    - 模型集成
        - 连续因变量：从N个模型中得到的单个预测进行简单平均
        - 分类因变量：由对这个观察值进行分类为该类别的估计器数量所占比例（投票数）确定，也可以计算平均概率值。
    - 降低方差
- 助推式集成
    - 逐步通过弱学习器生成偏差小的强学习器
        - 初始化为均匀权重，使用随机抽样生成一个训练集
        - 使用该训练集拟合一个学习器
        - 如果单个学习器的准确率大于接受阈值（例如在二元分类器中为50%，即比随机分类好），则保留该学习器，否则丢弃
        - 给误分类的样本更多的权重，给正确分类的样本更少的权重
        - 重复前面的步骤直到生成N个学习器
        - 合成的预测值是N个模型的个体预测值的加权平均，其中权值由个体学习器的准确性确定
    - AdaBoost算法：将弱学习器合并到一个强的学习器中，每个弱学习器在训练集中被赋予一个权重，该权重随着每轮迭代进行更新，同时给误分类的样本增加权重，以提高它们的“重要性”。
- 助推式集成：与装袋式集成的比较
    - 助推式串行计算，无法并行
    - 很弱的分类器将被放弃
    - 每轮迭代时样本的权重都不相同
    - 每个学习器都有一个不同的权重
    - 主要用于解决欠拟合的问题，金融应用中主要面临过拟合问题，应以装袋式集成为主。

lecture 7

- 金融机器学习的交叉验证
    - 交叉验证CV的目的是确定模型的泛化误差
    - CV法能够评估和控制过拟合
- CV法的基本思想：将数据分割成训练集(train set)和测试集(test set)，训练集和测试集不相交。使用训练集来训练模型，使用测试集来评估模型的性能。通过优化模型超参数来选择最优模型。
- CV法的关键假设：预测数据满足i.i.d，因此测试集与训练集仅仅的差异在于噪声大小，具有相同的结构规律。
- 在CV中调超参数，导致进行多次CV，但引入了多重测试偏误。很难反映泛化误差。因此标准CV采用重抽样CV法：[训练集 + 验证集] x 多次 + 测试集
- K折交叉验证：建模样本分为K个尺度相当的子集，每次选出一个子集做验证集，其余作为测试集，然后再选另个子集做验证集，其余做测试集，...，重复K次
    - 留一K折交叉验证 Leave only one K-fold CV : LOOCV：每次放一个样本作为验证集，共进行K次交叉验证
    - 重复K折交叉验证 repeated K-fold CV：重复的进行N次K折交叉验证，每次将建模样本都分为K个尺度相

当的子集，共进行N*K次交叉验证

- 蒙特卡洛交叉验证 Monte-Calo CV： 重复的对建模样本进行B次训练集/验证集劈分，劈分时采用固定的比例划分建模样本。
- 自助交叉验证 Bootstrap CV : BCV：有放回抽取B个与建模数据等长的样本作为训练集，未抽到的样本进入验证集（袋外）。BCV衡量的平均模型方差要低于K折CV，但平均有63.2%的样本会出现一次，模型偏差与2折交叉验证相似。
- 标准CV法失效的原因
  - 重抽样CV法仍然不能完全缓解多重测试偏误
  - 因为样本重叠，预测数据不再是i.i.d
  - 重抽样CV面对非i.i.d的后果
    - 信息泄露 leakage：训练集的信息泄漏到了验证集中
- 弥补传统CV法缺陷的办法
  - 净化法（数据层面）：Yj属于验证集，Yi属于训练集，且Yj和Yi有时间重叠，则在训练时清除Yi
  - 避免容易过拟合的分类器（算法层面）：从源头上直接避免模型过拟合，比如采用装袋式集成
    - 注意设置Early stop机制
    - 注意避免训练时样本重叠，采用DSB
    - 注意采用AvgUxN生成单个分类器，保证分类器的多样性
  - 净化-隔离CV法 Purged-CV
    - 将会与验证集内的数据产生重叠的训练集内的那部分数据清除（仅清除训练集中的数据，且一个区间全部清除）
    - 隔离机制 Embargo：在劈分数据集时，因为时间序列具有序列相关的性质，应该将验证集切分处之后一段时间内的数据也从训练集中删除。因为h不充分大，仍有可能Yj 约等于 Yi。 h可取0.01T
    - 附加了隔离机制的净化CV法称为净化-隔离CV法

