# 金融机器学习算法

第四讲

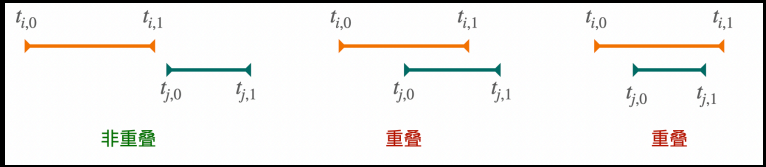
## 金融机器学习的样本权重

# 本讲主要内容

- 确定样本权重的动机
- 样本重叠与独特度
- 去重序贯抽样算法
- 回报率归因权重计算

# 确定样本权重的动机

- 不同样本对应的因变量 $y_i, y_j$ 很大可能不是独立的，但是机器学习的基础理论往往要求样本服从 $i.i.d$
- 因变量不独立（共享信息）：$y_i \sim [t_{i,0}\, t_{i,1}]$，$y_j \sim [t_{j,0}\, t_{j,1}]$，以下三种情况中，有两种情况会造成 $y_i, y_j$ 不独立




- 在金融应用信息重叠很常见：机器学习无法直接 P-A-P (血液检测)
- 通过设计权重来修正重叠

# 样本重叠度与独特度

- 样本重叠 (concurrent)：如果两个样本的因变量标签 $y_i, y_j$ 包含一项或多项共同回报 $r_{t-1,t}$。
- 样本独特 (uniqueness)：样本 $i$ 的因变量标签不予任何其他样本包含有共同回报 $r_{t-1,t}$
- 重叠矩阵: $[1_{t,i}]_{T\times N}$，$t = 1\cdots T$, $i = 1\cdots N$
- 重叠度：特定时点 $t$ 上重叠的样本数量 $c_t = \sum_{i=1}^{N} 1_{t,i}$
- 独特度：特定样本 $i$ 在特定时点上的独特程度 $u_{t,i} = \frac{1_{t,i}}{c_t}$
- 平均独特度: 特定样本 $i$ 的总体独特度评价 $\bar{u}_i = \frac{\sum_t u_{t,i}}{\sum_t 1_{t,i}}$

# 样本重叠度与独特度：$[t-1, t] \in [t_{i,0}, t_{i,1}]$

# 样本重叠度与独特度：例 1 （稀疏整齐）

| $t$ \ $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $t_1$ | × | | | | |
| $t_2$ | × | × | | | |
| $t_3$ | × | × | × | | |
| $t_4$ | | × | × | | |
| $t_5$ | | | × | | |
| $t_6$ | | | | × | |
| $t_7$ | | | | × | × |
| $t_8$ | | | | × | × |
| | | | | | |

$u_{t,i}$

| $c_t$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0.5 | 0.5 | 0 | 0 | 0 |
| 3 | 0.33 | 0.33 | 0.33 | 0 | 0 |
| 2 | 0 | 0.5 | 0.5 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0.5 | 0.5 |
| 2 | 0 | 0 | 0 | 0.5 | 0.5 |
| | 0.61 | 0.44 | 0.61 | 0.66 | 0.5 |

$\bar{u}_i$

# 样本重叠度与独特度：例 2 （稠密）

| $t$ \ $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $t_1$ | | | | × | | | |
| $t_2$ | | | × | × | | | |
| $t_3$ | × | | × | × | | | |
| $t_4$ | × | | × | × | | × | |
| $t_5$ | × | × | × | × | | × | |
| $t_6$ | | × | | | | × | × |
| $t_7$ | | × | | | | | × |
| $t_8$ | | × | | | × | | |
| $t_9$ | | | | | × | | |

$u_{t,i}$

| $c_t$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1/2 | 1/2 | 0 | 0 | 0 |
| 3 | 1/3 | 0 | 1/3 | 1/3 | 0 | 0 | 0 |
| 4 | 1/4 | 0 | 1/4 | 1/4 | 0 | 1/4 | 0 |
| 5 | 1/5 | 1/5 | 1/5 | 1/5 | 0 | 1/5 | 0 |
| 3 | 0 | 1/3 | 0 | 0 | 0 | 1/3 | 1/3 |
| 2 | 0 | 1/2 | 0 | 0 | 0 | 0 | 1/2 |
| 2 | 0 | 1/2 | 0 | 0 | 1/2 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1/2 | 0 | 0 |
| | 47/180 | 23/60 | 77/240 | 137/360 | 1/2 | 47/180 | 5/12 |

