金融机器学习算法

第三讲

金融机器学习中的标签

本讲主要内容

- 金融机器学习中打标签的动机
- 固定时限标签法
- 三限标签法
- 元标签法

金融机器学习打标签的动机

- 非监督机器学习允许直接从自变量集 X 中提取可预测的模式

$$X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(i)} \\ \vdots \\ X^{(I)} \end{pmatrix} = \begin{pmatrix} x_{11}, x_{12}, \cdots, \cdots, x_{1P} \\ \ddots \\ x_{i1}, x_{i2}, \cdots, \cdots, x_{iP} \\ \ddots \\ x_{I1}, x_{I2}, \cdots, \cdots, x_{IP} \end{pmatrix}$$

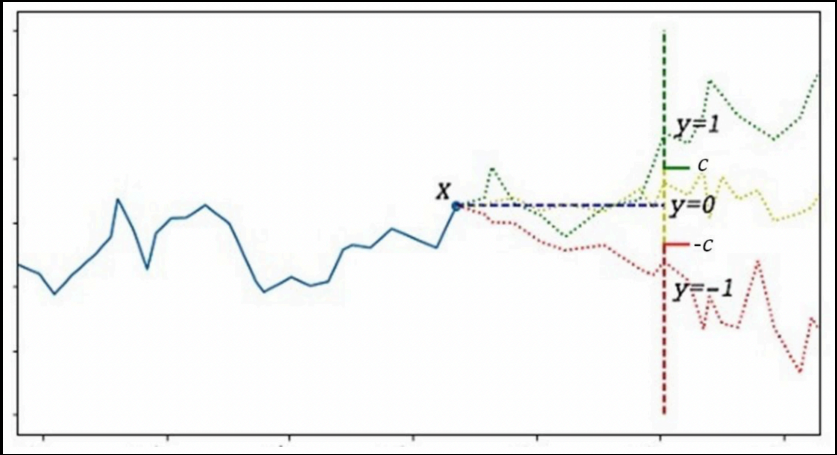
- 但监督型机器学习需要有因变量集 y，因变量的量化通过打标签完成

固定时限标签法

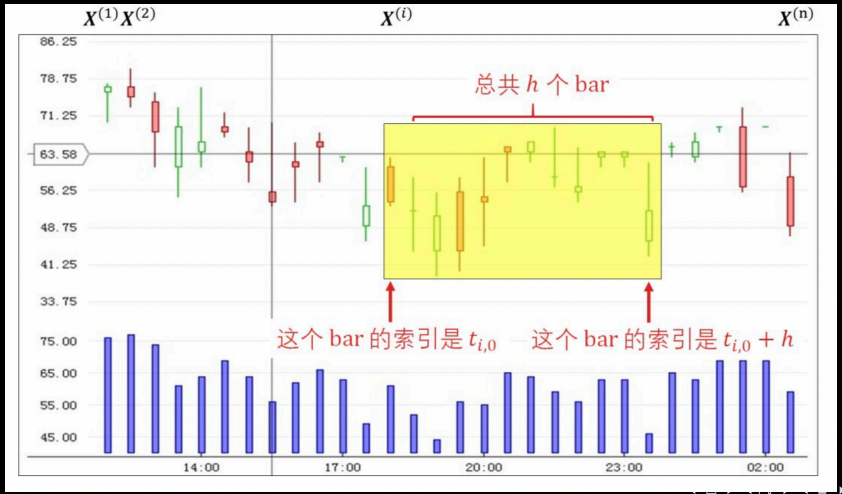
- 固定时限标签法 (Fixed-time horizon labeling, FH): 固定时间内对于某个股票，如果其收益高于阈值 c，那么被分为正例 (用 $+1$ 表示)；低于阈值 $-c$，那么被分为负例 (用 -1 表示)；如果在在 $-c$ 和 c 之间，被分为第三类 (用 0 表示)

$$y_i = \begin{cases} -1, & \text{如果 } r_{t_{i,0},t_{i,0}+h} < -c \\ 0, & \text{如果 } |r_{t_{i,0},t_{i,0}+h}| \leq c \\ +1, & \text{如果 } r_{t_{i,0},t_{i,0}+h} > c \end{cases}, \qquad r_{t_{i,0},t_{i,0}+h} = \frac{p_{t_{i,0}+h}}{p_{t_{i,0}}} - 1$$