lecture 8

- 回测的目的
  - 通过金融机器学习模型生成的投资策略在过去的表现来推断策略在未来的表现
  - 回测要包括完整的评估在特定场景下各种变量的效果，包括投资规模、投资周期、成本变化等
  - 回测不是实验，不能用来挖掘因果关系
- 回测的常见陷阱
  - 幸存者偏差：回测数据仅包含当前活跃资产，忽略了随时间推移由于破产、摘牌或被并购的资产
  - 前视偏差：使用了在历史上看还尚未公开的信息进行决策
  - 事后诸葛亮：事后错误的寻找因果关系来证实随机（不自知）的模式
  - 数据窥探：在测试集上训练模型
  - 交易成本：设定不切实际的交易成本进行模拟
  - 异常值控制：对极端情况情况筛选或裁剪
  - 做空：做空的可行性不进行正确的评估
  - 马尔科斯回测定律
    - 马尔科斯回测第一定律：回测不是研究工具，特征重要性才是
    - 特征重要性有助于人们理解机器学习算法获得的模式的本质，但并不涉及如何使用它们盈利。回测

是基于研究结果的基础上评价盈利的可能性
    - 马尔科斯回测第二定律：回测时研究就像开车时喝酒，不要在回测中去训练模型
    - 回测的意义是剔除不好的模型，而不是去通过不断回测改进他们。通过回测去调模型会产生选择偏差，从而浪费时间
    - 在所有研究流程结束后才能进行回测，回测结果不好必须重新开始研究
- 选择偏差风险
    - 选择性偏差通过反复回测，调整模型参数，最后筛选出一个策略表现最好的回测结果
    - 风险1：无法保证最好的回测结果不是来自于运气
    - 风险2：无法保证是否模型足够复杂以至于总有一套参数可以完美拟合历史数据（拟合了噪声）
    - 回测过拟合 Backtest Overfitting: BO：选择性偏差造成的效应
- 对选择偏差的评估算法：CSCV算法 **看PPT**
    - 组合对称交叉验证：通过交叉验证方法来评估回测过拟合的概率
    - 仅适用于评估模型的选择性偏差的风险，并不是完整的回测
    - 与开发模型时评估使用的CV不同，在开发完模型之后再使用
    - 需要对模型的所有超参数组合进行总体评估
    - 对称划分训练集和验证集，保证了样本内和样本外数据的平衡性
- 历史型单路径回测算法
    - 历史模拟 Historical simulation：假设策略在曾经的历史数据执行一遍
    - 合理性：只要避免使用后视数据，历史模拟表现可以看作是假设策略在历史执行中的实际表现
    - 作为当未来历史重演时策略的真实表现
    - 历史只有一次，所以模拟出也是一条路径上策略的表现，故称之为历史型单路径回测 Historical Simple Path: HSP
    - 因为仅在一条历史路径上进行回测，很容易陷入选择性偏差，导致策略回测过拟合
    - 一段回测历史中往往包含了多个明显不同的市场环境，比如包含快牛市、股灾、灾后反弹，HSP只能是这些环境按照历史顺序演化后策略的总体结果
    - 未来环境按不同顺序不同时间长度出现时，HSP无法给出评估结果
- 场景型交叉验证回测算法
    - 动机：克服HSP的选择性偏差问题，数据通过CV算法拆分为不同的环境，然后推断特定环境下策略未来的表现
    - 场景型交叉验证 Non-historical CV: NCV：对于每个回测场景，使用除该场景时间窗口内的所有数据训练模型，生成策略。在该场景中模拟策略效果
    - 首先模拟在各种不同的环境下未来的表现，然后再按这些环境出现的历史顺序重新拼装起来产生路径上的模拟效果
    - 优点
        - 使用CV的思想进行回测，可以支持多个不同的场景（k个测试集）
        - 生成策略的训练集样本大小一致，利用的信息量一致，后期有可比性
        - 每个场景对应一个唯一的测试集，不像HSP需要设定预热窗口。这样使得每个数据都能参与回测
    - 缺点
        - 没有明确的历史型解释，它的结果不是完整的模拟策略在过去的表现以反映未来
        - NCV结果经过拼装后还是只在一条路径（代表历史路径），只形成一次推断。当需要以历史为基础建立多种场景下的策略执行效果的统计分布时，NCV无法实现
        - 在回测时可能出现信息泄露问题
- 组合清除交叉验证回测算法 Combinatorial Purged Cross Validation: CPCV **看PPT**

- 是NCV的推广，将一条历史路径扩展到多条
- CPCV法基于Purged-CV来构建场景型交叉验证，避免了NCV的信息泄露问题