$\bar{u}_i$

# 去重序贯抽样算法

为什么抽样会重叠

- 假如一共有 $N$ 个代抽样本，一共放回式抽取 $I$ 次
- 假如非重叠的样本最大数量为 $K$
- 对于 $K$ 个样本中的任意一个样本，放回式抽取 $I$ 次后出现 $i$ 次的概率是
  $$\mathbb{P}(i) = C_I^i \left(\frac{1}{K}\right)^i \left(1 - \frac{1}{K}\right)^{I-i}$$
- $I \to \infty$，$P(i) \to \dfrac{(\frac{I}{K})^i e^{-\frac{I}{K}}}{i!}$，服从 $\lambda = \dfrac{I}{K}$ 的 Poisson 分布，均值为 $I/K > 1$
-

# 去重序贯抽样算法

**Algorithm 1:** 去重序贯抽样算法 (De-overlapping Sequential Bootstrap; DSB)

**Result:** 产生重叠度较低的抽样，尽量满足 $i.i.d$

确定重叠矩阵 $[1_{t,i}]_{T\times N}$，$t = 1\cdots T$, $i = 1\cdots N$，按 $i \sim U[1, I]$ 抽出第一个样本 $i_1$，抽样序列 $\phi^{(1)} = \{i_1\}$;

**while** $k \leqslant I$ **do**

(1) 计算 $u_{t,i}^{(k)} = \dfrac{1_{t,i}}{1 + \sum_{j\in\varphi^{(k-1)}} 1_{t,j}}$，$\bar{u}_i^{(k)} = \dfrac{\sum_t u_{t,i}^{(k)}}{\sum_t 1_{t,i}}$;

(2) 更新抽样的概率 $p_i^{(k)} = \dfrac{\bar{u}_k^{(k)}}{\sum_i \bar{u}_i^{(2)}}$;

(3) 抽取 $k$ 轮的样本 $i_k$, $\phi^{(k)} = \phi^{(k-1)} \bigcup \{i_k\}$

**end**

# 去重序贯抽样算法：例子

| $t$ \ $i$ | 1 | 2 | 3 |
|---|---|---|---|
| 1 | × | | |
| 2 | × | | |
| 3 | × | × | |
| 4 | | × | |
| 5 | | | × |
| 6 | | | × |

$u_{t,i}^{(1)}$

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| | 1 | 1 | 1 |
| | 1/3 | 1/3 | 1/3 |

$\bar{u}_i^{(1)}, p_i^{(1)}$

$u_{t,i}^{(2)}$

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 2 | 0.5 | 0.5 | 0 |
| 2 | 0 | 0.5 | 0 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| | 5/6 | 0.5 | 1 |
| | 5/14 | 3/14 | 6/14 |

$\bar{u}_i^{(1)}, p_i^{(1)}$

$u_{t,i}^{(3)}$

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 2 | 0.5 | 0.5 | 0 |
| 2 | 0 | 0.5 | 0 |
| 2 | 0 | 0 | 0.5 |
| 2 | 0 | 0 | 0.5 |
| | 5/6 | 0.5 | 0.5 |
| | 10/22 | 6/22 | 6/22 |

$\bar{u}_i^{(1)}, p_i^{(1)}$

$\phi^{(1)} = \{2\}$ $\phi^{(2)} = \{3,3\}$

# 回报归因权重计算

- 原理：回报的绝对值较大因变量对应的样本权重应该相应较大，但重叠性较大时权重应该减小
- 算法无论对初始标签还是元标签都适用
- 对于第 $i$ 个样本，标签的时间跨度为 $[t_{i,0}, t_{i,1}]$，则样本权重应该为

$$\tilde{w}_i = \left| \sum_{t=t_{i,0}}^{t_{i1}} \frac{r_{t-1,t}}{c_t} \right|$$

$$w_i = \frac{\tilde{w}_i I}{\sum_{j=1}^{I} \tilde{w}_j}$$