固定时限标签法



固定时限标签法


$X^{(1)}$ $X^{(2)}$
$X^{(i)}$
$X^{(n)}$
86.25
78.75
71.25
63.58
56.25
48.75
41.25
33.75
75.00
65.00
55.00
45.00
14:00
17:00
20:00
23:00
02:00
总共 h 个 bar
这个 bar 的索引是 $t_{i,0}$
这个 bar 的索引是 $t_{i,0}+h$

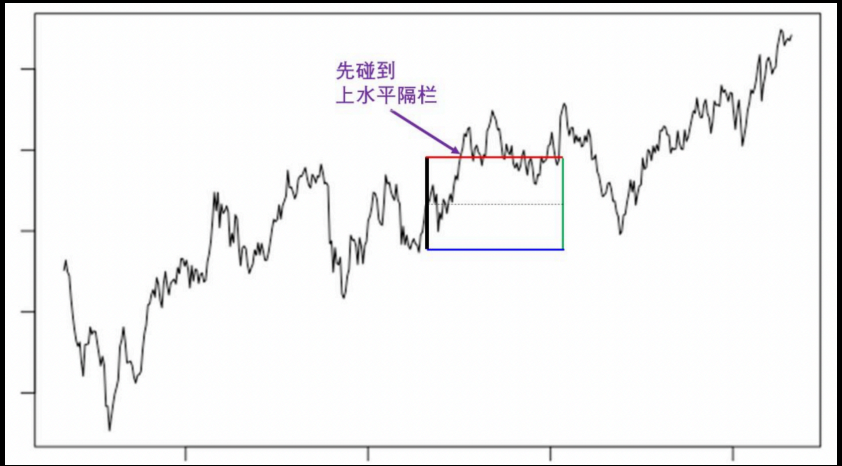
固定时限标签法

- 固定标签法的主要问题：阈值 c 选择为不变，但价格波动率却随时间变化，因此
  - 在波动率很大时，价格很容易突破 $[-c, +c]$，很少样本会被标注为 0，大量 1
  - 而波动率很小时，价格不容易突破 $[-c, +c]$，很多样本会被标注为 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 $\sigma_{t_{i,0}}$ 的动态阈值

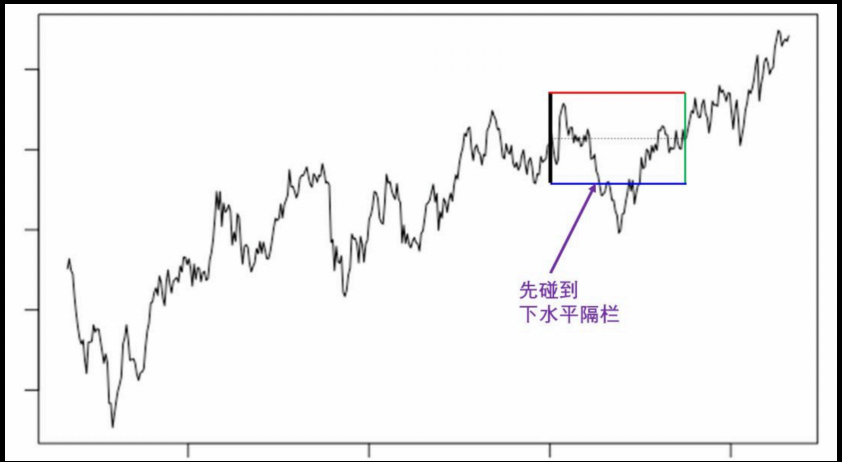
三限标签法

- 使用三限标签法的动机：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
- 三限标签法（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，
  - 水平界限用来止损止盈，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 Sign$(r_{t_{i,0},t_{i,0}+h})$

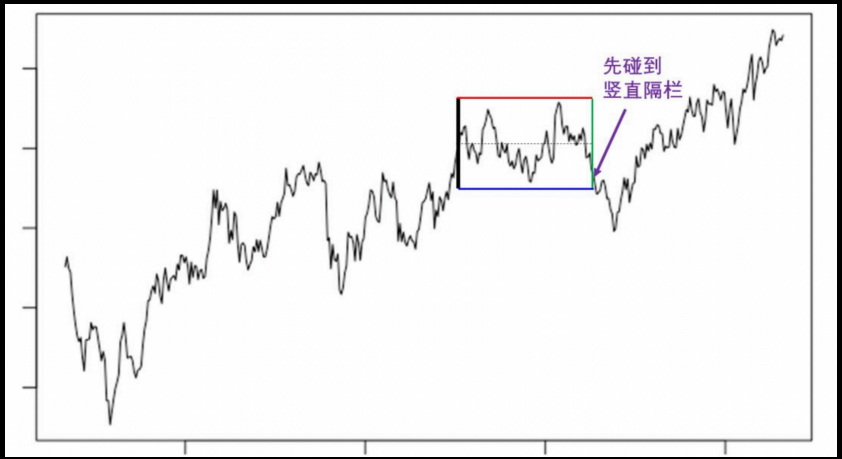
三限标签法



三限标签法



三限标签法

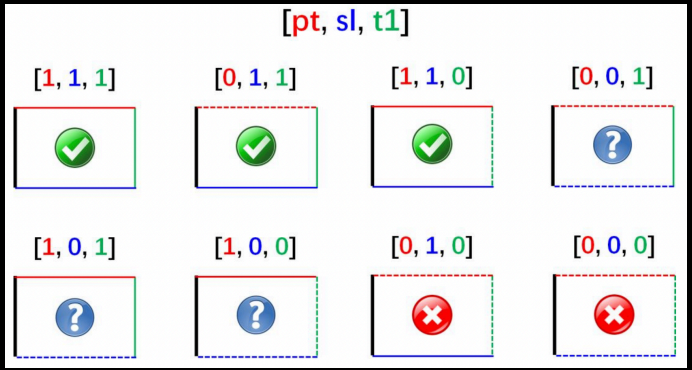


三限标签法

- 首次触及三个界限任意一个的时间使用 $t_{i,1}$ 表示，$t_{i,1} \in (t_{i,0}, t_{i,0} + h]$
- 相应的收益率为 $r_{t_{i,0}, t_{i,1}}$
- 上下两条水平界限不一定非要对称，使用二元组 $[ptSl1, ptSl2]$ 来量化界限的大小
- $ptSl_1$ 表示上界限的与 0 的距离
- $ptSl_2$ 表示下界限的与 0 的距离

三限标签法

三线标签法可以推广到比较灵活的形式，使用示性三元组 [pt, sl, tp] 来表示添加哪些界限



三限标签法

- $[1, 1, 1]$ 标准设置。希望实现盈利，但对损失和持有期限有最大限度
- $[0, 1, 1]$ 不设置止盈，要么止损退出，要么到持有期限退出。
- $[1, 1, 0]$ 仅因为需要止盈或止损时才会退出。
- $[0, 0, 1]$ 等价于 FH 标签法。
- $[1, 0, 1]$ 持有头寸直至获利或超过最长持有期，不考虑止损。
- $[1, 0, 0]$ 持仓直至获利，及时多年来一直亏损也不在乎。

元标签法

- 动机：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
- 元标签法 (Meta-Labeling) : 第一次，根据三限标签 $y^d = (+1, 0, -1)$ 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 $y^z = (1, 0)$，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。

元标签法

以做多头为例，首先进行第一次标签

- $y^d = 1$，当上水平界限先被触及
- $y^d = -1$，当下水平界限先被触及
- $y^d = 0$，当垂直界限先被触及

训练初级模型后，在所有预测 $\hat{y}^d = 1$ 的数据中，重新打二次标签

- $y^z = 1$，当上水平界限先被触及
- $y^z = 0$，当下水平界限或垂直界限先被触及

训练次级模型。

元标签法

使用元标签法的意义

- 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。这样可以同时兼顾查全率与查准率。
- 【思考】：初级模型选一个查全率为 100% 的模型，不可以吗?

元标签法

使用元标签法的优势

- 提升了模型的可解读性。先通过简单模型（如基本面或者人的看法）来确定头寸方向，随后再使用复杂模型（如机器学习模型）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尊预测不准确会叫人破产。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